Faculty of Informatics
Masaryk University
Czech Republic

Developing Similarity Search Technology

Habilitation Thesis
(Collection of Articles)

Vlastislav Dohnal

February 2011

Abstract

Similarity searching has become more and more popular, which was stimulated by the growth of diverse data archives available on-line that offer search services to users, and by the increasing complexity of data that must be searched. This issue has also been recognized by major Internet search engines, exemplified by Google, that recently enriched their image search services by allowing users to search for images by similarity. They usually apply the following procedure. Firstly, a candidate set of images is obtained by a regular text search in images' file names and associated textual tags. Then this set is reordered by images' content, expressed as color histograms, for example. Finally, this result is presented to the user. In this thesis, we focus on similarity searching – content-based retrieval. In this area, data items are retrieved by their content rather than by textual information associated with them. For example, images are searched by comparing their color histograms to the histogram posed as a query by a user. The principle of ranking search results may also be applied to further increase the user's satisfaction with the search results.

The problem of similarity searching, as is studied in this thesis, uses metric space as a convenient data model. The author's contributions range from centralized index structures via distributed ones up to a new indexing paradigm – self-organizing systems. These results represent the efficiency issue of similarity searching. On the other hand, the effectiveness of similarity searching can be expressed as efforts to model computerized similarity as closest as possible to the human perception of similarity. This problem is also tackled in this thesis by introducing a new query type.

This thesis is conceived as a collection of articles. The collection contains 24 contributions published as either a monograph, chapters in monographs, or journal and conference papers. The author's contribution is mainly in stating research issues and supervising students during solving the issues, but also by running some of experimental trials, collecting necessary input data, and writing and editing significant parts of articles. Concrete contributions to articles published are specified at the corresponding articles listed in sections titled "Articles in Collection". In general, the author's contribution is quantified as the ratio $1/x$, where x is the number of authors of that particular article.

Abstrakt

Současný růst popularity podobnostního hledání je stimulován jak zvyšující se diverzitou dat zpřístupňovaných různými archivy, tak i rostoucí složitostí dat v nich obsažených. Tato tendence byla rozpoznána i vyhledávači, např. společností Google, která nedávno obohatila své vyhledávání v obrázcích o službu podobnostního hledání podle obsahu. Tento vyhledávač ale podobné obrázky vyhledává stále tradičním způsobem, tj. podle textu obsaženého v názvu souboru nebo popisků asociovaných s obrázkem. Přidaná služba pak přeuspořádá výsledky textového hledání podle podobnosti, např. podle barevného histogramu, a výsledek vrátí uživateli. V této práci se soustřed'ujeme na podobnostní hledání podle obsahu (angl. „content-based retrieval"). Takový druh hledání identifikuje data podle jejich faktického obsahu než podle textových informací, které jsou k nim přidruženy. Příkladem může být vyhledání všech obrázků, jejichž barevný histogram se podobá histogramu, který byl uživatelem předložen jako dotaz. Techniky pro přeuspořádání výsledků (angl. „ranking") lze také aplikovat a sice z důvodu zvýšení spokojenosti uživatele s předloženými výsledky hledání.

V této práci se zabýváme podobnostním hledáním, které jako vhodný datový model používá metrický prostor. Autorův přínos lze zařadit nejen mezi centralizované a distribuované datové struktury ale i do nové oblasti, ve které se aplikují principy samoorganizace. Zmíněné přínosy jsou soustředěny na problematiku výkonnosti, která je měřena např. jako čas zpracování dotazu. Podobnostní hledání má ale i druhou stránku a to problematiku definice podobnosti. V ideální situaci předpis definující podobnost pro účely vyhledávacího systému přesně odpovídá podobnosti chápanou člověkem. Docílení tohoto stavu, bohužel, bývá často velmi obtížné. Uvedená problematika je rovněž studována v této práci, i když ne tak obsáhle jako stránka výkonnosti.

Tato habilitační práce je koncipována jako soubor uveřejněných vědeckých prací (§72 odst. 3 písmena b zákona o vysokých školách). Soubor je tvořen dvacetičtyřmi publikacemi ve formě monografie, kapitol v monografiích, časopiseckých článků nebo konferenčních příspěvků. Autorův přínos lze shrnout jako identifikace výzkumných témat a vedení studentů při realizaci výzkumných projektů, ale také jako formulování myšlenek a analýzy výsledků experimentálních měření do podoby textových zpráv. Konkrétní přínosy autora jsou uvedeny u citací prací, které tvoří tento soubor. Tyto citace jsou vždy souhrnně uvedeny v sekcích nazvaných „Sbírka článků" (angl. „Articles in Collection"). Obecně je autorův přínos kvantifikován jako podíl $1/x$, kde x vyjadřuje počet autorů konkrétního příspěvku.

Acknowledgments

Firstly, I would like to thank Pavel Zezula, my former supervisor, for his guidance and fruitful ideas. I also thank my colleagues, former as well as current ones, namely Michal Batko, Stanislav Bartoň and David Novák. I thank all the students that participated in the research area and who did the real hard work, namely Jan Sedmidubský.

Secondly, I appreciate the cooperation with partners from other universities who co-authored some of the papers, mainly the work with Giuseppe Amato from ISTI-CNR, Pisa, Italy.

Lastly, my thanks go to my wife, children and parents for their endless patience, and to my best friends, namely Jiří Barnat for helping me with the thesis formatting.

In Brno, February 2011 Vlastislav Dohnal

Contents

Preface

This thesis is conceived as a collection of articles. The collection contains 24 contributions published as either a monograph, chapters in monographs, or journal and conference papers. The author's contribution is mainly in stating research issues and supervising students during solving the issues, but also by running some of experimental trials, collecting necessary input data, and writing and editing significant parts of articles.

The focus of the thesis is on similarity searching from the perspective of efficiency but also from the effectiveness point of view. It summarizes recent advances in this area and witnesses systems for similarity searching maturing from small centralized solutions via distributed structures to large systems exhibiting self-organizing properties. The contribution in terms of effectiveness is in enriching the variety of similarity queries. This all together forms Part I. For clarity, the author's contribution is emphasized through out this part by adding a check mark (✓) as a margin note, as is depicted next to this sentence. ✓

Part II contains full versions of articles published. They are categorized to monographs, monograph chapters, journal and conference papers.

Part I

Commentary

Chapter 1

Introduction

In the Information Society, information holds the master key to economic influence and success. But the usefulness of information depends critically upon its quality and the speed at which it can be transferred. The scale of the problem is emphasized by the recent growth of digital libraries, data warehouses and other Internet resources that have arisen as a direct consequence of latest advances in computing, communication and storage. These repositories organize data of various domains such as multimedia, scientific observations, molecular biology, computer-aided design or even marketing and purchasing analysis. Traditional retrieval techniques, typically based upon sorting routines and hash tables, are not appropriate for newly-emerging data domains listed above. More flexible methods must be found instead and they have to take into account the needs of particular users and application domains.

The problem of searching can also be observed from the data volume point of view. In particular, almost everything we see, hear, read, write or measure will soon be available to a computerized information system. This can be currently proved by citing the Flickr blog [1] [2] which reported five billionth photo was uploaded in September 2010. They also stated that 3,000 new photos are uploaded each minute.

Ordinary retrieval techniques are inadequate in many of these newer data domains because linear sorting is simply impossible. For illustration, consider a collection of bit patterns compared using the Hamming distance, i.e., the number of bits by which a given pair of patterns differs. There is no way to sort all the patterns linearly so that, selecting any arbitrary member, the objects similar to it will be closer in the ordering than the others. The same applies to the spectrum of colors. Obviously, we can sort colors according to their similarity with respect to a specific hue, for example pink. But we cannot sort the set of all colors in such a way that, for each hue, its immediate neighbor is the hue most similar to it.

This is what has given rise to a novel indexing paradigm based upon distance. From a formal standpoint, the search problem is modelled in metric space. The collection of objects to be searched forms a subset of the metric space domain, and the distance measure applied to pairs of objects is a metric distance function. This approach significantly extends the scope of traditional search approaches and supports execution of similarity queries. From a technical point of view, the search is not done

[1] Flickr - Photo Sharing, `http://www.flickr.com/`

[2] `http://blog.flickr.net/en/2010/09/19/5000000000/`

on original data objects, such as images or video, directly. But some characteristics (features, descriptors) are extracted from them instead. Then, a similarity search system organizes such characteristics and executes queries over them. It implies that the same characteristics must be obtained from queries before processing.

