Beyond the Metric Space Model

Benjamin Bustos$^{\diamond}$ and Tomáš Skopal$^{\dagger *}$

$^{\diamond}$ Department of Computer Science, University of Chile. *bebustos@dcc.uchile.cl*

† Department of Software Engineering, FMP, Charles University in Prague. *skopal@ksi.mff.cuni.cz*

Introduction

The metric space model has represented a reasonable trade-off concerning the efficiency and effectiveness problem in similarity search. However, complex similarity models that do not satisfy the metric properties have been used in a wide variety of research domains like multimedia information retrieval, digital libraries, biological and chemical databases, time series analysis, and biometry [2]. All these domains require the management of very large data collections, but the algorithms and data structures for searching in metric spaces cannot be used directly, as they require to use nonmetric similarity measures. As the term *nonmetric* simply means that a similarity function does not satisfy some (or all) the properties of a metric, we restrict its definition to nonmetric similarity functions that are "context-free and static", that is, the similarity between two objects is constant regardless of the context (time, user, query, other objects in the collection, etc.).

Various psychological theories suggest that the metric axioms could substantially limit the expressive power of similarity functions [1]. In particular, reflexivity and non-negativity may be violated if different objects can be differently self-similar (e.g., the image of a leaf on a trunk can be viewed as positively self-dissimilar if one considers a similarity which measures the less similar parts of the objects, here the trunk and the leaf). Also, symmetry may be violated if a prototypical object could be less similar to an indistinct one than vice versa (e.g., a subset is included in its superset but not vice versa). However, the triangle inequality is the most questioned property. Some theories point out that similarity has not to be transitive, as shown by the well-known example that a man is similar to a centaur, the centaur is similar to a horse, but the man is completely dissimilar to the horse. Finally, another advantage of nonmetric measures is the increased freedom of similarity modeling – the designer of a similarity function (the *domain expert*) is not constrained by the metric postulates. Thus, the study of a nonmetric model for similarity search has become a very important research issue, both from the theoretical and the practical point of view.

Main issues and research problems

Nowadays, many research domains that use nonmetric functions for some kind of content-based similarity retrieval, like similarity queries (e.g., kNN), similarity joins, or classification, have

[*] This work has been supported by Czech Science Foundation grant GAČR 201/09/0683.

emerged. To illustrate that the need for efficient nonmetric similarity search is not of marginal importance but quite relevant, we present in the following several application domains employing the nonmetric model [2].

First, the area of *multimedia databases* was one of the first suitable environments for similarity search, while nonmetric functions for image retrieval have been used for a long time. For example, an empirical comparison of several functions (including nonmetric ones like χ^2 distance, Kullblack-Leibler divergence, and Jeffrey-divergence) was made for color and texture attributes in images, showing the superiority of nonmetric functions in image classification. Furthermore, fractional L_p distances ($p < 1$) have been suggested for robust image matching and retrieval. The nonmetric similarities found their assets also in the retrieval of video (probability-based edit distance), shapes (partial Hausdorff distance, dynamic time warping distance), audio (Kullback-Leibler divergence, nonmetric variant of the Earth mover's distance), web pages (nonmetric variant of the Earth mover's distance), XML (weighted tag similarity, level-based similarity), etc. Second, *scientific and medical databases* usually consist of various types of signals (1D, 2D, or even 3D) produced by measuring some physical phenomena. Because of the inherent noise in the signal, the employed similarity functions are often designed as robust to noise, and/or as locally sensitive, becoming thus nonmetric. In particular, there were several nonmetric functions used in medical image retrieval, like the minimum integral for comparing spine X-ray images, or the Jeffrey-divergence for comparing tomography images. A huge number of applications was proposed for time series (ECG, seismological signals, material engineering, sensor networks), where nonmetric functions like dynamic time warping distance or longest common subsequence were proved as effective. Last, in *computational biology* the usage of nonmetric similarity functions is omnipresent. Because most retrieval tasks (protein classification, prediction, etc.) need partial matching (local alignment), the metric functions, performing mostly global matching, cannot be successfully utilized. Instead, the local alignment techniques have been employed as similarities for proteins, like the Smith-Waterman algorithm, BLAST (sequences), or SAP, ProSup, STRUCTAL (structures). A number of nonmetric partial-matching techniques have been used also in general chemical retrieval.

Research problems

As nonmetric similarities lack the metric axioms, there is generally no a priori available information that could be used for indexing and efficient search. There are two main approaches to tackle this problem. First, the domain expert has to supply some other a priori knowledge about the similarity function that could be used to index a database, for example an alternative topological axiom. Second, given a black-box similarity, one could discover the properties from the analysis of the distance matrix computed on a fraction of the database. In the following overview, all of the state-of-the-art approaches fit one of the mentioned alternatives.

Key existing projects in the area

As we already motivated in the introduction, many application domains use complex nonmetric (dis)similarity functions for comparing objects. However, only a few distance-specific and general nonmetric index structures have been proposed so far [2]. We believe that this is a huge motivation to continue developing algorithms and data structures for performing efficient similarity search in complex domains, as applications from very different domains would benefit from advances in this area. In the following, we briefly summarize the existing approaches to efficient nonmetric similarity search, wrapped into a concise framework.

