Non-Metric Similarity Search Problems in Very Large Collections

Benjamin Bustos [#1], Tomáš Skopal [*2]

[#] *PRISMA Research Group, Department of Computer Science, University of Chile*
Av. Blanco Encalada 2120, 8370459 Santiago, Chile
[1] bebustos@dcc.uchile.cl

[*] *SIRET Research Group, Faculty of Mathematics and Physics, Charles University in Prague*
Malostranske nam. 25, 118 00 Prague, Czech Republic
[2] skopal@ksi.mff.cuni.cz

***Abstract*—This tutorial surveys domains employing non-metric functions for effective similarity search, and methods for efficient non-metric similarity search in very large collections.**

I. Introduction

Similarity search is a fundamental problem in many disciplines like multimedia databases, data mining, bioinformatics, computer vision, and pattern recognition, among others. The standard approach for implementing similarity search is to define a dissimilarity measure that satisfies the properties of a metric (strict positiveness, symmetry, and the triangle inequality), and then use it to query for similar objects in large data collections. The advantage of this approach is that there are many index structures (so-called metric access methods) that can be used to efficiently perform the queries. However, a recent survey [1] has shown that similarity measures not holding the metric properties have been widely used for content-based retrieval, because these (usually) complex similarity measures are more effective, i.e., they return better results.

The goal of this tutorial is to provide an interdisciplinary overview of non-metric similarity measures and their applications, focusing on their usage with very large data collections. We start the tutorial presenting the basics of similarity search and the motivation for using non-metric similarity measures. Next, we present many general non-metric measures, that can be used in a wide variety of application domains, and several different research areas that share the necessity of efficient similarity search algorithms in non-metric spaces. Then, the efficiency issue will be addressed, describing the current state-of-the-art in non-metric indexing, both for general and specific non-metric measures. Finally, we end the tutorial with a summary of the current techniques for searching with non-metric measures in large data collections, remarking the current challenges for the database community on this topic.

II. Outline

A. Introduction

The tutorial begins with an introduction to the topic of similarity measuring and search, with a focus on similarity spaces, the basics of similarity search [2], [3], and the topological properties [4], [5]. It continues motivating the use of non-metric functions for similarity search. This part ends with a discussion of the design and implementation issues of similarity functions, and how to turn a similarity into a dissimilarity function.

B. Non-metric similarity functions and application domains

We continue the tutorial giving details for a wide variety of non-metric similarity functions. Non-metric functions are often more robust than metric distances and they may consider local similarity, leading to better effectiveness in the similarity search result. Also, a black-box algorithm that computes some similarity score for an specific application can rarely be proved to satisfy the metric properties.

1) Non-metric functions: We start describing general-purpose non-metric measures such as Dynamic Partial Function [6], [7], Cosine measure and distance [8], Kullback-Leibler divergence, χ^2 distance [9], Dynamic Time Warping Distance [10], and Hausdorff Distance Variants [11].

2) Application domains and examples: We discuss specific application domains where non-metric similarity have been successfully used, such as:

- Multimedia databases (e.g., images [12], shapes [13], [14], [15], [16], audio [17], [18], [19], [20] digital libraries [21], and XML databases [22]).
- Medical and scientific databases (imaging data [23], [24], [25], time series [26], [27], [28], [29], and graph databases [30], [31], [32]).
- Biological and chemical databases (Primary and tertiary structure of proteins [33], [34], [35], [36], [37], and general molecules and compounds [38], [39]).
- Biometric databases (handwritten recognition [40] and face identification [41]).

C. Efficient search in non-metric spaces

Although there have been many approaches developed addressing efficient (fast) similarity search, they mostly share the general *lower-bounding concept*. That is, instead of computing the exact distance δ, a *tight lower-bound distance* δ_{LB}

($\delta_{LB}(q,x) \leq \delta(q,x)$) is computed and tried to filter the object x from the search. The crucial assumption here is the computation of δ_{LB} is much cheaper than that of δ.

1) The Metric Case: Before we discuss the general case of non-metric similarity, we present the access methods for the metric case – the *metric access methods* (MAMs) [42], [2]. Here, the similarity function is restricted to a metric distance δ, so that lower-bounding distance values are constructed by using precomputed distances to the so-called pivots (reference objects from database) and the metric postulates, especially the triangle inequality. We present the general principles of indexing metric spaces, including the ball-partitioning and hyperplane-partitioning schemes.