1.1 Challenges

✓ In this thesis, we seek the following challenges.

- **Dissemination** – by publishing articles about similarity searching at outstanding forums, we spread the knowledge about and technology used in similarity searching using metric spaces. Here, we would like to emphasis the first monograph studying the issue of metric similarity searching thoroughly and completely from basics to advances in existing solutions.
- **Efficiency** of similarity searching is an important issue to analyze due to enormous growth of data in volume. In this direction, we shift from the centralized paradigm to distributed one or even beyond to structures exhibiting self-organizing properties.
- **Effectiveness** of similarity searching is the other, orthogonal but closely related, issue. This term refers to the quality of similarity used, i.e., how big a discrepancy between human perception of similarity and the similarity modelled by a metric function is. We contributed in this area by proposing a new query type and an algorithm for solving the query inclusion problem.

The topics listed are presented in a concise way throughout this part. In Chapter 2, necessary definitions are introduced. Existing solutions for single computers are surveyed in Chapter 3. The efficiency problem is studied within Chapter 4 and Chapter 5. In Chapter 6, the third distinct contribution is sketched out. The thesis concludes in Chapter 7. Part II consists of articles referenced from Part I as author's contributions.

Chapter 2

Similarity Searching in a Nutshell

In contrast to traditional databases made up of simple attribute data, contemporary data is bulkier and more complex in nature. To deal with the increased bulk, data reduction techniques are employed as in [5]. These approaches typically result in high-dimensional vectors or other objects for which nothing beyond pairwise distances can be measured. Such data are sometimes designated *distance-only* data. A similar situation can occur with multimedia data. Here, the standard approach is to search not at the level of the actual multimedia objects, but rather using characteristic features extracted from these objects, e.g. color histogram obtained from an image. In such environments, an *exact match* has little meaning, and *proximity* concepts (*similarity*, *dissimilarity*) are typically much more fruitful for searching. Recent books or surveys that summarize the problem of similarity searching are available in [97, 117, 58, 31].

2.1 Metric Space

A useful abstraction for similarity is provided by the mathematical notion of *metric space* [69]. We consider the problem of organizing and searching large data sets from the perspective of *generic* or *arbitrary* metric spaces, sometimes conveniently labeled *distance spaces*. In general, the search problem can be described as follows:

Problem 2.1.1 *Let $\mathcal{D}$ be a domain, d a distance measure on $\mathcal{D}$, and $(\mathcal{D}, d)$ a metric space. Given a set $X \subseteq \mathcal{D}$ of n elements, preprocess or structure the data so that proximity queries are answered efficiently.*

From a practical point of view, X can be seen as a file (a data set or a collection) of objects that takes values from the domain $\mathcal{D}$, with d as the proximity measure, i.e., the distance function defined for an arbitrary pair of objects from $\mathcal{D}$. Though several types of similarity queries exist and others are expected to appear in the future, the basic types are known as the *similarity range* and the *nearest neighbor(s)* queries.

In a distance space, the only possible operation on data objects is the computation of a distance function on pairs of objects which satisfies the *triangle inequality*. In contrast, objects in a *coordinate space* – coordinate space being a special case of metric space – can be seen as vectors. Such spaces usually satisfy some additional properties that can be exploited in a storage and index structure design. Naturally, the distance between vectors can be computed, but each vector can also be uniquely located in

coordinate space. Further, vector representation allows us to perform operations like vector addition and subtraction. Thus, new vectors can be constructed from prior vectors. For more information, see [55, 16] for surveys of techniques that exploit the properties of coordinate space.

However, treating data collections as metric objects brings a great advantage in generality, because many data classes and information-seeking strategies conform to the metric view. Accordingly, a single metric indexing technique can be applied to many specific search problems quite different in nature. In this way, the important *extensibility* property of indexing structures is automatically satisfied. An indexing scheme that allows various forms of queries, or which can be modified to provide additional functionality, is of more value than an indexing scheme otherwise equivalent in power or even better in certain respects, but which cannot be extended.

2.2 Distance Measures

In the following, we present some examples of distance functions (metrics) used in practice on various types of data. Distance functions are often tailored to specific applications or a class of possible applications. In practice, distance functions are specified by domain experts, however, no distance function restricts the variety of queries that can be asked with this metric. Each distance function has to satisfy the following postulates:

$$\forall x, y \in \mathcal{D}, d(x, y) \geq 0 \qquad \text{non-negativity,}$$

$$\forall x, y \in \mathcal{D}, d(x, y) = d(y, x) \qquad \text{symmetry,}$$

$$\forall x, y \in \mathcal{D}, x = y \Leftrightarrow d(x, y) = 0 \qquad \text{identity,}$$

$$\forall x, y, z \in \mathcal{D}, d(x, z) \leq d(x, y) + d(y, z) \qquad \text{triangle inequality.}$$

The *Minkowski distance* functions form a whole family of metric functions, designated as the L_p metrics, because the individual cases depend on the numeric parameter p. These functions are defined on n-dimensional vectors of real numbers as:

$$L_p[(x_1, \ldots, x_n), (y_1, \ldots, y_n)] = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \,,$$

where the L_1 metric is known as the *Manhattan distance* (or the *City-Block distance*), the L_2 distance denotes the well-known Euclidean distance, and the $L_\infty = \max_{i=1}^{n} |x_i - y_i|$ is called the *maximum distance*, the *infinite distance* or the *chessboard distance*. The L_p metrics find use in a number of cases where numerical vectors have independent coordinates, e.g., in measurements of scientific experiments, environmental observations, or the study of different aspects of the business process.

Several applications using vector data have individual components, i.e., feature dimensions, correlated, so a kind of cross-talk exists between individual dimensions. Consider, for example, color histograms of images, where each dimension represents a specific color. To compute a distance, the red component, for example, must be

compared not only with the dimension representing the red color, but also with the pink and orange, because these colors are similar. The Euclidean distance L_2 does not reflect any correlation of features of color histograms. A distance model that has been successfully applied to image databases in [53], and that has the power to model dependencies between different components of features, is provided by the *quadratic form distance* functions in [57, 103]. In this approach, the distance measure of two n-dimensional vectors is based on an $n \times n$ positive semi-definite matrix $M = [m_{i,j}]$, where the *weights* $m_{i,j}$ denote how strong the connection between two components i and j of vectors $\vec{x}$ and $\vec{y}$ is, respectively. The following expression represents a generalized quadratic distance measure d_M, where the superscript T denotes vector transposition:

$$d_M(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \cdot M \cdot (\vec{x} - \vec{y})} \, .$$

The quadratic form distance measure may be computationally expensive, depending upon the dimensionality of the vectors. Color image histograms are typically high-dimensional vectors consisting of 64 or 256 distinct colors (vector dimensions).

The closeness of sequences of symbols (strings) can be effectively measured by the *edit distance*, also called the *Levenshtein distance* [71]. The distance between two strings $x = x_1 \cdots x_n$ and $y = y_1 \cdots y_m$ is defined as the minimum number of atomic edit operations (insert, delete, and replace) needed to transform string x into string y. The generalized edit distance function assigns weights (positive real numbers) to individual atomic operations. Hence, the distance between strings x and y is the minimum value of the sum of weighted atomic operations needed to transform x into y. If the weights of insert and delete operations differ, the edit distance is not symmetric and therefore not a metric function. However, the weight of the replace operation can differ. An excellent survey on string matching can be found in [83].

Let us now focus on a distance function to sets. Assuming two sets A and B, *Jaccard's coefficient* is defined as

$$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \, .$$

This distance function is simply based on the ratio between the cardinalities of intersection and union of the compared sets. An application of this metric to vector data is called the *Tanimoto similarity* measure [70], the distance version of which can be defined as:

$$d_{TS}(\vec{x}, \vec{y}) = 1 - \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\|^2 + \|\vec{y}\|^2 - \vec{x} \cdot \vec{y}} \, ,$$

where $\vec{x} \cdot \vec{y}$ is the scalar product of $\vec{x}$ and $\vec{y}$, and $\|\vec{x}\|$ is the Euclidean norm of $\vec{x}$.

2.3 Similarity Queries

A similarity query is defined explicitly or implicitly by a query object q and a constraint on the form and extent of proximity required, typically expressed as a distance. The response to a query returns all objects which satisfy the selection conditions.

