RESEARCH AND DEVELOPMENT
ERCIM News No.43 - October 2000 [contents]

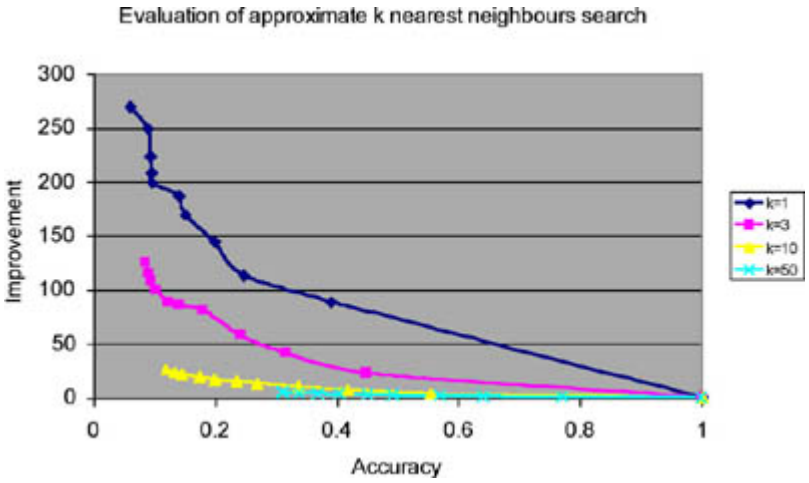
Approximate Similarity Search

by Giuseppe Amato

Similarity searching is fundamental in various application areas. Recently it has attracted much attention in the database community because of the growing need to deal with large volume of data. Consequently, efficiency has become a matter of concern in design. Although much has been done to develop structures able to perform fast similarity search, results are still not satisfactory, and more research is needed. The performance of similarity search for complex features deteriorates and does not scale well to very large object collections. Given the intrinsically interactive nature of the similarity-based search process, the efficient execution of elementary queries has become even more important, and the notion of approximate search has emerged as an important research issue.

Contrary to traditional databases, where simple attribute data are used, the standard approach to searching modern data repositories, such as multimedia databases, is to perform search on characteristic features that are extracted from information objects. Features are typically high dimensional vectors or some other data items, the pairs of which can only be compared by specific functions. In such search environments, exact match has little meaning; thus, concepts of similarity are typically applied. In order to illustrate this, let us consider an image data repository. It is clear that images are not atomic symbols, so equality is not a particularly realistic predicate. Instead, search tends to be based on similarity, because resemblance is more important than perfectly matching bit patterns. On the other hand, all that is similar is not necessarily relevant, so this paradigm tends to entail the retrieval of false positives that must be manually discarded. In other words, the paradigm rejects the idea that queries may be expressed in terms of necessary and sufficient conditions that will determine exactly which images we wish to retrieve. Instead, a query is more like an information filter, defined in terms of specific image properties, which reduce the user's task by providing only a small number of candidates to be examined. It is more important that candidates that are likely to be of interest are not excluded than it is that possibly irrelevant candidates be included.

We have investigated the problem of approximated similarity search for the range and nearest neighbour queries in the environment of generic metric spaces. From a formal point of view, the mathematical notion of metric space provides a useful abstraction of similarity or nearness. We modified existing tree-based similarity search structures to achieve approximate results at substantially lower costs. In our proposal, approximation is controlled by an external parameter of proximity of regions that allows avoiding access to data regions that possibly do not contain relevant objects. When the parameter is zero, precise results are guaranteed, and the higher the proximity threshold, the less accurate the results are and the faster the query is executed.



In order to have good quality results an accurate proximity measure is needed. The technique that we use to compute proximity between regions adopts a probabilistic approach: given two data regions, it is able to determine the probability that the intersection of these two regions contains relevant data objects. In fact, note that, even if two regions overlap, there is no guarantee that objects are contained in their intersection. Extensive experimental tests have shown a high reliability of this approach that gave a substantial contribution to the quality of the approximate results and to the efficiency of the approximate similarity search algorithm. We applied this idea for the similarity range and the nearest neighbours queries and verified its validity on real-life data sets. Improvements of two orders of magnitude were achieved for moderately approximated search results.

The main contributions of our approach can be summarised as follows:

• A unique approximation approach has been applied to the similarity range and the nearest neighbours queries in metric data files. Previous designs have only considered the nearest neighbours search, sometimes even restricted to one neighbour.

• The approximation level is parametric, and precise response to similarity queries is achieved by specifying zero proximity threshold.

• The approach to computation of proximity keeps the approximated results probabilistically bound. Experimental results demonstrate high precision and improvements of efficiency.

• We have experimentally demonstrated the importance of precise proximity measures, the application of which can lead to effective and efficient approximations of similarity search.

• Though implementation is demonstrated by extending the M-tree, the approach is also applicable to other similarity search trees at small implementation costs.

In the future, we plan to properly compare all existing approaches to approximation in uniform environment. We also hope to develop a system-user interface and apply the approach to real image and video archives. Finally, we intend to study the cases of iterative similarity search and complex approximated similarity search.

Other people that have contributed to this research are Pavel Zezula, Pasquale Savino, and Fausto Rabitti.

Please contact:
Giuseppe Amato - IEI-CNR
Tel: +39 050 315 2906
E-mail: G.Amato@iei.pi.cnr.it