# Information Theory For Data Management

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

Suresh Venkatasubramanian
University of Utah
suresh@cs.utah.edu

## 1. INTRODUCTION

We are awash in data. The explosion in computing power and computing infrastructure allows us to generate multitudes of data, in differing formats, at different scales, and in inter-related areas. Data management is fundamentally about the harnessing of this data to extract information, discovering good representations of the information, and analyzing information sources to glean structure. Data management generally presents us with cost-benefit tradeoffs. If we store more information, we get better answers to queries, but we pay the price in terms of increased storage. Conversely, reducing the amount of information we store improves performance at the cost of decreased accuracy for query results. *The ability to quantify information gain or loss* can only improve our ability to design good representations, storage mechanisms, and analysis tools for data.

Information theory provides us with the tools to quantify information in this manner. It was originally designed as a theory of data communication over noisy channels. However, it has more recently been used as an abstract domain-independent technique for representing and analyzing data. For example, entropy measures the degree of disorder in data and mutual information captures the idea of noisy relationships among data. In general, viewing information theory as a tool to express and quantify notions of information content and information transfer has been very successful as a way of extracting structure from data [14, 3, 9, 5, 7, 8, 2].

In this tutorial, we will explore the use of information theory as part of a data representation and analysis toolkit. We will do this with illustrative examples that span a wide range of topics of interest to data management researchers and practitioners. We will also examine the computational challenges associated with information-theoretic primitives, indicating how they might be computed efficiently.

## 2. INFORMATION THEORY BASICS

The tutorial will start with an introduction to the relevant concepts in information theory. Starting with the notion of a discrete distribution and its relation to histograms, we will introduce notions such as entropy, the KL-distance, conditional entropy and mutual information.

There are two alternate views of mutual information that are informative from a data management perspective.

**I: Measuring Strength Of Signal.** Suppose $f : X \to Y$ is an invertible function. Then given $y = f(x)$, we can recover $x$ with certainty. If we introduce some noise into the system then, for a given $y$, we might no longer be able to recover a unique $x$. Rather a specific $y$ might imply a *distribution* over possible values of $x$.

Mutual information captures this notion of a noisy mapping. For instance, a functional relationship common in database systems is a key: a single value of the key uniquely defines the other fields in a tuple. But if the mapping is corrupted, then we can measure the mutual information of the mapping between the key and the tuple to quantify the degree to which the mapping is deterministic. As the mapping gets more and more corrupt, the mutual information decreases, and finally becomes 0.

**II: Complexity of Representation** A simple transformation allows us to rewrite the mutual information between variables $X$ and $Y$ as the "average KL-distance" to the center of a cluster, where each $p(y|x)$ is a vector. This means that mutual information can play a role similar to the sum-of-squares cost measure used in $k$-means clustering, and acts as a measure of the complexity of representation of a cluster (informally, the average number of bits needed to write down a description of the cluster).

## 3. APPLICATIONS

In this unit, we will illustrate the conceptual roles played by information-theoretic quantities through a series of applications in data management. This list of examples is by no means exhaustive, but it is representative of the diversity of problems where information-theoretic tools have proved useful.

*Measuring Information Content.* The entropy of a distribution characterizes the average number of bits needed to write down an element of the distribution. More entropy indicates that the distribution contains more information, and less entropy indicates that it has less.

Arenas and Libkin [6] exploit this idea to quantify the notion of *redundancy* in normal forms. Intuitively, a normal form should be non-redundant in order to avoid update anomalies, and they show that BCNF (and other forms like 4NF, 5NF etc.) is indeed non-redundant, using entropy as


*VLDB '09,* August 24-28, 2009, Lyon, France

a measure of information content.

Measuring information content via entropy works when items are either the same or completely distinct. But if data items can be similar to each other, then it is important to distinguish between items that are close and items that are far away. If we cluster the data into clusters, then the resulting distribution on cluster sizes gives us a more general notion of information content. This is exploited in work by Dai *et al* [7].

*Data Linkage In Schema Matching.* A key problem in data-driven schema matching is determining which columns of data are more likely to be associated with each other than others. Three recent works [3, 12, 8] approach the problem of schema matching from different perspectives, but are linked by the idea that the mutual information between attributes, viewed as a measure of similarity, can be exploited to drive schema match discovery in heterogeneous sources.