The most common type of similarity query is the *range query* $R(q, r)$. The query is specified by a query object $q \in \mathcal{D}$ and a query radius r as the distance constraint. The

query retrieves all objects found within distance r of q, formally:

$$R(q, r) = \{o \in X, d(o, q) \leq r\}.$$

If the search radius is zero, the range query $R(q, 0)$ is called a *point query* or *exact match*.

An alternative way to search for similar objects is to use *nearest neighbor queries*. Specifically, $kNN(q)$ query retrieves the k nearest neighbors of the object q. If the collection to be searched consists of fewer than k objects, the query returns the whole database. Formally, the response set can be defined as follows:

$$kNN(q) = \{R \subseteq X, |R| = k \wedge \forall x \in R, y \in X - R : d(q, x) \leq d(q, y)\}.$$

When several objects lie at the same distance from the k-th nearest neighbor, the ties are solved arbitrarily.

A query that is defined only by a distance threshold is *similarity join*. It retrieves all pairs of objects (x, y) from the data sets X, Y, respectively, whose distance does not exceed the given distance threshold μ. The purpose of such a query can be in *data cleaning* or *data integration* when many Internet resource are federated. A nice survey over similarity joins is in [61].

Many other types similarity queries, as well as combination of basic types, can be defined. Some alternatives are surveyed in [62, 117].

2.4 Articles in Collection

The author has contributed to similarity searching by co-authoring the following survey articles. The most valuable one is the monograph on similarity searching, which has been the first publication dedicated and thoroughly studying the topic of similarity searching.

- Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity Search: The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer, 2006.

  Author's contribution: $1/4$, sections 2-7 of chapter 1, chapters 2 and 3, editorial supervision.

- Pavel Zezula, Michal Batko, and Vlastislav Dohnal. *Encyclopedia of Database Systems*, chapter Indexing Metric Spaces, pages 1–4. Database Management and Information Retrieval. Springer-Verlag, New York, 2009.

  Author's contribution: $1/3$, sections "Historical Background" and "Scientific Fundamentals".

- Pavel Zezula, Vlastislav Dohnal, and Michal Batko. *Encyclopedia of Computer Science and Engineering*, chapter File Organizations, pages 1219–1227. Wiley-Interscience, Hoboken, NJ, USA, 2009.

  Author's contribution: $1/3$, sections 2, 3 and 4 (excluding 4.3).

Chapter 3

Centralized Index Structures

Data management is important for any information system. In this chapter, we provide the reader with an overview of data organization principles used in metric spaces. Each of the principles is accompanied with a brief survey of main representatives of existing index structures. In overall, this chapter focuses on centralized solutions solely.

In metric spaces, data structures are based on either data partitioning principles or data transformation to a coordinate system or a combination of both. By the data transformation, we mean exploiting pre-computed distances. In the following sections, we tackle them separately.

3.1 Data Partitioning

Partitioning, in general, is one of the most fundamental principles of any storage structure, aiming at dividing the search space into sub-groups, so that once a query is given, only some of these groups are searched. A *ball partitioning* [110] uses a selected object, called a *pivot*, and a threshold on distance to establish a partitioning into two regions – the objects closer to the pivot than the threshold and the other objects. For illustration, see Figure 3.1. A *generalized hyperplane partitioning* [110] uses two pivots to split data objects to a set of objects closer to the first pivots and to a set of objects closer to the second pivots, i.e. two Voronoi-like cells are obtained. In [116], an *excluded middle partitioning* is proposed. It is based on the ball partitioning but adds another parameter on distance that defines a middle area which forms the third partition.

The ball partitioning is used in Vantage Point Tree [115] and its sequels [17, 38] including the Excluded-middle Vantage Point Forest [116]. Another family of ball-partitioning-based structures is formed by Burkhard-Keller Tree [23] and its improvement Fixed Queries Tree [3]. These structures are suited for distance functions producing low range of values, since their partitioning principle creates as many as subset as the number of distinct distance values. The concept of Fixed Queries Tree was further improved in [2, 28, 27].

The structures exploiting the generalized hyperplane partitioning are represented by Bisector Trees [66, 88, 87, 22], Voronoi Tree [42], and Generalized Hyperplane Tree (GHT) [110]. An example of GHT is depicted in Figure 3.2.

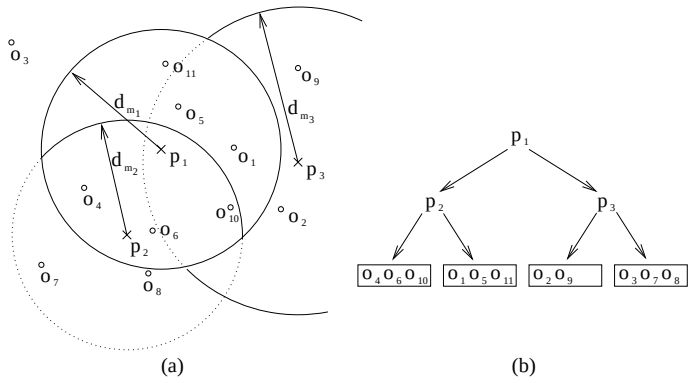

Figure 3.1: (a) Recursive ball partitioning of a metric space, (b) corresponding binary tree.

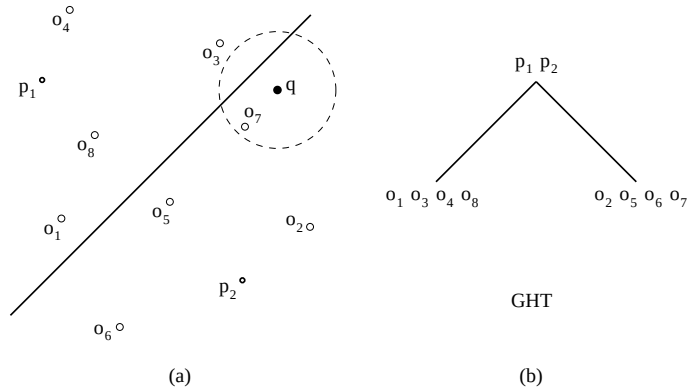

Figure 3.2: Generalized Hyperplane Tree (GHT): (a) a range query requiring access to both subsets of the hyperplane partition, (b) the corresponding structure of the tree.

3.2 Pre-computed Distances

When distance computations become expensive, a sound objective is to reduce their number to a minimum. To give efficient answers to similarity search queries, [104] have suggested using pre-computed distances between data objects. Having such distances, we can state formal rules that filter out objects irrelevant to the query.

The Approximating and Eliminating Search Algorithm (AESA) [111, 112] uses a complete matrix of pair-wise distances between all objects in the database. As an optimization, a half of the matrix lying below the diagonal need to be stored, since the distance function satisfies the symmetry property. Later, a linear version that selects only a few objects and stores only distance between the pivots and the objects, was proposed [78, 79, 77]. An alternative representation of pre-computed distance as a graph structure is used in Spaghettis [26]. Approaches that store pre-computed distances approximately and in the form of permutations are the Ordering Permutations [25] and PP-index [50]. Amato and Savino proposed a technique inspired by text retrieval and implements an inverted file for metric data [1].

The principle of storing and using pre-computed distances may be effective for data sets of small cardinality. But space requirements and search complexity become overwhelming for larger files.

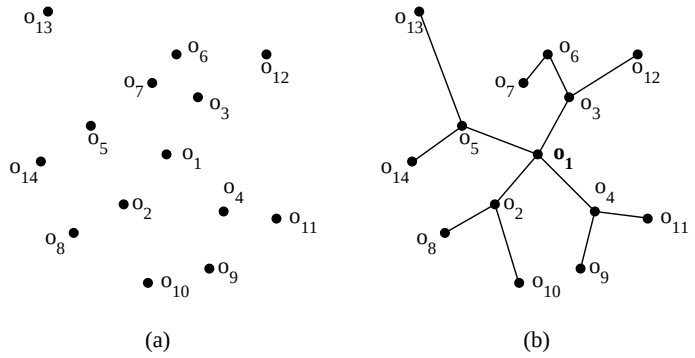

Figure 3.3: An example of SAT: (a) the data set, (b) SAT structure with the root o_1.

