

Approximate similarity search in metric spaces

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
der Universität Dortmund
am Fachbereich Informatik

von

Giuseppe Amato

Dortmund

2002

Tag der mündliche Prüfung: 14.06.2002

Dekan: i. V. Prof. Dr. Bernhard Steffen

Gutachter: Prof. Dr. Norbert Fuhr
Prof. Ing. Pavel Zezula
Prof. Dr. Joachim Biskup

To my wife Pina and my sons Niccolò and Giacomo

Table of Contents

Table of Contents	vii
Abstract	ix
Acknowledgements	xi
List of Symbols	xiii
List of Figures	xvii
1 Introduction	1
1.1 Similarity search	2
1.1.1 Similarity and distance functions	3
1.1.2 Vector spaces	4
1.1.3 Metric spaces	4
1.2 Approximate similarity search	5
1.3 Contribution of this thesis	7
1.3.1 Approximate similarity search	8
1.3.2 Proximity of ball regions in metric spaces	9
1.4 Outline of the thesis	10
2 Similarity search in metric spaces: overview and preliminaries	13
2.1 Introduction	13
2.2 Similarity search and its applications	15
2.3 From vector spaces to generic metric spaces	20
2.3.1 Metric spaces	24
2.3.2 Metric ball regions	25
2.3.3 Similarity and distance functions	25
2.3.4 Statistical information on data	31

2.4	Similarity queries	38
2.5	Data sets used in this thesis	40
3	Access methods for similarity search	43
3.1	Introduction	43
3.2	Access methods for similarity search	44
3.2.1	Sequential scan	45
3.2.2	Hashing	46
3.3	Tree-based access methods and similarity search algorithms	49
3.3.1	General structure	49
3.3.2	General similarity search algorithms	52
3.4	Specific tree-based access methods	57
3.4.1	One dimensional access methods: B-Trees	58
3.4.2	Point access methods: k-d-Trees and quad-Trees	60
3.4.3	Spatial access methods: R-Trees	64
3.4.4	Metric access methods: M-Trees	67
4	Proximity of ball regions in metric spaces	71
4.1	Introduction	71
4.2	Formal definition of proximity	72
4.3	Application considerations	74
4.4	Computational issues	77
4.5	Heuristics for an accurate measurement of the proximity	79
4.5.1	Definition of the heuristics	80
4.5.2	Computational complexity of the heuristics	91
4.6	Validating the approaches to the proximity measure	91
4.6.1	Experiments and comparison measures	92
4.6.2	Discussion on the experimental results	94
5	Approximate similarity search	103
5.1	Introduction	103
5.2	Approximate similarity search issues	104
5.3	Survey of existing approaches to approximate similarity search	106
5.3.1	First category: approaches able to reduce the size of data objects	106
5.3.2	Second category: approaches able to reduce the data set that needs to be examined	107
5.3.3	Approximate nearest neighbors searching using BBD trees	108
5.3.4	Approximate range searching using BBD trees	113
5.3.5	Approximate nearest neighbors searching using angle property	117

5.3.6	PAC nearest neighbor searching	122
6	Four new techniques for approximate similarity search in metric spaces	125
6.1	Introduction	125
6.2	Overview of the approaches	126
6.3	Generic approx. similarity search algorithms	130
6.3.1	Generic approximate range search algorithm	131
6.3.2	Generic approximate nearest neighbors search algorithm . . .	133
6.4	Efficiency and accuracy measures	134
6.4.1	Efficiency	135
6.4.2	Accuracy	136
6.5	Experimentation settings	140
6.6	Method 1: Approximate similarity search using relative error of distances	141
6.6.1	Results	146
6.7	Method 2: Approximate similarity search using distance distribution .	154
6.7.1	Results	158
6.8	Method 3: Approximate similarity search using the slowdown of distance improvement	162
6.8.1	Approximating the improvement of distances by a regression curve	166
6.8.2	Results	169
6.9	Method 4: Approximate similarity search using the region proximity .	174
6.9.1	Results	176
6.9.2	Further observations	185
6.10	Cross comparisons	187
6.10.1	Range queries	188
6.10.2	Nearest neighbors queries	191
6.10.3	Global considerations	193
6.11	Comparison with other techniques	194
7	Conclusions	199
7.1	Approximate similarity search in metric spaces	200
7.2	Proximity of metric ball regions	201
7.3	Future directions	202
	Bibliography	204

Abstract

There is an urgent need to improve the efficiency of similarity queries. For this reason, this thesis investigates approximate similarity search in the environment of metric spaces. Four different approximation techniques are proposed, each of which obtain high performance at the price of tolerable imprecision in the results. Measures are defined to quantify the improvement of performance obtained and the quality of approximations. The proposed techniques were tested on various synthetic and real-life files. The results of the experiments confirm the hypothesis that high quality approximate similarity search can be performed at a much lower cost than exact similarity search. The approaches that we propose provide an improvement of efficiency of up to two orders of magnitude, guaranteeing a good quality of the approximation.

