DOI: 10.1145/1592761.1592796

BY FRANK NIELSEN

Steering Self-Learning Distance Algorithms

THE CONCEPT OF DISTANCE EXPRESSES THE DISTORTION measure between any pair of entities lying in a common space. Distances are at the very heart of geometry, and are ubiquitous in science, needless to say in computational science. From the vantage point of physics, distances may be interpreted as the smallest amount of energy required to go from one location to the other, or to morph from one state to the other. Unfortunately, there is a lot of confusion in popular press about what is exactly meant by using the wording "distance." For example, people quite often interchange "distance" with "metric" without caring much about the implicitly underlying mathematical properties: In this case, to know whether the triangle inequality axiom is satisfied or not. A great deal of efforts was achieved by Deza and Deza[6] in 2006 by publishing the first dictionary of distances presenting succinctly but unambiguously the various properties of distortion measures (such as, metrics, semi-metrics, distances, quasi-distances, divergences, etc.), and listing an extensive although non-exhaustive catalog of principal distances with their domain of applications encountered in both natural sciences (biology, chemistry, physics, and cosmology), and computer sciences (coding theory, data mining, and audio/video processing).

Algorithm designers and researchers in computational sciences daily face the daunting task of choosing the most appropriate distance functions for solving their specific problems at hand. It is clearly understood nowadays that the usual flatland Euclidean distance is rarely appropriate for solving tasks on high-dimensional heterogeneous datasets that are rather lying on curved manifolds. A simple toy argument is to consider the task of averaging rotation matrices. Rotation matrices are orthonormal matrices of unit determinant. Unfortunately, the center of mass of a set of matrices, for example, the centroid defined as the arithmetic mean of rotation matrices is not a rotation matrix so that a regularization method is required to cast the average matrix to the closest rotation matrix. Consider yet another example: partial 3D shape retrieval. In partial shape retrieval, a user queries a database of 3D objects with a given part. Solving this problem requires to consider an oriented distance to break the symmetry rule. That is, one would like the distance part to object to be greater than the distance object to part, for all parts belonging to the given object. Indeed, to clarify this point, consider the distance of a 3D wagon to a 3D train model consisting of a locomotive attached to several wagon units. This distance wagon to train should be strictly greater than the distance of the same 3D train to the said wagon. This kind of asymmetric property is fulfilled by the relative entropy distance, known also as the Kullback-Leibler divergence that acts on statistical distributions. Liu et al.[9] built an efficient and accurate 3D part search engine inspired by probabilistic text analysis technique by considering 3D objects as documents covering a small number of topics called "shape topics." They experimentally showed that the relative entropy distance behaves significantly better than the Euclidean

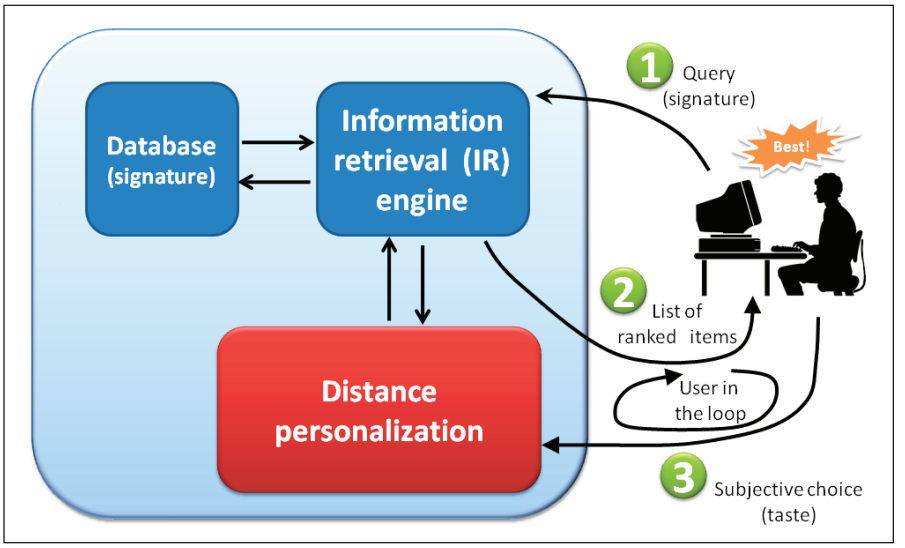
or vector space model weighted cosine distances. It is natural to ask oneself whether this Kullback-Leibler distance is the best ultimate distance function for 3D search engines or not?

