Chapter 7

Conclusions

In this thesis, we have dealt with the issue of approximate similarity search in metric spaces. Even though there are several existing access methods that aim at increasing performance of similarity search, current technologies cannot be considered satisfactory and the issue of approximate similarity search has emerged as a relevant research topic. This approach offers higher performance at the price of imprecision in the results.

In its more general definition, similarity search only needs a similarity function to be defined in order to compare data. Metric spaces are thus very suitable as a framework for this problem. Contrary to other formalizations (for instance vector spaces), metric spaces do not pose any restriction on the possible representation of data. They only require that distance functions satisfy the metric postulates: symmetry, positiveness, reflexivity, and triangular inequality.

In the thesis we defined four new approaches for approximate similarity search in metric spaces. In addition, since one of the techniques is based on the measurement of the proximity of ball regions defined in metric spaces, we have also proposed some heuristics to compute this proximity efficiently and accurately.

7.1 Approximate similarity search in metric spaces

We have investigated techniques that relax the problem of similarity search allowing higher performance to be achieved at the price of some imprecision in the result sets. Our analysis shows that imprecision can be effectively controlled, while still guaranteeing very high performance compared to search algorithms on traditional access methods for similarity search. We conclude that approximate similarity search may represent a valid solution to the intrinsic inefficiency of the similarity search problem.

We have proposed and extensively tested four different techniques. All of them were implemented as search algorithms on M-Tree [CPZ97] access methods. However, given their generality, applying them to other tree search structures would be straightforward.

We have specified a number of measurements to assess the tradeoff between the speedup achieved and the quality of approximation. Experimental results on real-life data files are very promising, and efficiency improvement of two orders of magnitude has been achieved for acceptable approximations.

The performance and accuracy of our techniques, especially that of the strategies based on the exploitation of distance distribution (see Section 6.7) and proximity (see Section 6.9), are higher than those obtained by other approximate similarity search techniques proposed in the past (see Section 5.3). In addition our techniques are more flexible. In fact, they are defined for metric spaces, which also includes the case of vector spaces, and can be used both for nearest neighbors queries and range queries. As far as we know, there are no other techniques that can be applied at the same time to range and nearest neighbors queries. Most existing techniques can only be

applied to nearest neighbors search, and in some cases to a single nearest neighbor search.

7.2 Proximity of metric ball regions

We have also proposed techniques for the efficient and effective estimation of the proximity of metric ball regions. The approximate similarity search approach proposed, that gives the best performance relies on this measurement.

The proximity of two ball regions was formalized, using a probabilistic approach, as the probability that a random object belongs to the intersection of the two ball regions. Given the computational difficulty of exactly evaluating the proximity, some heuristics were defined that relax the original problem in such a way that a very efficient and accurate estimation was guaranteed.

In accordance with our objectives, the methods proposed are flexible and only require that the distance functions are metric. The effectiveness of the methods is high and depends only on the overall distance distribution. The computational complexity of the techniques proposed is linearly proportional to the granularity of distance distribution samples, thus it is also applicable at run-time. The storage overhead for maintaining the distance distribution functions needed are low. Since our methods only assume the generic metric postulates, our results are automatically valid for the important class of vector spaces.

We have also shown that the precision of proximity evaluation is determinant for the accuracy of the approximate similarity search algorithm. In fact, in terms of efficiency and accuracy, the performance of proximity based similarity search algorithms using a trivial measurement of proximity is much lower than using our approaches.

7.3 Future directions

Many issues in approximate similarity search remain to be investigated. We have proposed a general framework that can be used in all applications where data are represented in a metric space and we have applied some objective measures to assess the accuracy and efficiency of the approaches proposed. However, different applications may have different requirements in terms of efficiency and accuracy. For instance, in image database systems, image similarity is highly subjective and the image similarity search application has low criticality, therefore missing some images in the results set is not a big problem. On the other hand, other applications such as for instance, stock quotes analysis or DNA sequences retrieval, have well defined ways to measure similarity between objects, and the accuracy of the result set might have a high impact. Approaches that take into account application considerations need to be designed using ad-hoc techniques. Their performance should be tested by using application-driven measures.

Incremental approximation techniques should also be investigated. The techniques that we have developed cannot incrementally use results obtained in previous queries. After a query has been processed, if the user wants to refine it (i.e. approximate less), a new query must be executed using the new approximation parameters, since it is currently not possible to exploits results obtained in the previous sessions. The possibility of incrementally reusing previous results would be very important and useful in highly interactive systems.

It would also be interesting to investigate how approximate similarity search algorithms can be used when processing complex similarity queries involving combinations of several similarity predicates. This is the case for instance of systems that perform

multi-feature searches, where similarity with respect to several features is evaluated at the same time in the same query.

We have also defined techniques for efficient and effective estimation of the proximity between ball regions and we have applied it as a basis for an approximate similarity search algorithm. However, other applications can benefit from this measurement. As an example, we can consider node allocation strategies, used to decide what is the best way to allocate nodes on the disk in order to have higher search performance, or partition strategies, that can be used to decide what is the best way to partition entries of a node when a split has to be performed. In both cases, the probability that two regions will be accessed together during query execution can help to take an optimal decision. This probability can be inferred by using the proximity measurement. Investigating other applications where the measurement of the proximity can be exploited and refining the techniques for computing it accordingly would also be an interesting line for future research direction.