3.3 Combined Approaches

This is a very popular solution to index structures because it takes advantages of both the principles sketched above. Representatives are tree structures based on partitioning or clustering, and hashing structures. An evolution of Vantage Point Trees to a hybrid approach was proposed in [17, 18]. It is called the Multi Vantage Point Tree and stores several distances to pivots on the path from the root of the tree to a leaf for each object accommodated in that leaf node. Next, a generalization of the Generalized Hyperplane Tree that uses an unlimited number of pivots in data partitioning is called the Geometric Near-neighbor Access Tree [19].

The Spatial Approximation Tree (SAT) family is formed by a bunch of structures that approximate the Delaunay graph obtained by connecting neighboring Voronoi cells. The original SAT was proposed in [82, 84]. An illustration of its principle is given in Figure 3.3. Later dynamic versions and further improvements were introduced [85, 86, 6].

The List of Clusters [29] is a representative of a structure based on clustering. It creates a linked list of clusters. First, it creates a cluster from the whole data set. Then, the second cluster is identified in the remaining objects. The procedure is repeated until the data set is empty. The List of Clusters was later improved in [30, 76]. A variant of this concept but capable of evaluating similarity joins was proposed as the List of Twin Clusters [93].

A very popular and probably the most cited structure is Metric Tree (M-tree), proposed in [37]. It was designed as a dynamic structure that supports secondary memory for data storage. This structure clusters data objects hierarchically and forms a tree. It has a large variety of extensions and improvements: a bulk-loading algorithm [35], a slimming-down algorithm known as the Slim-tree [65], new insertion algorithms [108, 74], and the Pivoting M-tree [105]. It has also been recently extended to support non-metric functions and called NM-tree [107]. An M-tree suited for evaluation of queries over more metrics was proposed in [24].

A representative of locality hashing principles in metric space indexing is called the Distance Index (D-index) [43]. Its configuration algorithms were later extended in [4, 81]. Another representative is called the Metric Index (M-index), proposed in [90].

3.4 Articles in Collection

The author has contributed to similarity searching by co-authoring the following monograph, where the principles of constructing index structures are precisely detailed and defined and most of existing centralized index structures are surveyed.

- Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity Search: The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer, 2006.

  Author's contribution: $1/4$, sections 2-7 of chapter 1, chapters 2 and 3, editorial supervision.

Chapter 4

Distributed Index Structures

The growth rate of data collections in practically any domain of application of similarity searching now exceeds the growth rate of processor speeds. An example is an archive of nucleic acid sequences[1] which continues to grow at an exponential rate, doubling every 15 months. In February 2003, the database contained over 29 billion nucleotide bases in more than 23 million sequences [15]. Though metric access methods are able to speedup retrieval considerably, the processing time is not negligible and it grows linearly with the size of the searched collection. Thus the consequent degradation in performance of centralized access methods requires investigation of search methods that organize the database in a distributed manner. In this chapter, we present current approaches to distributed similarity search processing. This issue was studied in the co-authored survey papers [121, 44, 47, 48].

4.1 Centralized Control

For local area networks, a suitable distributed computation paradigm is to have a centralized directory that distributes the computations over the network, so it can obtain precise statistics about the overall load of the network as well as individual computer nodes. A representative designed for Grid infrastructures was proposed in [12] and is called M-Grid. This structure dedicates one of the computer nodes to a data directory, so this node routes incoming data and queries to particular nodes for processing. Since the directory is based on hashing, CPU costs are kept as low as possible. Thus this node may not become a bottleneck. Its advantage is simple design, so it can be easily implemented on modern architectures such as MapReduce [41] too.

4.2 Decentralized Control

Due to the centralized-control systems are prone to failures of the centralized node, researches put much more effort to decentralized solutions based on *scalable and distributed data structures* (SDDS) and *peer-to-peer network* (P2P) paradigms. SDDS was proposed in [73] for simple search keys like numbers and strings. Data objects are stored in a distributed file on network nodes. SDDS properties are *scalability* (data

[1] `http://www.ncbi.nlm.nih.gov`

migrate to new network nodes gracefully), *no hot-spot* (no master site that must be accessed for resolving queries), and *independence* (file accesses never require atomic updates on multiple nodes). P2P networks adds to SDDS new requirements that overcome problems of unreliability in the underlying network, namely: *fault tolerance* (failure of a network node is not fatal) and *redundancy* (data components are replicated on multiple nodes to increase availability).

Though, there are many existing approaches that eliminate the centralized control. Systems designed for handling metric data are twofold: mapping-based and native approaches. The mapping-based solutions select a set of pivots using which the metric space is transformed to a coordinate space. Then techniques for organizing high-dimensional vector data, such as CAN [94] or Chord [109], can be applied. The corresponding representatives are MCAN [52] and M-Chord [92]. On the other hand, the native approaches use the metric data objects directly without any transformation and applied data partitioning or clustering techniques. Distributed versions of Vantage Point Tree and Generalized Hyperplane Tree were proposed in [13, 14] under the names VPT* and GHT*. For a complete survey of distributed structures, please refer to [10, 89].

4.2.1 Splitting Nodes

Even distributed structures need a storage management technique on each node of the network to store data locally. In case any specific schema is not implemented within a distributed system, centralized indexes are applied. In [46], we study the issue of splitting centralized indexes. During populating the distributed system with new data objects, the system gracefully expands to new nodes, so it has to split the data stored on overflowing nodes. To optimize the process, we proposed and compare several techniques to split the M-tree.

4.3 Demonstration Applications

A prototype application of similarity searching in a collection of images is presented in [91]. It extracts MPEG-7 descriptors, namely color histogram, color structure, scalable color, shape and texture. The resulting descriptors form a single 280-dimensional vector for each of images. Then the vectors are organized in the M-Chord distributed indexing structure. A publicly available demonstration application[2] that indexes 100 million images from Flickr is described in [11]. Possibilities of using metric indexes in the area of biometrics are summarized in [119].

4.4 Articles in Collection

The articles listed below are four survey-like papers (two monograph chapters and two journal papers) and five conference papers (one of them is a demonstration paper).

[2] http://mufin.fi.muni.cz/imgsearch/

Surveys

- Pavel Zezula, Vlastislav Dohnal, and David Novák. *Global Data Management*, chapter Towards Scalability of Similarity Searching, pages 277–300. IOS Press, Amsterdam, The Netherlands, 2006.

  Author's contribution: $1/3$, sections 2 and 3, final editing.

- Vlastislav Dohnal, Claudio Gennaro, and Pavel Zezula. *Computational Intelligence in Medical Informatics*, chapter Efficiency and Scalability Issues in Metric Access Methods, pages 235–264. Springer-Verlag Berlin Heidelberg, Berlin, Germany, 1 edition, 2008.

  Author's contribution: $1/3$, section 2, first half section 3, and section 4.

- Vlastislav Dohnal and Pavel Zezula. Similarity searching in structured and unstructured p2p networks. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 400–416, 2009.

  Author's contribution: $1/2$, section 3, 4 and 5, final editing.

- Vlastislav Dohnal and Pavel Zezula. Real-life performance of metric searching. *SIGSPATIAL Special*, 2(2):28–31, 2010.

  Author's contribution: $1/2$, pages 2 and 3.

Conference papers

- Michal Batko, Vlastislav Dohnal, and Pavel Zezula. M-grid: Similarity searching in grids. In *Proceedings of International Workshop on Information Retrieval in Peer-to-Peer Networks, ACM CIKM 2006*, pages 17–24, Arlington, 2006. ACM Press.

  Author's contribution: $1/3$, main idea, prototyping, experimental evaluation.

- David Novák, Michal Batko, Vlastislav Dohnal, and Pavel Zezula. Scaling up the image content-based retrieval. In *Second DELOS Conference on Digital Libraries*, pages 1–10, Pisa, Italy, 2007. Information Society Technologies.

  Author's contribution: $1/4$, section 3 and 4 (excluding 3.3 and 4.2), partly prototyping.

- Vlastislav Dohnal, Jan Sedmidubský, Pavel Zezula, and David Novák. Similarity searching: Towards bulk-loading peer-to-peer networks. In *1st International Workshop on Similarity Search and Applications (SISAP 2008)*, pages 87–94, Los Alamitos CA, Washington, Tokyo, 2008. IEEE Computer Society.

  Author's contribution: $1/4$, idea of all splitting algorithms, major editing.