An important factor when indexing by MAMs is the distance (similarity) distribution within the metric dataset, which also determines its *indexability*. We present two indexability measures, the intrinsic dimensionality [42] and the ball-overlap factor [43]. High values of these measures indicate the dataset is poorly structured, so that efficient indexing and search is not possible due to lack of distinct clusters of similar objects.

2) Mapping: Most of the approaches to efficient non-metric search have been based on a kind of transformation of the non-metric search problem into a metric one. The early approaches even treated the non-metric similarity as it would be a metric distance, turning thus an exact metric search technique into an approximate non-metric search technique. Among the early approaches, we name various embeddings into (Euclidean) vector spaces [44], [45], [46], [47], [48], where the entire dataset is (expensively) transformed. The other, more recent methods, do not transform the dataset itself, as they rather transform the non-metric distance to become (behave as) a metric one. This could be done by modifying the distance function itself (as in the TriGen algorithm [43]), or by relaxing the triangle inequality property (as in the Constant Shifting Embedding [49]).

There also appeared approaches that completely abandoned the metric space paradigm, replacing it by a different one. For example, the fuzzy similarity approach [50], [51] employed the power of fuzzy logic to efficiently search non-metric spaces. Similarly, the ptolemaic indexing [52] replaced the triangle inequality by the more powerful Ptolemy inequality, leading to efficient similarity search under an expensive distance function. Recently, a combinatorial framework for general similarity search was proposed [53], where a "comparison oracle" is used instead of the distance function, while the lower-bounding is computed by the so-called disorder inequality.

3) General non-metric access methods: Furthermore, we present some pioneer access methods that natively provide general non-metric search. The early approaches were implemented as clustering and classification problems [54], [55], [7], [56]. The QIC-M-tree [57], on the other hand, allows to query using a non-metric distance on a metric index (the M-tree) constructed using a metric distance that lower-bounds the query distance. Recently, the NM-tree [58] combined the M-tree and the TriGen algorithm, allowing the user to set the retrieval precision at query time. The Local Constant Embedding [59] uses the constant shifting embedding together with a suitable clustering of the dataset into groups that allows a more effective lower-bounding in each group.

4) Distance-specific non-metric access methods: Besides the techniques designed to search by arbitrary non-metric distances, there was a number of techniques proposed, assuming a specific distance and/or data representation. The inverted file [60] is an example of a distance-specific index which efficiency is constrained by the usage of the cosine similarity and sparse query vectors. The inverted file was also an inspiration for the IGrid [61], an index for non-metric L_p-like distances. In time series retrieval, the popular measures are the dynamic time warping distance and the longest common subsequence. Both measures are non-metric, while there have been techniques proposed for their efficient (metric) lower-bounding [62], [63]. Finally, the area of bioinformatics is a huge repository of specific approaches to non-metric similarity search. In particular, FASTA [64] and BLAST [65] are popular techniques for non-metric similarity search in protein databases.

D. Conclusions

Non-metric similarity search is widely used in very different domains, spanning over many areas of interdisciplinary research. This includes multimedia databases, time series, medical, scientific, chemical and bioinformatic tasks, among others. As the similarity search problems have started to originate from more complex domains than before, the database community will have to take into consideration the non-metric case. In the following, we outline five challenges for future research in non-metric similarity indexing:

1) Scalability: Today, the common mean of non-metric searching is the sequential scan of the database. In the future, the lack of access methods could be a bottleneck for the domain experts, as the database sizes tend to increase even in domains that used to be not data intensive (e.g., protein engineering). Thus, it is necessary to design indexing techniques specifically designed to tackle with non-metric spaces.

2) Indexability: We have mentioned the indexability concepts related to metric spaces. Although a non-metric problem may be turned into a metric one and then use the metric indexability concepts, this might not be the best solution of the non-metric space analysis. One way to better analyze a non-metric space could be to combine multiple indexability concepts, each of them applicable to different mappings of the original problem (e.g., to metric/ptolemaic space or fuzzy logic).