The most promising of the proposed techniques exploits the measurement of the proximity of ball regions in metric spaces. The proximity of two ball regions is defined as the probability that data objects are contained in their intersection. This probability can be easily obtained in vector spaces but is very difficult to measure in generic metric spaces, where only distance distribution is available and data distribution cannot be used. Alternative techniques, which can be used to estimate such probability in metric spaces, are thus also proposed, discussed, and validated in the thesis.

Acknowledgements

It would have been very difficult to produce this thesis without the help and support of a number of people. In particular, I should like to express my gratitude to the following colleagues, friends and members of my family.

Above all, I am particularly grateful to Norbert Fuhr, who trusted in my research activity and encouraged me. He gave me the possibility of pursuing my PhD at his university and accepted to review my thesis.

Of course I must also thank, Costantino Thanos, the head of my research group, who offered me the opportunity to work on this thesis. I should like to acknowledge my appreciation to him for believing in me and encouraging me in the difficult world of scientific research.

Then, Pavel Zezula and Pasquale Savino for their invaluable contributions to the development of the research work described in this thesis. Their enthusiasm, clarity and above all willingness to listen and to provide suggestions have been of great importance to me. I must especially thank Pavel Zezula for also have accepted to be my reviewer.

Joachim Biskup for his patience in reading and reviewing this thesis, and his valuable suggestions. Piero Maestrini, the director of my research institute, for encouraging me in the development of this research activity.

I should like to thank a number of colleagues for their support. Fausto Rabitti and Claudio Gennaro for their willingness to discuss and exchange opinions on issues related to the topics of my thesis. Donatella Castelli and Paola Venerosi for their continuous support and encouragement. Carol Peters for her encouragement and precious suggestions regarding the text. Umberto Straccia for his patience and courage in sharing an office with me during this time. Stefano Chessa for his great friendship and openness. We took our first steps in the research world together and we have shared several important lessons in life.

My parents for always supporting and believing in me. My brother Sandro for teaching me to run rather than walk in order to achieve more. My sister-in-law Imma with whom I have had many important and motivating discussions on the meaning and difficulties of life.

And finally Pina for her limitless patience in assisting me continuously and sustaining me in the production of this thesis. During this period, she has not only helped me by creating a warm and happy home life but most important of all, she has given me two wonderful sons, Niccolò and Giacomo, thus filling our lives with joy.

List of Symbols

Symbol	Description
\mathcal{B}	Generic ball region.
$\mathcal{B}(O, r)$	Ball region with center O and radius r .
$cost(\mathbf{oper})$	Cost of executing search operation \mathbf{oper} .
d or $d(O_1, O_2)$	Distance function.
dim	Number of dimensions in a vector space.
$d_{it}^{O_q, k}(iter)$	Discrete function returning the distance of the current k -th object from the query object O_q at the iteration $iter$ of the k nearest neighbors search algorithm.
d_m	Maximum distance in the distance bounded metric space.
d_{xy}	Distance between objects O_x and O_y .
\mathcal{D}	Domain of the metric space.
\mathcal{DS}	Data set containing objects of the domain \mathcal{D} .
D_{XY}	Continuous random variable corresponding to the distance $d(\mathbf{O}_x, \mathbf{O}_y)$, with \mathbf{O}_x and \mathbf{O}_y random objects of \mathcal{D} .
EP	Error on the position, used to determine the accuracy of approximate nearest neighbors algorithms.