- Michal Batko, Vlastislav Dohnal, David Novák, and Jan Sedmidubský. Mufin: A multi-feature indexing network. In *Proceedings of the Second International Workshop on Similarity Search and Applications (SISAP 2009)*, pages 158–159, Washington, DC, USA, 2009. IEEE Computer Society.

Author's contribution: $1/4$, partly prototyping, article editorial revision.

- Pavel Zezula, Michal Batko, Vlastislav Dohnal, David Novák, and Jan Sedmidubský. Similarity search in large collections of biometric data. In *NATO RTO Modelling and Simulation Group Symposium*, pages 1–13, Brussels, 2009.

  Author's contribution: $1/5$, section 5 on biometrics.

Chapter 5

Self-organizing Search Systems

Distributed index structures presented in the previous chapter provide a good deal of scalability, so efficiency of a retrieval system is sufficient. Nonetheless, they can be barely applied to wide-area networks such as mobile networks, where the rate of joining or disconnecting nodes is very high. A promising approach is to apply principles of self-organizing which imply the following properties: *scalability* (with an exponential increase of nodes and their interactions, the system must be able to finish operations within acceptable bounds), *adaptability* (resources of the system change over time, so techniques to diminish degradation in system's performance mechanisms and to adapt to changes must be provided), and *robustness* (there is no single point of failure; nodes of the systems can automatically detect and recover from failures).

From the computer network's point of view, a self-organizing search system consists of many computer nodes interacting with each other to create a desired outcome – answering user queries. Thus, the structure of a self-organizing system often appears without an explicit pressure from outside the system, so the constraints on organization are internal to the system and result from interactions among the participating nodes. Moreover, there is no need for centralized control. Individual nodes communicate directly and exchange information locally, so there is no global view of the entire systems.

5.1 Metric Social Network

A *social network* is a term that is used in sociology since the 1950s and refers to a social structure of people, related either directly or indirectly to each other through a common relation or interest [113]. Using this notion, our approach places the peers of the distributed access structure in the role of people in the social network and creates relationships among them according to the similarity of the particular peer's data. The query processing then represents the search for the community of people – peers related by a common interest – similar data.

On the concepts of the social network we designed a data-oriented metric social network [98] that according to the terminology stated in [80] is as knowledge-cognitive network. The metric social network is used for similarity searching using metric spaces. As for the navigation, social networks exhibit the *small world network topology* [114] where most pairs of nodes are reachable by a short chain of interme-

diates – usually the average pairwise path length is bound by $\log n$. Therefore it is anticipated that a small amount – around six – of transitions is needed to find the community of peers holding the answer to a query posed at any of the participating peers in the network. The correspondence of this hypothesis to the network structure created by our system is analyzed in [7].

Further improvement of this concept is available in [99], where a brand new adaptive query routing algorithm was proposed. Its adaptability is built on top of confusability of queries stored in query history. Thus previous querying influence routing of new queries. Random factors were introduced to query routing algorithm in [100]. They help the system to answer new, previously unseen, queries satisfactorily. The system's recall to test queries is also analyzed. A prototype of Metric Social Network applied to an image database is presented in [8].

Next generation of query routing algorithm that is not restricted to limited values of radii when range queries are evaluated, is presented in [45]. It is based on a *confusability* of queries which is defined as a gradual function of two range queries. Algorithms implementing a joining procedure of newly connecting nodes are introduced in [102]. The new node selects so-call *bootstrap queries* to advertise its own data to the system. In [101], algorithms for estimating quality of a node's knowledge about the system are presented. Each node independently monitors its activity and based on changes in results of querying, it modifies the estimated quality of knowledge.

A general framework for implementing self-organizing search systems is introduced in [9], a monograph chapter. This general model give clear guidelines that advice developers implementing a self-organizing search system. Besides the model, a survey of existing systems based on self-organizing principles is presented.

5.2 Other Systems

In this section, we survey existing approaches related to self-organizing systems. The MRoute [56] is an unstructured P2P network for indexing and searching in multimedia data. The data is considered as objects of a metric space. MRoute transforms the objects to binary vectors using preselected pivots. Then, Routing Index [39], originally developed for organizing text documents, is applied to index these binary vectors. SIMPEER [49] introduces a super-peer architecture where the super-peers collect information about the content of peers connected to the particular super-peer. This approach uses a metric technique iDistance [63] for clustering data and creating a concise representation of a peer stored at a super-peer. However, the super-peers can be observed as a bottleneck of the whole system due to their centralized-control nature. Another work [72] presents preliminary results on Metric Overlay Networks (MONs) inspired by the idea of Semantic Overlay Networks (SONs [40]). The basic idea of SONs is that peers in an unstructured network are joined into semantically close groups. However, the scalability issue of MON is unclear. The authors used a dataset of 68,000 images. From our experience [98], such systems applied to small networks exhibit good results, but on a large scale the results deteriorate radically.

5.3 Articles in Collection

The articles listed below are conference papers and one monograph chapter (the last item).

- Stanislav Bartoň, Vlastislav Dohnal, Jan Sedmidubský, and Pavel Zezula. Gauging the evolution of metric social network. In *5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007) held at 33rd International Conference on Very Large Data Bases (VLDB 2007)*, pages 1–12, Vienna, 2007. VLDB Endowment.

  Author's contribution: 1/4, partly designing the algorithms and analyzing experimental results.

- Jan Sedmidubský, Stanislav Bartoň, Vlastislav Dohnal, and Pavel Zezula. Querying similarity in metric social networks. In *Network-Based Information Systems, First International Conference, NBiS 2007*, number vol. 4658 in Lecture Notes in Computer Science, page 278, Berlin, 2007. Springer.

  Author's contribution: 1/4, partly the idea of algorithms, result analyses, final editing.

- Jan Sedmidubský, Stanislav Bartoň, Vlastislav Dohnal, and Pavel Zezula. Adaptive approximate similarity searching through metric social networks. In *24th International Conference on Data Engineering (ICDE 2008)*, pages 1424–1426, Los Alamitos CA, 2008. IEEE Computer Society.

  Author's contribution: 1/4, definition of replaceability and confusability functions, writing a majority of the article.

- Jan Sedmidubský, Vlastislav Dohnal, Stanislav Bartoň, and Pavel Zezula. A self-organized system for content-based search in multimedia. In *IEEE International Symposium on Multimedia (ISM 2008)*, pages 322–327, Los Alamitos, California 90720-1314, 2008. IEEE Computer Society.

  Author's contribution: 1/4, definition of confusability function, section 3.

- Stanislav Bartoň, Vlastislav Dohnal, Jan Sedmidubský, and Pavel Zezula. Building self-organized image retrieval network. In *Proceeding of the 2008 ACM workshop on Large-Scale distributed systems for information retrieval (LSDS-IR'08)*, pages 51–58, USA, 2008. ACM.

  Author's contribution: 1/4, preparation and conversion of a data collection, section 2 and subsections 3.2, 4.2 and 4.3, final editing.

- Vlastislav Dohnal and Jan Sedmidubský. Query routing mechanisms in self-organizing search systems. In *2nd International Workshop on Similarity Search and Applications (SISAP 2009)*, pages 132–139, Los Alamitos, CA 90720-1314, 2009. IEEE Computer Society.

Author's contribution: $1/2$, specification of all confusability functions, preparation of experimental data, sections describing the algorithms, editing the article.

- Jan Sedmidubský, Vlastislav Dohnal, and Pavel Zezula. Feedback-based performance tuning for self-organizing multimedia retrieval systems. In *International Conference on Advances in Multimedia (MMEDIA 2010)*, pages 102–108, Los Alamitos, CA 90720-1314, 2010. IEEE Computer Society.

  Author's contribution: $1/3$, idea of confidence evaluation algorithm, writing some parts of the article and editing it.

- Jan Sedmidubský, Vlastislav Dohnal, and Pavel Zezula. On building a self-organizing search system for multimedia retrieval. In *International Workshop on Multimedia and Semantic Technologies (MUST 2010)*, Red Hook, NY 12571, USA, 2010.

  Author's contribution: $1/3$, idea of bootstrapping algorithms, preparation of data for experiments, writing and editing.

- Stanislav Bartoň, Vlastislav Dohnal, Jan Sedmidubský, and Pavel Zezula. *Computational Social Network Analysis: Trends, Tools and Research Advances*, chapter Towards Self-organizing Search Systems, pages 49–80. Computer Communications and Networks. Springer, New York, NY, USA, 2010.