3) Implementation Specificity: One approach to achieve efficient and scalable non-metric search is the design of made-to-measure access methods, where each method is designed specifically for a particular similarity function. The inverted file and the cosine measure are examples of this approach. While specific access methods cannot be reused for other similarity functions, it might turn out that this is the only solution within the restrictions given by the non-metric similarity search approach.

4) Efficiency vs. Effectiveness: Given the characteristics of searching in non-metric spaces, this may imply focusing the research on approximate and probabilistic techniques. These techniques, at least in the case of metric spaces, have been shown to give a good trade-off between efficiency and effectiveness regarding to similarity search.

5) Extensibility: Because of the simple assumptions on the syntax of pairwise similarity function, the mechanisms of similarity search might be applied also in contexts where the (dis)similarity function is interchangeable with another "syntactically compatible" aggregating/scoring function.

ACKNOWLEDGMENT

This research has been supported in part by Czech Science Foundation project 201/09/0683.

REFERENCES

[1] T. Skopal and B. Bustos, "On nonmetric similarity search problems in complex domains," *ACM Computing Surveys*, to appear.

[2] P. Zezula, G. Amato, V. Dohnal, and M. Batko, *Similarity Search: The Metric Space Approach (Advances in Database Systems)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[3] H. Samet, *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann, 2006.

[4] M. A. Khamsi and W. A. Kirk, Eds., *An Introduction to Metric Spaces and Fixed Point Theory*. Wiley-Interscience, 2001.

[5] P. Corazza, "Introduction to metric-preserving functions," *American Mathematical Monthly*, vol. 104, no. 4, pp. 309–23, 1999.

[6] C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in *Proc. 8th International Conference on Database Theory (ICDT'01)*. London, UK: Springer-Verlag, 2001.

[7] K.-S. Goh, B. Li, and E. Chang, "DynDex: A dynamic and non-metric space indexer," in *Proc. 10th ACM International Conference on Multimedia (MM'01)*. ACM Press, 2002, pp. 466–475.

[8] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[9] Y. Rubner, J. Puzicha, C. Tomasi, and J. Buhmann, "Emipirical evaluation of dissimilarity measures for color and texture," *Computer Vision and Image Understanding*, vol. 84, no. 1, pp. 25–43, 2001.

[10] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. AAAI Workshop On Knowledge Discovery in Databases*, 1994, pp. 229–248.

[11] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[12] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[13] I. Bartolini, P. Ciaccia, and M. Patella, "WARP: Accurate retrieval of shapes using phase of fourier descriptors and time warping distance," *IEEE Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 142–147, 2005.

[14] B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranić, "Feature-based similarity search in 3D object databases," *ACM Computing Surveys*, vol. 37, no. 4, pp. 345–387, 2005.

[15] J. Pu, Y. Kalyanaraman, S. Jayanti, K. Ramani, and Z. Pizlo, "Navigation and discovery in 3D cad repositories," *IEEE Computer Graphics and Applications*, vol. 27, no. 4, pp. 38–47, 2007.

[16] B. Bustos, D. Keim, D. Saupe, and T. Schreck, "Content-based 3D object retrieval," *IEEE Computer Graphics and Applications*, vol. 27, no. 4, pp. 22–27, 2007.

[17] E. Pampalk, "Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns. implementation submitted to the 3rd Annual Music Information Retrieval eXchange (MIREX'06)." in *Proc. International Symposium on Music Information Retrieval*, 2006.

[18] C. A. Ratanamahatana and P. Tohlong, "Speech audio retrieval using voice query," in *Proc. 9th International Conference on Asian Digital Libraries (ICADL'06)*, ser. LNCS 4312. Springer-Verlag, 2006, pp. 494–497.

[19] I. S. H. Suyoto, A. L. Uitdenbogerd, and F. Scholer, "Effective retrieval of polyphonic audio with polyphonic symbolic queries," in *Proc. International Workshop on Multimedia Information Retrieval (MIR'07)*. New York, NY, USA: ACM, 2007, pp. 105–114.

[20] C. Parker, A. Fern, and P. Tadepalli, "Learning for efficient retrieval of structured data with noisy queries," in *Proc. 24th International Conference on Machine Learning (ICML'07)*. ACM Press, 2007, pp. 729–736.

[21] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)," *IEEE Transactions on Dependable and Secure Computing*, vol. 3, no. 4, pp. 301–311, 2006.