Symbol	Description
ϵ	Relative error on distances or upper bound of the relative error on distances.
$\epsilon(r_x, r_y, d_{xy})$	Absolute error of $X_{d_{xy}}^{appr}(r_x, r_y)$ with respect to $X_{d_{xy}}^{actual}(r_x, r_y)$.
$\epsilon'_\mu(d_{xy})$	Average of $\epsilon(r_x, r_y, d_{xy})$ varying r_x and r_y .
$\epsilon''_\mu(r_x, r_y)$	Average of $\epsilon(r_x, r_y, d_{xy})$ varying d_{xy} .
$\epsilon'_\sigma(d_{xy})$	Variance of $\epsilon(r_x, r_y, d_{xy})$ varying r_x and r_y .
$f(x)$	Overall distance density.
$f_O(x)$	Density of distances with respect to object O .
$f_X(x), f_Y(y)$	Density functions of continuous random variables X and Y .
$f_{XY}(x, y)$	Joint density function of continuous random variables X and Y .
$f_{XY D_{XY}}(x, y, d_{xy})$	Joint conditional density function of continuous random variables X and Y given D_{XY} .
$F(x)$	Overall distance distribution.
$F_O(x)$	Distribution of distances with respect to object O .
IE	Improvement of efficiency, used to determine the performance of approximate search algorithms.
k	Number of objects retrieved in a nearest neighbors query.
\mathcal{M}	Metric space. $\mathcal{M} = (\mathcal{D}, d)$, such that distance function d is a metric.
$\mathbf{nearest}(O_q, k)$	Set of objects returned by the nearest neighbors search algorithm.
$\mathbf{nearest}^{x_p, x_s}(O_q, k)$	Set of objects returned by the approximate nearest neighbors search algorithm with approximation parameters x_p and x_s .
N or N_i	node of a tree.

Symbol	Description
NE	Number of exact results, used to determine the accuracy of approximate range search algorithms.
$O, O_x, O_y, O_z, O_i, O_j$	Objects of the metric space or centers of ball regions.
O_q	Query object.
oper	Exact similarity search operation. It can be either range (O_q, r_q) or nearest (O_q, k).
oper ^A	Approximate version of oper .
p_i	Pointer to a record in an entry of a tree node.
Q, Q_1, Q_2, Q_3	Query regions.
r, r_x, r_y, r_i	Radii of ball regions.
range (O_q, r_q)	Set of objects returned by the range search algorithm.
range ^{x_p} (O_q, r_q)	Set of objects returned by the approximate range search algorithm with approximation parameter x_p .
$reg_d(iter)$	Continuous function that approximates $d_{it}^{O_q, k}(iter)$, obtained by using linear regression.
r_q	Radius of the query region.
$\mathcal{R}, \mathcal{R}_i$	Region.
x_s	Parameter for the approximate stop condition.
x_p	Parameter for the approximate pruning condition.
X	Continuous random variable corresponding to the distance $d(\mathbf{O}, \mathbf{O}_x)$, with \mathbf{O} and \mathbf{O}_x random objects of \mathcal{D} .

Symbol	Description
$X(\mathcal{B}(O_x, r_x), \mathcal{B}(O_y, r_y))$	Proximity of ball regions $\mathcal{B}(O_x, r_x)$ and $\mathcal{B}(O_y, r_y)$
$X_{d_{xy}}(r_x, r_y)$	Overall proximity of any pairs of regions having radii r_x and r_y , and whose distance between centers is d_{xy} .
$X_{d_{xy}}^{actual}(r_x, r_y)$	Overall proximity computed using the formal definition.
$X_{d_{xy}}^{appr}(r_x, r_y)$	Overall proximity computed using one of the proposed heuristics.
$X^{trivial}(\mathcal{B}(O_x, r_x), \mathcal{B}(O_y, r_y))$	Proximity of ball regions $\mathcal{B}(O_x, r_x)$ and $\mathcal{B}(O_y, r_y)$ computed using a trivial technique.
Y	Continuous random variable corresponding to the distance $d(\mathbf{O}, \mathbf{O}_y)$, with \mathbf{O} and \mathbf{O}_y random objects of \mathcal{D} .
$ exp $	Absolute value of expression exp .
$\ v\ $	Euclidean norm of vector v .
$\#S$	Cardinality of set S .

List of Figures

2.1	Shape of a ball region $\mathcal{B}(O, r)$, in a two dimensional vector space, when respectively L_1, L_2, L_6, L_∞ , are used as distance functions.	28
2.2	Density of data in a two dimensional vector space	36
2.3	Density of distances from the object O_i	37
2.4	Overall distance density functions of the data sets used for the experiments	41
3.1	B-Trees structure example, supposing that the maximum number of entries in a node is three	59
3.2	k-d-Trees structure example	61
3.3	Point quad-Trees structure example	63
3.4	R-Trees structure example, supposing that the maximum number of entries in a node is two	66
3.5	M-Trees structure example, supposing that the maximum number of entries in a node is two	68
4.1	Use of the proximity measure for region splitting (a), the allocation of objects on disks (b), and approximate similarity search (c)	75
4.2	Area bounded by the triangular inequality	80
4.3	Comparison between $f_{X,Y D_{XY}}(x, y d_{xy})$ and $f_{XY}(x, y)$	81
4.4	The four heuristics proposed to compute region proximity	84
4.5	Cases to be taken into account when defining bounding functions	88
4.6	Average and variance of errors given d_{xy} in HV1	95