  Author's contribution: $1/4$, partly designing the general model, major text writing and editing.

Chapter 6

Querying Effectiveness

Executing queries is the most fundamental operation and all index structures optimize it from the computation point of view. However, querying has also the effectiveness perspective which can be expressed as user's satisfaction. This has two dimensions: result quality and supported type of queries. The former one expresses whether the user likes or dislikes the result presented by the system, which is naturally mainly influenced by the similarity function used during searching. But a bad impression can be also caused by displaying only images identical to the query. So the expected variability of response is very limited if any. This issue is tackled in the following section. The latter dimension concerns the variety of supported queries, which is shortly analyzed in Section 6.2.

6.1 Result Quality

The effectiveness of a similarity-based retrieval system that uses a metric space model, is high related to the quality of the distance function used. Distance function is usually tuned by an expert in the domain of application. Thus it should somehow reflect user's expectations from the search. The discrepancy between the user-perceived and a system-defined similarity is called the *semantic gap*. There are techniques to decrease the semantic gap. One possibility is to allow the user to express his or her preferences by changing parameters of the distance function [32]. To ensure correctness of the result, a user-defined function should satisfy the *lower-bounding property* [36]. It states that the distance function used to index the data must return values not greater than the user-defined function. If this condition is not met, another function that lower-bounds both the user-defined and index functions must be defined. It is consequently used for querying. Another approach can be based on combining more distance functions into a single scoring function, which must be a distance function too. A retrieval system may then selectively employ more or fewer of the functions to fulfil the user's requirements. Feedback style of querying can be easily implemented in this way. For example, $\mathcal{A}_0$ algorithm [51] or some other more efficient solutions such as the threshold algorithm [20] can be applied for federating results obtained from more independent index structures, each built for that particular distance. A currently popular approach is based on ranking the result set, i.e., reordering the objects retrieved before presenting them to the user. The similarity searching is then enriched with a text-based

ranking, for example. Such a ranking can use tags associated with photos [21].

6.1.1 Eliminating Duplicates

Similarity searching used to suffer from low quality response to user queries since the databases contained a low number of objects – hundreds of thousands. This was caused by the sparsity of the data space, so the user that asked for 10 most similar objects, obtained something dissimilar from the query. Recently, we have witnessed the tremendous growth of digital data, mainly multimedia, available in Internet archives. So this new situation has lead to the increase of result quality, i.e. the user is almost in 100% cases served with similar objects and rarely he or she becomes unsatisfied by not finding anything relevant. From the data point of view, we can say that the data space became "denser", which has introduced a new and quite unexpected behavior. Some queries retrieve only almost identical objects, which is again unpleasant. Please remark that this issue is different from the term *semantic gap*.

To overcome the "dense space" problem, we proposed the *Distinct k-Nearest Neighbors query* (kDNN) in [106]. It builds on the classic nearest neighbor query, but, at the same time, it excludes all objects that are too similar to any of the already reported objects. Such an approach is quite robust with respect to the database size, while the robustness can be tuned by setting what still is and what already is not understood as "too similar". The kDNN(q) query can be defined as a subset R of the database X, such that $|R| = k$ and

$$\forall x, z \in \mathbb{R}, \forall y \in X - R : d(x, z) \geq \phi \wedge (d(x, q) \leq d(y, q) \vee \exists w \in R : d(y, w) < \phi),$$

where ϕ is a user-defined *separation distance*. The objects within the distance are considered as duplicates. Note that for $\phi = 0$ we get the classic kNN query.

6.2 Query Containment

There are many situations in which the searching for objects that are globally similar to the given query is not sufficient. This is especially true for image database where a user may require to search for images containing an object exposed in the query image. Thus, the retrieval system should solve the query containment problem. In the following, we focus on query containment issue in case of images. There are two main streams in research. Firstly, global image descriptors, such as color histogram, are used. Secondly, local image descriptors, such as SIFT or SURF, are applied. The local descriptors characterizes small parts of an image, thus many descriptors (even thousands) are usually extracted from an image, which is a major disadvantage of local descriptors when comparing to global ones, where only one descriptor is obtained.

In case of global descriptors, a solution proposed in [96] segments all database images as well as query images into chunks of pre-defined size and a global descriptor is extracted from each chunk. A searching procedure then finds the correspondence between the database images and the query image. For good retrieval results of subimages segments of an image must overlap significantly, so this technique must be tuned properly. For this reason, local descriptors are favorable.

In [68], the authors present a solution to storage implementation of the LSH-coded descriptors that allows searching for sub-images in linear time. However, this implementation does not handle the spatial position of features. Another retrieved information reduction [95] uses the properties of PCA-SIFT descriptors [67] directly. In particular, their hierarchical ordering and bit representation of each feature are used. This leads to the most-significant bit index files, which are memory-oriented structures storing bit prefixes of PCA-SIFT descriptors. Neither this technique indexes the spatial information. In [33], the geometric min-hash (GmH) algorithm is proposed. It is based on the original min-hash [34], but it incorporates the spatial context of features. From the searching point of view, it identifies regions in an image in which the identical or almost-identical groups of features with respect to the query image occur. However, only small spatial surrounding is considered. The method presented in [64] finds small logos in a natural image collection. SIFT features are reduced using the multi-probe locality sensitive hashing [75]. To determine the geometrically close features the RANSAC algorithm [54] is applied.

6.2.1 Proximity-based Order-respecting Intersection

Our approach to query containment in image database is based on a proximity-based order-respecting intersection, so-called ε-intersection. It is based on local image descriptors. Before querying the most important local descriptors are extracted from a query image. Then, a regular range query is executed for each of these descriptors. The queries with radius ε are evaluated on a database containing all local descriptors extracted from indexed images. After grouping the descriptors retrieved by original image identification, spatial distribution of descriptors is checked by projecting the descriptors to x and y axes independently. A special ORD function is used to rank the result. The formal definition of ε-intersection is available in [59] whereas a prototype implementation of this approach is presented in [60].

6.3 Articles in Collection

The following list of articles consists of one paper defining a new query type and two papers tackling the query containment problem – one of them is a demonstration paper.

- Tomáš Skopal, Vlastislav Dohnal, Michal Batko, and Pavel Zezula. Distinct nearest neighbors queries for similarity search in very large multimedia databases. In *11th ACM International Workshop on Web Information and Data Management (WIDM 2009)*, pages 11–14, New York, USA, 2009. ACM.

  Author's contribution: $1/4$, prototyping all algorithms, experiment planning and evaluation.

- Tomáš Homola, Vlastislav Dohnal, and Pavel Zezula. Proximity-based order-respecting intersection for searching in image databases. In *8th International Workshop on Adaptive Multimedia Retrieval, AMR'2010*, Linz, 2010.

Author's contribution: $1/3$, writing the article, editing.

- Tomáš Homola, Vlastislav Dohnal, and Pavel Zezula. Sub-image searching through intersection of local descriptors. In *3rd International Conference on Similarity Search and Applications (SISAP 2010)*, pages 127–128, New York, 2010. ACM Press.

  Author's contribution: $1/3$, partly prototyping on indexing structures, editing the article.

Chapter 7

Conclusions

In this thesis, we surveyed the author's contributions to the area of similarity searching when the metric space is applied as a convenient data model. The main contribution is a pioneering book on similarity searching which is currently cited more than 180 times. The other contributions consist of advances in distributed structured index networks and self-organizing solutions, and finding solutions to new query types.

The unifying idea of all contributions of this thesis is bringing similarity search toys to mature content-based information retrieval systems. This collection of articles consists of 24 articles including one monograph and five monograph chapters.

Bibliography

[1] Giuseppe Amato and Pasquale Savino. Approximate similarity search in metric spaces using inverted files. In *Proceedings of the 3rd International Conference on Scalable Information Systems (InfoScale 2008)*, pages 1–10, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[2] Ricardo A. Baeza-Yates. Searching: an algorithmic tour. In Allen Kent and James G. Williams, editors, *Encyclopedia of Computer Science and Technology*, pages 331–359. Marcel Dekker, Inc., 1997.

[3] Ricardo A. Baeza-Yates, Walter Cunto, Udi Manber, and Sun Wu. Proximity matching using fixed-queries trees. In Maxime Crochemore and Dan Gusfield, editors, *Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching (CPM 1994), Asilomar, California, USA, June 5-8, 1994*, Lecture Notes in Computer Science, pages 198–212. Springer, Berlin, 1994.