*Data Anonymization.* Anonymization of data prior to publishing is a key area of interest right now. Central to research in this area is the quantification of privacy loss, or leakage of information. Information-theoretic operators play a crucial role here: informally, the Kullback-Leibler distance between the prior knowledge about data and the posterior knowledge (after anonymization) is a measure of the amount of information *leaked* to a potential adversary. The conditional entropy has also been used in this context [1, 10, 11]. Modeling the background knowledge of the adversary is another challenging problem that has been tackled with information-theoretic methods [13].

# 4. ESTIMATION

Information-theoretic concepts are effective in practice because of concurrent work in the area of streaming and sampling that allows us to estimate these quantities efficiently over large data sets.

We will first present an overview of methods for estimating entropy and mutual information from large data sets. We will then turn to techniques for clustering data using information-theoretic principles, describing both hierarchical and divisive methods[15], as well as approaches that work well in a large-data setting[4, 8].

# 5. CONCLUSION

We see two main learning outcomes from this tutorial. In the short term, we expect that this tutorial, by providing an information theory toolkit, will lead to a more effective use of information theory in a principled fashion in data management applications. Taking a more long-term view, we hope that understanding the role of information theory in the modeling, representation and analysis of data will lead to a better understanding and utilization of the tradeoff between cost and benefit when designing data management systems.

# 6. REFERENCES


[1] Agrawal, D., and Aggarwal, C. C. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems* (2001), pp. 247–255.

[2] Ahmadi, B., Hadjieleftheriou, M., Seidl, T., Srivastava, D., and Venkatasubramanian, S. Type-based categorization of relational attributes. In *Proc. 12th International Conference on Extending Database Technology (EDBT)* (2009).

[3] Andritsos, P., Miller, R. J., and Tsaparas, P. Information-theoretic tools for mining database structure from large data sets. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data* (2004), pp. 731–742.

[4] Andritsos, P., Tsaparas, P., Miller, R., and Sevcik, K. LIMBO: Scalable Clustering of Categorical Data. *Lecture Notes In Computer Science* (2004), 123–146.

[5] Arenas, M., and Libkin, L. An information-theoretic approach to normal forms for relational and xml data. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (2003), pp. 15–26.

[6] Arenas, M., and Libkin, L. An information-theoretic approach to normal forms for relational and xml data. *J. ACM 52*, 2 (2005), 246–283.

[7] Dai, B. T., Koudas, N., Ooi, B. C., Srivastava, D., and Venkatasubramanian, S. Rapid identification of column heterogeneity. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining* (2006), pp. 159–170.

[8] Dai, B. T., Koudas, N., Srivastava, D., Tung, A. K. H., and Venkatasubramanian, S. Validating multi-column schema matchings by type. In *24th International Conference on Data Engineering (ICDE)* (2008), pp. 120–129.

[9] Dalkilic, M. M., and Roberston, E. L. Information dependencies. In *PODS* (2000), pp. 245–253.

[10] Evfimevski, A., Gehrke, J., and Srikant, R. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the ACM SIGMOD/PODS Conference* (June 2003), pp. 211–222.

[11] Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. Privacy preserving mining of association rules. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (2002), pp. 217–228.

[12] Kang, J., and Naughton, J. F. On schema matching with opaque column names and data values. In *SIGMOD* (2003), pp. 205–216.

[13] Martin, D. J., Kifer, D., Machanavajjhala, A., Gehrke, J., and Halpern, J. Y. Worst-case background knowledge for privacy-preserving data publishing. In *ICDE* (2007), pp. 126–135.

[14] Pantel, P., Philpot, A., and Hovy, E. H. An information theoretic model for database alignment. In *Proc. 17th Intnl. conf. on scientific and statistical database management (SSDBM)* (2005), pp. 14–23.

[15] Tishby, N., Pereira, F., and Bialek, W. The information bottleneck method. In *Proc. 37-th Annual Allerton Conference on Communication, Control and Computing* (1999), pp. 368–377.