Chapter 1

Introduction

The performance capabilities of computer systems are rapidly increasing. Faster CPUs, larger secondary storage, and larger network bandwidth make it possible to handle more complex data types. Nowadays, people buy computers attracted by the promise of using audio, image, video, and 3D data. In professional and scientific areas such as medicine, computational biology, signal processing, or finance, new large data types in the form, for example, of time series, huge sequences of symbols, or raw streams of data are processed, analyzed and contrasted.

A very important issue on any kind of data management is searching. Traditional database systems efficiently search for structured records by using the *exact match* paradigm. However, the new data types, cannot be effectively represented as structured records. In these cases, the *similarity search* paradigm [JMM95] is used instead of exact match. Similarity searching consists in retrieving data that are similar to a given query. The measure of similarity is specifically defined with respect to the target application.

Techniques for improving performance of exact match search cannot be used for similarity search, and even if other alternative approaches were proposed, current

techniques are far from satisfactory [WSB98, BGRS99]. Therefore, the issue of *approximate similarity search* [AM95, AMN+98, CP00, FTAA00, FTAA01, IM98, PL99, PAL99, ZSAR98, Kle97, Cha97, KOR99, PMD01] has recently become an important research topic. Approximate similarity search provides an improvement in similarity search performance at the price of some imprecision in the results.

In this thesis, we have investigated techniques for approximate similarity search when data are represented in generic metric spaces. The techniques proposed offer an improvement of efficiency, with respect to exact similarity search techniques, up to two orders of magnitude, guaranteeing a high degree of accuracy.

In particular, one of the approximate similarity search techniques that is proposed relies on the measurement of the so-called *proximity of regions*. This measurement provides an estimation of the amount of objects shared by two regions of the space. In generic metric spaces, proximity of regions is not easily computed, given the lack of a coordinate system. We have also developed some techniques for computing the proximity of ball regions defined in metric spaces. The techniques proposed are efficient and their estimation of the proximity is extremely accurate.

1.1 Similarity search

Traditional database systems maintain files consisting of sequences of homogeneous records. In these databases the search paradigm used is the *exact match search* paradigm. Even more sophisticated searches, where a range of possible values is specified, such as a numerical range or a string prefix, rely on the possibility of exact comparison among attribute values. The exact match search divides the whole set of records into a subset that contains all records that satisfy the query, and another

subset that contains the records that do not satisfy the query.

On the other hand, a *similarity search* query orders the whole set of managed data with respect to the similarity of the individual items to the query. Therefore, ideally, the result set returned by a similarity search contains all items contained in the database along with the degree of similarity of each item to the query.

In order to limit the size of the returned result set, two different techniques are typically considered: cutting off items whose similarity to the query is below a specified threshold, or specifying the maximum number of items to be retrieved. Correspondingly, there are two basic types of similarity queries: *range queries* and *nearest neighbors queries*. A range query is composed of a query object and a similarity threshold. All data whose similarity to the query object is greater than the threshold satisfy the query. A nearest neighbors query, on the other hand, is composed of a query object and a value k indicating the maximum number of elements to be retrieved. The k elements most similar to the query satisfy the query.

1.1.1 Similarity and distance functions

At the basis of the similarity search process lies the ability of measuring the similarity between two items. This is obtained by defining a *similarity function*. There is not a general, universal definition of this function, since it depends on the specific application. Similarity functions are typically defined by application experts depending on the aspect that interests them.

Note that in some applications a *distance function* can be easily and more intuitively defined than a similarity function. However, it is easy to obtain a similarity function given a distance function and vice versa. In fact, similarity functions and

distance functions have an exactly opposite behaviour: the smaller the distances the higher the similarity.

The similarity function is typically considered as a black box provided by the application expert. General techniques developed for similarity search should not be tied to a particular similarity function. However, the similarity functions might be required to satisfy some specific properties.

In the following we briefly introduce two typical frameworks used to handle similarity search.

1.1.2 Vector spaces

Several similarity search applications represent data as high dimensional vectors. As an example consider image databases. Typically, the similarity between images is assessed by comparing features extracted by them, instead of using the raw images themselves. For instance, a typical image feature is a color histogram. This can easily be represented as a vector, where the value contained in each dimension is the density of the color associated with the dimension as a consequence of the quantization of the color space [Smi97].