[4] Tao Ban. Using genetic algorithm to balance the d-index algorithm for metric search. In Masumi Ishikawa, Kenji Doya, Hiroyuki Miyamoto, and Takeshi Yamakawa, editors, *Neural Information Processing*, volume 4985 of *Lecture Notes in Computer Science*, pages 264–273. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-69162-4_28.

[5] Daniel Barbará, William DuMouchel, Christos Faloutsos, Peter J. Haas, Joseph M. Hellerstein, Yannis E. Ioannidis, H. V. Jagadish, Theodore Johnson, Raymond T. Ng, Viswanath Poosala, Kenneth A. Ross, and Kenneth C. Sevcik. The new jersey data reduction report. *IEEE Data Engineering Bulletin*, 20(4):3–45, 1997.

[6] Marcelo Barroso, Nora Reyes, and Rodrigo Paredes. Enlarging nodes to improve dynamic spatial approximation trees. In *Proceedings of the Third International Conference on SImilarity Search and APplications (SISAP 2010)*, pages 41–48, New York, NY, USA, 2010. ACM.

[7] Stanislav Bartoň, Vlastislav Dohnal, Jan Sedmidubský, and Pavel Zezula. Gauging the evolution of metric social network. In *5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007) held at 33rd International Conference on Very Large Data Bases (VLDB 2007)*, pages 1–12, Vienna, 2007. VLDB Endowment.

[8] Stanislav Bartoň, Vlastislav Dohnal, Jan Sedmidubský, and Pavel Zezula. Building self-organized image retrieval network. In *Proceeding of the 2008 ACM workshop on Large-Scale distributed systems for information retrieval (LSDS-IR 2008)*, pages 51–58, USA, 2008. ACM.

[9] Stanislav Bartoň, Vlastislav Dohnal, Jan Sedmidubský, and Pavel Zezula. *Computational Social Network Analysis: Trends, Tools and Research Advances*, chapter Towards Self-organizing Search Systems, pages 49–80. Computer Communications and Networks. Springer, New York, NY, USA, 2010.

[10] Michal Batko. *Scalable and Distributed Similarity Search*. PhD thesis, Masaryk University, May 2006.

[11] Michal Batko, Vlastislav Dohnal, David Novák, and Jan Sedmidubský. Mufin: A multi-feature indexing network. In *Proceedings of the Second International Workshop on Similarity Search and Applications (SISAP 2009)*, pages 158–159, Washington, DC, USA, 2009. IEEE Computer Society.

[12] Michal Batko, Vlastislav Dohnal, and Pavel Zezula. M-grid: Similarity searching in grids. In *Proceedings of International Workshop on Information Retrieval in Peer-to-Peer Networks, ACM CIKM 2006*, pages 17–24, Arlington, 2006. ACM Press.

[13] Michal Batko, David Novák, Fabrizio Falchi, and Pavel Zezula. On scalability of the similarity search in the world of peers. In *Proceedings of First International Conference on Scalable Information Systems (INFOSCALE 2006), Hong Kong, May 30 – June 1, 2006*, pages 1–12, New York, NY, USA, 2006. ACM Press.

[14] Michal Batko, David Novák, Fabrizio Falchi, and Pavel Zezula. Scalability comparison of peer-to-peer similarity search structures. *Future Generation Computer Systems*, 24(8):834–848, October 2008.

[15] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. Genbank: update. *Nucleic Acids Research*, 32:Database Issue D23–D26, 2004.

[16] ERRORinAuthors Bib298. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Comput. Surv.*, 33(3):322–373, 2001.

[17] Tolga Bozkaya and Meral Z. Özsoyoglu. Distance-based indexing for high-dimensional metric spaces. In Joan Peckham, editor, *Proceedings of the ACM International Conference on Management of Data (SIGMOD 1997), Tucson, Arizona, USA, May 13-15, 1997*, pages 357–368. ACM Press, 1997.

[18] Tolga Bozkaya and Meral Z. Özsoyoglu. Indexing large metric spaces for similarity search queries. *ACM Transactions on Database Systems (TODS 1999)*, 24(3):361–404, 1999.

[19] Sergey Brin. Near neighbor search in large metric spaces. In Umeshwar Dayal, Peter M. D. Gray, and Shojiro Nishio, editors, *Proceedings of the 21th International*

Conference on Very Large Data Bases (VLDB 1995), Zurich, Switzerland, September 11-15, 1995, pages 574–584. Morgan Kaufmann, 1995.

[20] Nicolas Bruno and Hui (Wendy) Wang. The threshold algorithm: From middleware systems to the relational engine. *IEEE Transactions on Knowledge and Data Engineering*, 19(4):523–537, 2007.

[21] Petra Budikova, Michal Batko, and Pavel Zezula. Improving the image retrieval system by ranking. In *Proceedings of the Third International Conference on SImilarity Search and APplications (SISAP 2010)*, pages 123–124, New York, NY, USA, 2010. ACM.

[22] Edouard Bugnion, Shi Fhei, Thomas Roos, Peter Widmayer, and Felizitas Widmer. A spatial index for approximate multiple string matching. In Ricardo A. Baeza-Yates and N. Ziviani, editors, *Proceedings of the 1st South American Workshop on String Processing (WSP 1993), Belo Horizonte, Brazil, September 13-15, 1993*, pages 43–53, 1993.

[23] Walter A. Burkhard and Robert M. Keller. Some approaches to best-match file searching. *Communications of the ACM (CACM 1973)*, 16(4):230–236, 1973.

[24] Benjamin Bustos and Tomáš Skopal. Dynamic similarity search in multi-metric spaces. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval (MIR 2006)*, pages 137–146, New York, NY, USA, 2006. ACM.

[25] Edgar Chávez, Karina Figueroa, and Gonzalo Navarro. Effective proximity retrieval by ordering permutations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(9):1647–1658, 2008.

[26] Edgar Chávez, José L. Marroquín, and Ricardo A. Baeza-Yates. Spaghettis: An array based algorithm for similarity queries in metric spaces. In *Proceedings of the 6th International Symposium on String Processing and Information Retrieval & International Workshop on Groupware (SPIRE/CRIWG 1999), Cancun, Mexico, September 21-24, 1999*, pages 38–46. IEEE Computer Society, 1999.

[27] Edgar Chávez, José L. Marroquín, and Gonzalo Navarro. Overcoming the curse of dimensionality. In *Procedings of the European Workshop on Content-Based Multimedia Indexing (CBMI 1999), Toulouse, France, October 25-27, 1999*, pages 57–64, 1999.

[28] Edgar Chávez, José L. Marroquín, and Gonzalo Navarro. Fixed Queries Array: A fast and economical data structure for proximity searching. *Multimedia Tools and Applications*, 14(2):113–135, 2001.

[29] Edgar Chávez and Gonzalo Navarro. An effective clustering algorithm to index high dimensional metric spaces. In *Proc. 7th International Symposium on String Processing and Information Retrieval (SPIRE)*, pages 75–86. IEEE CS Press, 2000.

[30] Edgar Chávez and Gonzalo Navarro. A compact space decomposition for effective metric indexing. *Pattern Recogn. Lett.*, 26(9):1363–1376, 2005.

[31] Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. Searching in metric spaces. *ACM Comput. Surv.*, 33(3):273–321, 2001.

[32] Jan Chomicki. Querying with intrinsic preferences. In Christian S. Jensen, Keith G. Jeffery, Jaroslav Pokorný, Simonas Saltenis, Elisa Bertino, Klemens Böhm, and Matthias Jarke, editors, *Proceedings of the 8th International Conference on Extending Database Technology (EDBT 2002)*, Lecture Notes in Computer Science, pages 34–51. Springer, 2002.

[33] Ondřej Chum, M. Perdoch, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *Proceedings of Computer Vision and Pattern Recognition (CVPR 2009)*, pages 17–24, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[34] Ondřej Chum, James Philbin, Michael Isard, and Andrew Zisserman. Scalable near identical image and shot detection. In *Proceedings of the 6th ACM international conference on Image and video retrieval (CIVR 2007)*, pages 549–556, New York, NY, USA, 2007. ACM.

