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

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 **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 $\mathbf{range}(O_q, r_q)$ or $\mathbf{nearest}(O_q, k)$.
$\mathbf{oper}^A$	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.
$\mathbf{range}(O_q, r_q)$	Set of objects returned by the range search algorithm.
$\mathbf{range}^{x_p}(O_q, r_q)$	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**
$X(\mathcal{B}(O_x, r_x), \mathcal{B}(O_x, r_y))$	Proximity of ball regions $\mathcal{B}(O_x, r_x)$ and $\mathcal{B}(O_x, r_y)$
$X_{d_{xy}}(r_x, r_y)$	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.
$X^{trivial}(\mathcal{B}(O_x, r_x), \mathcal{B}(O_x, r_y))$	Proximity of ball regions $\mathcal{B}(O_x, r_x)$ and $\mathcal{B}(O_x, r_y)$ 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

(page left intentionally blank)