When data are represented in vector spaces, similarity can be assessed by measuring the distance between two vectors, using a function of the Minkowski family, such as for instance, the Euclidean distance.

1.1.3 Metric spaces

In certain applications requiring similarity search either data cannot be effectively represented as vectors, or distance (similarity) cannot be effectively defined in terms

of functions of the Minkowski family. Consider again, for instance, color histogram similarity. Minkowski functions compare values in each dimension (colors) independently of the others. However it has been shown that color histogram similarity can be better assessed by using the *quadratic form* distance [FSA+95, HSE+95, NBE+93], which does not belong to the Minkowski family.

In such cases data are typically represented in *metric spaces*. The advantage of metric spaces is that there are no specific requirements on the representation of data, while the only constraint for the distance function is that it should comply to the metric postulates: it should satisfy the symmetric, positivity, identity, and triangular inequality properties.

Note that, when distances are measured by using Minkowski functions, vector spaces are in fact specific metric spaces. Therefore, techniques developed for generic metric spaces can also be applied to them. However metric spaces are generally more difficult to handle than vector spaces. This is due to the fact that vector spaces provide more information, such as geometric properties and coordinate systems.

Given their generality, in this thesis, we suppose that data are represented in a metric space; our results are thus widely applicable.

1.2 Approximate similarity search

Though the problem of similarity search seems to be very well defined and several storage structures have been proposed – see for example the survey in [CNBYM01] for metric spaces and the numerous methods for the vector spaces surveyed in [GG98] – current technology cannot certainly be considered stable. In fact, in some situations, performance of existing access methods is worse than a sequential scan of the whole

data set [BKK97, WSB98, BGRS99].

Given this inefficiency problem, the notion of *approximate similarity search* [AM95, AMN+98, CP00, FTAA00, FTAA01, IM98, PL99, PAL99, ZSAR98, Kle97, Cha97, KOR99, PMD01] has emerged as important research issue. The basic idea behind the approximate similarity search is that queries are processed faster, at the price of some imprecision in search results.

In general, approaches to approximate similarity search are motivated by the following three observations. First, typically a good data partitioning of metric data sets is simply not possible, so resulting data regions have typically a high overlap, and many regions must be accessed to answer a single query. Second, similarity-based search processes are intrinsically iterative. In many cases, users redefine queries depending on the results of previous searches. In such cases, an efficient execution of elementary queries is of particular importance and users easily accept some imprecision, especially in the initial and intermediate search results, if much faster responses can be achieved. Third, introducing some controlled imprecision in the result of a similarity search query may not be noticed by users or will be accepted when the increase in performance obtained is high. Note that similarity is an intuitive and subjective measurement. People, for instance, judge the similarity between two images differently. However, when similarity search algorithms are designed, the similarity measure must be defined using a rigorous mathematical formula and intuitiveness and subjectivity is lost.

Approaches to approximate similarity search can be classified into two broad categories [FTAA01]:

(1) approaches that reduce the size of data objects;

(2) approaches that reduce the data set that needs to be examined.

Approaches of the first category are mainly based on techniques of dimensionality reduction, and assume that a few dimensions are enough to represent the most relevant information of data items.

Two classes of approaches can be distinguished in the second category:

(a) approaches based on heuristics that stop the search algorithms before natural termination, when the current result set is judged to be satisfactory;

(b) approaches based on heuristics that do not access regions judged as not to containing promising results.

The advantage of the second category of approximate algorithms is that the original data representation is used and the heuristics may be tuned by using specific approximation parameters at query time: the higher the performance, the lower the accuracy of the approximation. Approaches belonging to the first category, on the other hand, modify the original data set and no tuning can be typically performed at query time.

1.3 Contribution of this thesis

This thesis proposes *four new techniques for approximate similarity search*, starting from the general case that data are represented in a metric space.

In addition we have developed some *techniques to measure the proximity of ball regions in generic metric spaces* efficiently and accurately. This measurement gives

the amount of data shared by two regions. One of the algorithms proposed exploits the measurement of the proximity of ball regions in its approximation strategy.

In the following we list the main issues that will be developed in the thesis.