[22] I. Sanz, M. Mesiti, G. Guerrini, and R. Berlanga, "Fragment-based approximate retrieval in highly heterogeneous xml collections," *Data & Knowledge Engineering*, vol. 64, no. 1, pp. 266–293, 2008.

[23] W. Tsang, A. Corboy, K. Lee, D. Raicu, and J. Furst, "Texture-based image retrieval for computerized tomography databases," in *Proc. 18th IEEE Symposium on Computer-based Medical Systems (CBMS'05)*. IEEE Computer Society, 2005, pp. 593–598.

[24] G. Schaefer, S. Zhu, and S. Ruszala, "Visualization of medical infrared image databases," in *Proc. 27th IEEE Annual Conference on Engineering in Medicine and Biology*. IEEE, 2005, pp. 634–637.

[25] S. Saha and S. Bandyopadhyay, "MRI brain image segmentation by fuzzy symmetry based genetic clustering technique," in *Proc. IEEE Congress on Evolutionary Computation (CEC'07)*. IEEE, 2007, pp. 4417–4424.

[26] L. Chen, M. T. zsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *PROC. ACM International Conference on Managment of Data (SIGMOD'05)*. ACM, 2005, pp. 491–502.

[27] V. Tuzcu and S. Nas, "Dynamic time warping as a novel tool in pattern recognition of ECG changes in heart rhythm disturbances," in *IEEE International Conference on Systems, Man and Cybernetics*. IEEE Computer Society, 2005, pp. 182–186.

[28] M. Vlachos, G. Kollios, and D. Gunopulos, "Elastic translation invariant matching of trajectories," *Machine Learning*, vol. 58, no. 2–3, pp. 301–334, 2005.

[29] X. Zuo and X. Jin, "General hierarchical model (ghm) to measure similarity of time series," *SIGMOD Record*, vol. 36, no. 1, pp. 13–18, 2007.

[30] X. Yan, P. Yu, and J. Han, "Substructure similarity search in graph databases," in *Proc. ACM International Conference on Management of Data (SIGMOD'05)*. ACM, 2005, pp. 766–777.

[31] H. He and A. Singh, "Closure-tree: An index structure for graph queries," in *Proc. 22nd International Conference on Data Engineering (ICDE'06)*. IEEE Computer Society, 2006, p. 38.

[32] Y. Ke, J. Cheng, and W. Ng, "Efficient correlation search from graph databases," *IEEE Transactions on Data Knowledge and Engineering*, vol. 20, no. 12, pp. 1601–1615, 2008.

[33] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model for evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345–352, 1978.

[34] S. Henikoff and J. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10 915–10 919, 1992.

[35] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[36] P. Lackner, W. A. Koppensteiner, M. J. Sippl, and F. S. Domingues, "Prosup: a refined tool for protein structure alignment," *Protein Engineering*, vol. 13, no. 11, pp. 745–752, November 2000.

[37] A. R. Ortiz, C. E. Strauss, and O. Olmea, "Mammoth (matching molecular models obtained from theory): An automated method for model comparison," *Protein Science*, vol. 11, no. 11, pp. 2606–2621, November 2002.

[38] D. D. Robinson, P. D. Lyne, and W. G. Richards, "Partial molecular alignment via local structure analysis," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 2, pp. 503–512, 2000.

[39] N. Nikolova and J. Jaworska, "Approaches to measure chemical similarity: a review," *SAR & Combinatorial Science*, vol. 22, no. 10, pp. 1006–1026, 2003.

[40] S.-H. Cha, S. Yoon, and C. Tappert, "On binary similarity measures for handwritten character recognition," in *Proc. 8th International Conference on Document Analysis and Recognition (ICDAR'05)*. IEEE Computer Society, 2005, pp. 4–8.

[41] X. Lu and A. Jain, "Deformation modeling for robust 3d face matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1346–1357, Aug. 2008.

[42] E. Chávez, G. Navarro, R. Baeza-Yates, and J. Marroquín, "Searching in metric spaces," *ACM Computing Surveys*, vol. 33, no. 3, pp. 273–321, 2001.

[43] T. Skopal, "Unified framework for fast exact and approximate search in dissimilarity spaces," *ACM Transactions on Database Systems*, vol. 32, no. 4, pp. 1–46, 2007.