It turns out that this subtle question cannot be settled in a static way as it depends on the considered input database and on the not-yet-known online queries to be processed in the future. Otherwise, adversorial input sets could be purposely designed to prove the sub-optimality of any prescribed distance function. That is, distances need to be tuned up for every single input set by a built-in learning process. Further, these algorithmic distances need to be dynamically maintained as objects are added, edited, or deleted in the database. This dynamic paradigm of selecting distances bears much similarity with the recent concept of self-improved algorithms[1] that devote some of their computational time to learn distribution characteristics of the input data sets to be able to speed up the overall process.

Since the space of potential distance functions is uncountably infinite, designing self-learning distance algorithms need to proceed first by choosing a small set of axiom rules (such as symmetry, or triangle inequality) specifying the type of distances, and yielding parameterized distance families for each class. For example, back to 1991, Csiszár1 axiomatically derived the 1D parametric family of so-called Bregman-Csiszár distances by generalizing the principles of orthogonal projection measures in least-square-type optimization problems. This generalization let us discovered some counter-intuitive facts a priori, such as the non-necessarily commutative property of orthogonal projections. The Bregman-Csiszár parametric family includes the Kullback-Leibler and the Itakura-Saito divergences at its extremities. Thus learning the best Bregman-Csiszár distance for a given input amounts to find its best member subject to problem-specific constraints.

In information retrieval (IR) systems such as the former partial shape search engine, a set of features playing the role of signature are first extracted from every single input element, and an overall appropriate distance function is properly defined on the signature space.

Learning Distance



In modern information retrieval (IR) systems, users are able to dynamically steer the distance learning process of search algorithms to reflect accurately and efficiently their subjective tastes: First, a user query the IR engine (1), and a list of top ranked items are returned (2), from which the user selects his/her best matches (3), thus allowing the system to personalize accordingly the distance function for future queries.

Then, given a query object, its signature data point is first computed (feature extractor) and its nearest-neighbor is searched for among all input signatures. In practice, better classification methods such as the k-NN rule that consists in taking the majority class of the k nearest neighbors (NN), or using kernel machines such as popular support vector machines (SVMs) are employed. Geometrically speaking, the input signatures yields a partition of the signature space into discrete elementary proximity volumes, called Voronoi cells that represent the locus of signature points closer to the cell's signature than to any other input signature. Such a discrete Voronoi diagram implicitly encodes the shape of signature data points. Interestingly, Voronoi diagrams have been recently generalized to the parameterized family of Bregman divergences[10] as well, unifying both the classic ordinary Euclidean diagram with entropic statistical diagrams into a single unifying framework.

Furthermore, we seek for efficiency reasons to reduce the number of input signatures to keep only its most representative elements. This is achieved by using a technique called vector quantization that clusters the points into groups such as to minimize the overall intra-cluster distance, while maximizing the inter-cluster distance. The seminal centroid-based k-means cluster algorithm originally established in 1957 by Lloyd has also been recently generalized to its provably most generic class of distance measures by a breakthrough result of Banerjee et al.[2] in 2005: Namely, the class of Bregman divergences.