4.7	Average and variance of errors given d_{xy} in HV2	96
4.8	Average and variance of errors given d_{xy} in UV	97
4.9	Comparison between the errors of the trivial method and the parallel method given r_x and r_y in HV1	98
4.10	Comparison between the errors of the trivial method and the parallel method given r_x and r_y in HV2	99
4.11	Comparison between the errors of the trivial method and the parallel method given r_x and r_y in UV	100
5.1	Partitions, data regions, and query regions	105
5.2	Possible regions in a BBD tree: a) dim -dimensional rectangle and b) set theoretic difference of two rectangles.	110
5.3	Overview of the approximate nearest neighbors search algorithm using BBD trees	112
5.4	Range queries using BBD tree: a) exact behaviour and b) approximate behaviour	116
5.5	Angle between objects contained in a ball region and a query object with respect to the center of the ball region	118
5.6	Objects belonging to children whose center is in the closest half of the parent node are more likely to contain nearest neighbors	120
5.7	If the query region does not intersect promising portions of the data region, this is discarded.	121
6.1	Comparison between the 10 nearest neighbors obtained by the precise and the proximity based approximate algorithms for two specific queries, using 0.01 as proximity threshold in the HV1 data set.	129
6.2	The relative distance error is not a reliable measure of the approximation accuracy. Even though the relative distance error is small, almost all objects are missed by the approximate search algorithm.	139

6.3	The region $\mathcal{B}(O_i, r_i)$, its parent region $\mathcal{B}(O_p, r_p)$, the query region $\mathcal{B}(O_q, r_q)$, and the reduced query region $\mathcal{B}(O_q, r_q/(1 + \epsilon))$	145
6.4	Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the fraction of exact results (NE). Range queries, HV1 data set.	147
6.5	Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the fraction of exact results (NE). Range queries, HV2 data set.	148
6.6	Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the fraction of exact results (NE). Range queries, UV data set.	149
6.7	Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the position error (EP). Nearest neighbor queries, HV1 data set.	150
6.8	Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the position error (EP). Nearest neighbor queries, HV2 data set.	151
6.9	Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the position error (EP). Nearest neighbor queries, UV data set.	152
6.10	An estimation of the fraction of the objects closest to O_q , whose distances from O_q are smaller than $d(O_q, O_c^k)$, can be obtained by using $F_{O_q}(x)$	156
6.11	Improvement of efficiency (IE) as a function of the derivative threshold (x_s) and the position error (EP). Nearest neighbor queries, HV1 data set.	159
6.12	Improvement of efficiency (IE) as a function of the derivative threshold (x_s) and the position error (EP). Nearest neighbor queries, HV2 data set.	160

6.13 Improvement of efficiency (IE) as a function of the derivative threshold (x_s) and the position error (EP). Nearest neighbor queries, UV data set.	161
6.14 Trend of $d_{it}^{O_q,k}(iter)$, when k is 3, in HV1.	163
6.15 Trend of $d_{it}^{O_q,k}(iter)$, when k is 3 in HV1, and two possible regression curves.	168
6.16 Improvement of efficiency (IE) as a function of the derivative threshold (x_s) and the position error (EP). Nearest neighbor queries, HV1 data set.	170
6.17 Improvement of efficiency (IE) as a function of the derivative threshold (x_s) and the position error (EP). Nearest neighbor queries, HV2 data set.	171
6.18 Improvement of efficiency (IE) as a function of the derivative threshold (x_s) and the position error (EP). Nearest neighbor queries, UV data set.	172
6.19 Overlap between the query region and data regions: not all data regions that overlap the query region share objects with it.	175
6.20 Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the fraction of exact results (NE). Range queries, HV1 data set.	177
6.21 Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the fraction of exact results (NE). Range queries, HV2 data set.	178
6.22 Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the fraction of exact results (NE). Range queries, UV data set.	179
6.23 Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the position error (EP). Nearest neighbor queries, HV1 data set.	180

6.24 Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the position error (EP). Nearest neighbor queries, HV2 data set.	181
6.25 Improvement of efficiency (IE) as a function of the proximity threshold (x_p) and the position error (EP). Nearest neighbor queries, UV data set.	182
6.26 Comparison of the trivial and probabilistic approximation techniques	186
6.27 Comparison of the approximation methods that support range queries in the various data sets.	189
6.28 Comparison of all approximation methods for nearest neighbor queries in the various data sets.	190
6.29 Average trend of the distance of the current k-th object from the query object during the exact nearest neighbor search execution in HV1 . . .	193

(page left intentionally blank)