1.3.1 Approximate similarity search

All the methods proposed belong to the category of approximation algorithms that reduce the data set that needs to be examined (category 1). Two of the proposed techniques are based on heuristics of early terminations (class a). The other two are based on heuristics for discarding unpromising regions (class b).

A brief description of these heuristics follows:

First method The first technique discards regions, containing data, guaranteeing that the maximum relative error on distances introduced, when regions containing qualifying data are discarded, is smaller than a threshold defined by the user. This method can be used both for range and nearest neighbors queries.

Second method The second technique of approximation stops the search algorithm when the current result set belongs to a user-defined percentage of the most similar data in the whole database. This method can be used for nearest neighbors queries only.

Third method Similarity search algorithms are based on an iterative process where a current result set is improved at every iteration. The third approximation method stops the search algorithm when the improvement of the current result set slows down. This method can be used for nearest neighbors queries only.

Fourth method The fourth approximation method discards regions when it is judged that they do not contain qualifying objects. This estimation is obtained by measuring the proximity of ball regions. This corresponds to the amount of data shared by two ball regions in a metric space. This method can be used both for range and nearest neighbors queries.

The results were extremely promising. The improvement of efficiency offered by these techniques reaches some orders of magnitude in correspondence with high accuracy of the retrieved result sets. Our techniques are more general than other existing techniques for approximate similarity search. In fact, they can be used in metric spaces and some can be used both for range and nearest neighbors queries. Most of the existing technique are, on the other hand, typically limited to vector spaces and to nearest neighbors queries. Their performance is generally worse than that of our techniques. As far as we know, there are no other techniques that can be used for both range and nearest neighbor queries.

1.3.2 Proximity of ball regions in metric spaces

It is easy to infer that a large overlap between regions does not always imply that a large amount of data are shared by them. The amount of data contained in the intersections depends on the data distribution. Therefore, there may be regions with a large intersection and few data in common, but also regions with a small intersection and many data in common. This happens when the intersection covers a dense area of the data space.

In this thesis, the measurement of the amount of data shared by two ball regions, called the *proximity of regions*, is analyzed and techniques for its quantification are

proposed. The proposed techniques are developed for general metric spaces, which naturally subsume the case of vector spaces. These techniques satisfy the following criteria: (1) the proximity is measured with sufficient precision; (2) the computational cost is low; (3) it can be applied to different metrics and data sets; (4) storage overhead is moderate.

Note that in a generic metric space it is not easy to obtain such measurements since we are not able, for instance, to compute volumes or areas. Therefore, the problem of the proximity of ball regions is analyzed using a probabilistic approach. After an analysis of the computational difficulties of the proximity measurement, heuristics to compute it in an effective and efficient way are proposed. An extensive validation was performed to prove the high accuracy of the proposed approaches. It was also shown that the accuracy of the proximity measurement has a high impact on the application involved. In fact, it was observed that, when the probabilistic approach was used, the performance of the approximate similarity search algorithm based on the proximity was much higher than the same algorithm when a trivial measurement of the proximity was adopted.

1.4 Outline of the thesis

Chapter 2 introduces the issues of similarity search. We discuss applications that need this approach instead of exact match search, discuss differences between handling similarity search using vector spaces or with generic metric spaces, and introduce similarity search queries. Finally, the data sets used for the tests performed to validate the approaches proposed in the thesis are described.

Chapter 3 describes the most important access methods for improving performance

of similarity search. In particular tree-based access methods were described in detail, as the results of this thesis are mainly applicable to them. These access methods were discussed giving a general definition of their structure and their similarity search algorithms. Specific techniques for one dimensional data, multi dimensional data, spatial data, and metric data are then presented.

Chapter 4 discusses the measurement of the proximity of ball regions in metric spaces. This measurement is needed by one of the approximate similarity search algorithms proposed later. Given the difficulty of computing this measurement efficiently in metric spaces, we propose some heuristics to estimate it efficiently and accurately. Tests performed to validate the techniques proposed are discussed and analyzed.

Chapter 5 deals with the issue of approximate similarity search. After an introduction to this technique, some of the most important existing approaches are described.

Chapter 6 proposes four new techniques for approximate similarity search in metric spaces. These techniques are defined by relaxing similarity search algorithms on tree-based access methods for metric data so that the trade-off between the accuracy of results and performance can be effectively controlled. The results obtained from extensive testing are discussed and analyzed.