One can legitimately ponder whether such self-learning distance algorithms are indeed the best suited strategy to get the optimal solution. These tweaked algorithms are indeed presumably the best whenever the objective or loss functions are unambiguously defined from the input datasets. But no one would doubt on the subjective part of defining the "closest" 3D shape to a given collection of 3D shapes. Human perception then plays a determinant role, and answers are all but subjective, reflecting the different tastes of individuals. Therefore another recent line of research is to let users steer themselves the distance learning process by loosely entering preferences. These personal user preferences are entered either by clicking on the best subjective ranked item in a list of top matches, or by providing prior information such as "I find these two images quite similar but these two others are rather far apart, etc." that are handled as like/dislike constraints. These semi-supervised learning problems become a hot topic in machine learning as attested by the increasing number of publica-

tions related to this area.

For example, Bar-Hillel and Weinshall[3] described such a semi-supervised learning algorithm where users give positive/negative equivalence constraints denoting intra-cluster/inter-cluster pairs of points. The problem one faces then is to extract as precisely (for example, numerically) and reliably as possible the information provided by users. The thesis of Hertz[7] provides an excellent review of distance learning techniques starting from the most common Mahalanobis metric learning algorithms that generalize the usual Euclidean metric to more flexible non-parametric distance learning methods. For example, the non-parametric "DistBoost" distance[4] is derived from signed margin values of binary classifiers combined altogether in the spirit of a machine learning technique called boosting. The recent edited book of Basu et al4 on constrained clustering further describes many other semi-supervised clustering techniques that support user feedback.

Learning distances play also a crucial role in the field of collaborative filtering. The seminal idea of collaborative filtering was originally presented by Goldberg et al.[7] in 1992 at Xerox Palo Alto Research Center in their experimental mail system called Tapestry. Nowadays, collaborative filtering is used in many commercial systems including Amazon book store and eBay auction site. The underlying idea of collaborative filtering is that information filtering is much more effective when humans are part of the filtering task too. That is, both user-steered content-based filtering and group-steered collaborative filtering are used to quickly retrieve insightful documents given a huge set of annotations gleaned from many users.

In the near future, we envision a whole new generation of scaleable personalized information retrieval systems driven by novel algorithms incorporating self-learning built-in distance modules, and providing light user interfaces. These brand new search engines would be able to better listen to the voice of their users, and more importantly give adequate feedbacks to the full information retrieval engine about users/groups subjective tastes, all at the clicks of a mouse. C

References

1. Ailon, N., Chazelle, B., Comandur, S., and Liu, D. Self-improving algorithms. In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms* (SIAM press, Philadelphia, PA), 261-270, 2006.
2. Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with Bregman divergences, *J. Machine Learning Research 6*, (2005), 1705 – 1749.
3. Bar-Hillel, A., and Weinshall, D. Learning distance function by coding similarity. In *Proceedings of 24th International Conference on Machine Learning* (Madison, WI, 2007), Omnipress, 65-72.
4. Basu, S., Davidson, I., and Wagstaff, K. *Constrained Clustering: Advances in Algorithms, Theory and Applications*, Chapman & Hall/CRC, 2008.
5. Csiszár, I. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals of Statistics 19*, 4, (1991) 2032-2066.
6. Deza, E., Deza, M. M. *Dictionary of distances*, Elsevier, 2006.
7. Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. Using collaborative filtering to weave an information tapestry. *Comm. ACM 35*, 12, (Dec. 1992), 61-70.
8. Hertz, T. Learning distance functions: Algorithms and applications. PhD thesis, 2006.
9. Liu, Y., Zha, H., and Qin, H. Shape topics: A compact representation and new algorithms for 3D partial shape retrieval. In *Proceedings of IEEE Computer Vision and Pattern Recognition Conference* (CA 2006), 2025-2032.
10. Nielsen, F., Boissonnat, J.-D., and Nock, R. *Bregman Voronoi Diagrams: Properties, Algorithms and Applications*. INRIA Research Report 6154 (Sophia-Antipolis, France 2007).

Frank Nielsen (frank.nielsen@acm.org) is in the Fundamental Research Department of Sony Computer Science laboratories, Inc., in Shinagawa-Ku, Tokyo, Japan, and in the Computer Science Department (LIX) of Ecole Polytechnique, Palaiseau, France.