The state of the art

Figure 1 presents a framework of the state-of-the-art techniques for efficient nonmetric similarity search. A domain expert defines a similarity search problem, the techniques that transform some complex raw data into the descriptors of the universe, and the domain-specific, nonmetric similarity function. We assume that the data collection is large enough and/or the similarity function is computationally expensive enough so that one wants to avoid the brute force solution (sequential scan).

The framework shows that the domain expert has several options depending on certain characteristic of the particular problem. On the one hand, if the problem uses a specific nonmetric distance function, the expert may use an efficient specific index in case it is available. A specific index is usually more efficient than a general one, so this is the best case for the overall performance (i.e., best efficiency and effectiveness). On the other hand, if the problem is modeled as black-box similarity or there is no specific index for the used nonmetric distance, there are two main options. First, one could use mapping of the problem into another space/paradigm. This may imply losing discriminative power or effectiveness of the (dis)similarity function. Second, one may use some of the few general nonmetric index structures/algorithms. The advantage of this last option is that it provides an all-in-one solution requiring none or little parameterization. Its disadvantage is that it will provide only approximate search, thus decreasing the retrieval effectiveness.

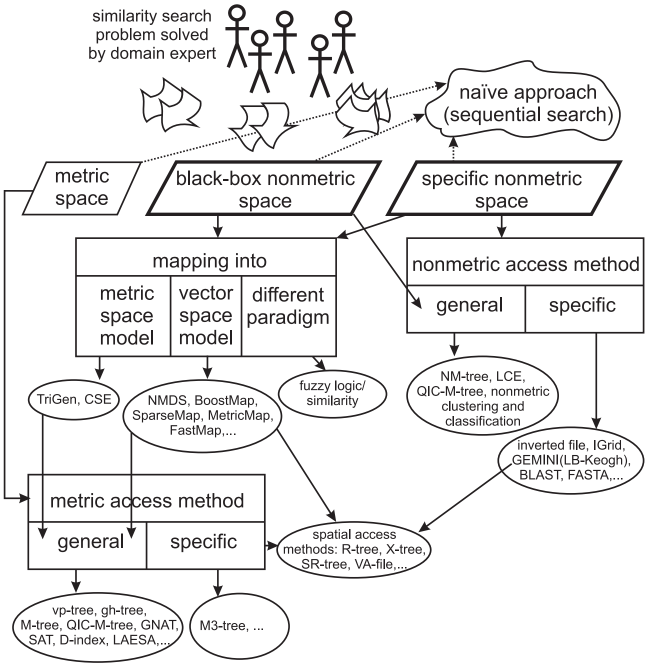

Figure 1: Framework of nonmetric similarity search

In particular, a nonmetric search problem can be mapped into a simpler one, for example, into the metric search problem. For instance, the TriGen algorithm adds a user-defined amount of triangle inequality into any nonmetric, which becomes partially or fully metric. Then, an existing metric access method (metric index) can be used for the indexing and search. The cost of this approach is higher intrinsic dimensionality of the target space and also more or less approximate search. Furthermore, the constant shifting embedding, the embeddings into various vector spaces (under L_p distances), and the fuzzy similarity approaches represent alternative techniques based on mapping the problem into a simpler one.

Additionally, several distance-specific indexes have been proposed, assuming specific properties of the similarity function, or even their precise definition. For example, the inverted file is here a perfect example, since it is an efficient structure for searching using the well-known cosine measure. Another example is the IGrid, that adapts the inverted file to work efficiently with a kind of nonmetric L_p distances. Some domain-specific indexing schemes were proposed also for the dynamic time warping distance, for the longest common subsequence, and for a number of biological similarity functions. Finally, there have emerged methods for black-box nonmetric similarity search, based on the use of some kind of mapping. In particular, the NM-tree combines the M-tree (a metric index) and the TriGen algorithm; the QIC-M-tree requires a metric lower-bound distance to the query nonmetric distance; the local constant embedding applies a kind of partitioning of the database and the constant shifting embedding. Last but not least, there have been introduced several nonmetric classification and clustering techniques applicable for approximate similarity search (nearest neighbor search, respectively).

Challenges to the future

This section presents some open questions and research problems associated to the development of a nonmetric space model for similarity search:

Alternative topological properties: Identifying alternative topological properties easily implementable into many similarities, but less restrictive than the metric axioms, could be very useful for indexing purposes.

Difficulty of the problem: Similarly to the metric space model, an open question is how one could measure how intrinsically difficult is to search in a given nonmetric space.

Efficiency vs. effectiveness: Apart from very few exceptions, almost all nonmetric index structures proposed so far return an approximated result for each query. Thus, the study of the "efficiency vs. effectiveness" trade-off in nonmetric spaces is a very relevant research issue.

Interdisciplinary collaboration: Fostering the interaction between domain practitioners and the database community seems to be a key issue, in order to develop and promote the efficient similarity search techniques into the real-world applications.

References

[1] S. Santini and R. Jain. Similarity measures. *IEEE Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

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