[44] C. Faloutsos and K.-I. Lin, "Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets," in *Proc. ACM International Conference on Management of Data (SIGMOD'95)*. New York, NY, USA: ACM, 1995, pp. 163–174.

[45] X. Wang, J. T. L. Wang, K. I. Lin, D. Shasha, B. A. Shapiro, and K. Zhang, "An index structure for data mining and clustering," *Knowledge Information Systems*, vol. 2, no. 2, pp. 161–184, 2000.

[46] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios, "Boostmap: A method for efficient approximate similarity rankings," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*. IEEE, 2004, pp. 486–493.

[47] R. N. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function I," *Psychometrika*, vol. 27, no. 2, pp. 125–140, 1962.

[48] V. Athitsos, M. Hadjieleftheriou, G. Kollios, and S. Sclaroff, "Query-sensitive embeddings," in *Proc. ACM International Conference on Management of Data (SIGMOD'05)*. New York, NY, USA: ACM Press, 2005, pp. 706–717.

[49] V. Roth, J. Laub, J. M. Buhmann, and K. R. Müller, "Going metric: Denoising pairwise data," in *Proc. International Conference on Neural Information Processing Systems (NIPS'02)*, 2002, pp. 817–824.

[50] P. Vojtáš and A. Eckhardt, "Using tuneable fuzzy similarity in non-metric search," in *Proc. 2nd International Workshop on Similarity Search and Applications (SISAP'09)*. IEEE, 2009, pp. 163–164.

[51] A. Eckhardt, T. Skopal, and P. Vojtáš, "On fuzzy vs. metric similarity search in complex databases," in *Proc. 8th Conference on Flexible Query Answering Systems (FQAS'09)*, ser. LNAI, vol. 5822. Springer, 2009, pp. 64–75.

[52] M. L. Hetland, "Ptolemaic indexing," `arXiv:0911.4384 [cs.DS]`, 2009.

[53] Y. Lifshits, "Combinatorial framework for similarity search," in *Proc. 2nd International Workshop on Similarity Search and Applications (SISAP'09)*. IEEE Computer Society, 2009, pp. 11–17.

[54] G. Becker and M. Potts, "Non-metric biometric clustering," in *Proc. Biometrics Symposium*, 2007, pp. 1–6.

[55] M. Ackermann, J. Blömer, and C. Sohler, "Clustering for metric and non-metric distance measures," in *Proc. 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'08)*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2008, pp. 799–808.

[56] D. Jacobs, D. Weinshall, and Y. Gdalyahu, "Classification with non-metric distances: Image retrieval and class representation," *IEEE Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 583–600, 2000.

[57] P. Ciaccia and M. Patella, "Searching in metric spaces with user-defined and approximate distances," *ACM Database Systems*, vol. 27, no. 4, pp. 398–437, 2002.

[58] T. Skopal and J. Lokoč, "NM-tree: Flexible approximate similarity search in metric and non-metric spaces," in *Proc. 19th International Conference on Database and Expert Systems Applications (DEXA'08)*, ser. LNCS 5181. Springer-Verlag, 2008, pp. 312–325.

[59] L. Chen and X. Lian, "Efficient similarity search in nonmetric spaces with local constant embedding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 3, pp. 321–336, 2008.

[60] M. Berry and M. Browne, *Understanding Search Engines, Mathematical Modeling and Text Retrieval*. Siam, 1999.

[61] C. Aggarwal and P. Yu, "The IGrid index: reversing the dimensionality curse for similarity indexing in high dimensional space," in *Proc. 6th ACM International Conference on Knowledge Discovery and Data Mining (KDD'00)*. New York, NY, USA: ACM Press, 2000, pp. 119–129.

[62] E. Keogh and C. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358–386, 2005.

[63] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multi-dimensional time-series with support for multiple distance measures," in *Proc. 9th ACM International Conference on Knowledge Discovery and Data Mining (KDD'03)*. New York, NY, USA: ACM Press, 2003, pp. 216–225.

[64] D. Lipman and W. Pearson, "Rapid and sensitive protein similarity searches," *Science*, vol. 227, pp. 1435–1441, Mar. 1985.

[65] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct 1990.