[35] Paolo Ciaccia and Marco Patella. Bulk loading the M-tree. In *Proceedings of the 9th Australasian Database Conference (ADC 1998), Perth, Australia, February 2-3, 1998*, Australian Computer Science Communications, pages 15–26. Springer, 1998.

[36] Paolo Ciaccia and Marco Patella. Searching in metric spaces with user-defined and approximate distances. *ACM Transactions on Database Systems (TODS 2002)*, 27(4):398–437, 2002.

[37] Paolo Ciaccia, Marco Patella, and Pavel Zezula. M-tree: An efficient access method for similarity search in metric spaces. In Matthias Jarke, Michael J. Carey, Klaus R. Dittrich, Frederick H. Lochovsky, Pericles Loucopoulos, and Manfred A. Jeusfeld, editors, *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB 1997), Athens, Greece, August 25-29, 1997*, pages 426–435. Morgan Kaufmann, 1997.

[38] Tzi cker Chiueh. Content-based image indexing. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994), Santiago de Chile, Chile, September 12-15, 1994*, pages 582–593. Morgan Kaufmann, 1994.

[39] Arturo Crespo and Hector Garcia-Molina. Routing indices for peer-to-peer systems. In *Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS 2002)*, pages 1–23, Washington, DC, USA, 2002. IEEE Computer Society.

[40] Arturo Crespo and Hector Garcia-Molina. Semantic overlay networks for p2p systems. In *Proceedings of the 3rd International Workshop on Agents and Peer-to-Peer Computing (AP2PC 2004), New York, NY, USA, July 19, 2004*, Lecture Notes in Computer Science, pages 1–13. Springer, 2004.

[41] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Comm. ACM*, 51(1):107–113, 2008.

[42] Frank K. H. A. Dehne and Hartmut Noltemeier. Voronoi trees and clustering problems. *Information Systems (IS 1987)*, 12(2):171–175, 1987.

[43] Vlastislav Dohnal, Claudio Gennaro, Pasquale Savino, and Pavel Zezula. D-Index: Distance searching index for metric data sets. *Multimedia Tools and Applications*, 21(1):9–33, 2003.

[44] Vlastislav Dohnal, Claudio Gennaro, and Pavel Zezula. *Computational Intelligence in Medical Informatics*, chapter Efficiency and Scalability Issues in Metric Access Methods, pages 235–264. Springer-Verlag Berlin Heidelberg, Berlin, Germany, 1 edition, 2008.

[45] Vlastislav Dohnal and Jan Sedmidubský. Query routing mechanisms in self-organizing search systems. In *2nd International Workshop on Similarity Search and Applications (SISAP 2009)*, pages 132–139, Los Alamitos, CA 90720-1314, 2009. IEEE Computer Society.

[46] Vlastislav Dohnal, Jan Sedmidubský, Pavel Zezula, and David Novák. Similarity searching: Towards bulk-loading peer-to-peer networks. In *1st International Workshop on Similarity Search and Applications (SISAP 2008)*, pages 87–94, Los Alamitos CA, Washington, Tokyo, 2008. IEEE Computer Society.

[47] Vlastislav Dohnal and Pavel Zezula. Similarity searching in structured and unstructured p2p networks. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 400–416, 2009.

[48] Vlastislav Dohnal and Pavel Zezula. Real-life performance of metric searching. *SIGSPATIAL Special*, 2(2):28–31, 2010.

[49] Christos Doulkeridis, Akrivi Vlachou, Yannis Kotidis, and Michalis Vazirgiannis. Peer-to-peer similarity search in metric spaces. In Christoph Koch, Johannes Gehrke, Minos N. Garofalakis, Divesh Srivastava, Karl Aberer, Anand Deshpande, Daniela Florescu, Chee Yong Chan, Venkatesh Ganti, Carl-Christian Kanne, Wolfgang Klas, and Erich J. Neuhold, editors, *VLDB 2007: 33rd International Conference on Very Large Data Bases, September 23–27 2007, University of Vienna, Austria*, pages 986–997. ACM, 2007.

[50] Andrea Esuli. PP-Index: Using permutation prefixes for efficient and scalable approximate similarity search. In *Proceedings of the 7th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR 2009)*, pages 17–24, 2009.

[51] Ronald Fagin. Combining fuzzy information from multiple systems. In *Proceedings of the 15th ACM Symposium on Principles of Database Systems (PODS 1996), Montreal, Canada, June 3-5, 1996*, pages 216–226. ACM Press, 1996.

[52] Fabrizio Falchi, Claudio Gennaro, and Pavel Zezula. A content-addressable network for similarity search in metric spaces. In *Databases, Information Systems, and*

Peer-to-Peer Computing, International Workshops, DBISP2P 2005/2006, Trondheim, Norway, August 28–29, 2005, Seoul, Korea, September 11, 2006, Revised Selected Papers, Lecture Notes in Computer Science, pages 98–110. Springer, August 2007.

[53] Christos Faloutsos, Ron Barber, Myron Flickner, James L. Hafner, Wayne Niblack, Dragutin Petkovic, and William Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems (JIIS 1994)*, 3(3/4):231–262, 1994.

[54] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[55] Volker Gaede and Oliver Günther. Multidimensional access methods. *ACM Computing Surveys (CSUR 1998)*, 30(2):170–231, 1998.

[56] Claudio Gennaro, Matteo Mordacchini, Salvatore Orlando, and Fausto Rabitti. MRoute: A peer-to-peer routing index for similarity search in metric spaces. In *Proceedings of the 5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007)*, pages 1–12, September 2007.

[57] James L. Hafner, Harpreet S. Sawhney, William Equitz, Myron Flickner, and Wayne Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI 1995)*, 17(7):729–736, July 1995.

[58] Gisli R. Hjaltason and Hanan Samet. Index-driven similarity search in metric spaces. *ACM Trans. Database Syst.*, 28(4):517–580, 2003.

[59] Tomáš Homola, Vlastislav Dohnal, and Pavel Zezula. Proximity-based order-respecting intersection for searching in image databases. In *8th International Workshop on Adaptive Multimedia Retrieval, AMR'2010*, Linz, 2010.

[60] Tomáš Homola, Vlastislav Dohnal, and Pavel Zezula. Sub-image searching through intersection of local descriptors. In *3rd International Conference on Similarity Search and Applications (SISAP 2010)*, pages 127–128, New York, 2010. ACM Press.

[61] Edwin H. Jacox and Hanan Samet. Metric space similarity joins. *ACM Trans. Database Syst.*, 33(2):1–38, 2008.

[62] H. V. Jagadish, Alberto O. Mendelzon, and Tova Milo. Similarity-based queries. In *Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 1995)*, pages 36–45. ACM Press, 1995.

[63] H. V. Jagadish, Beng Chin Ooi, Kian-Lee Tan, Cui Yu, and Rui Zhang. iDistance: An adaptive B^+-tree based indexing method for nearest neighbor search. *ACM Transactions on Database Systems (TODS 2005)*, 30(2):364–397, 2005.

[64] Alexis Joly and Olivier Buisson. Logo retrieval with a contrario visual query expansion. In *Proceedings of the seventeen ACM international conference on Multimedia (MM 2009)*, pages 581–584, New York, NY, USA, 2009. ACM.

[65] Caetano Traina Jr., Agma J. M. Traina, Bernhard Seeger, and Christos Faloutsos. Slim-Trees: High performance metric trees minimizing overlap between nodes. In Carlo Zaniolo, Peter C. Lockemann, Marc H. Scholl, and Torsten Grust, editors, *Proceedings of the 7th International Conference on Extending Database Technology (EDBT 2000), Konstanz, Germany, March 27-31, 2000*, Lecture Notes in Computer Science, pages 51–65. Springer, 2000.

[66] Iraj Kalantari and Gerard McDonald. A data structure and an algorithm for the nearest point problem. *IEEE Transactions on Software Engineering (TSE 1983)*, 9(5):631–634, 1983.

[67] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of Computer Vision and Pattern Recognition (CVPR 2004)*, pages 506–513. IEEE Computer Society, 2004.

[68] Yan Ke, Rahul Sukthankar, and Larry Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM Multimedia*, pages 869–876, 2004.

[69] John L. Kelly. *General Topology*. D. Van Nostrand, New York, 1955.

[70] Teuvo Kohonen. *Self-Organization and Associative Memory*. Springer, 1984.

[71] V.I. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17, 1965.

[72] Alessandro Linari and Marco Patella. Metric overlay networks: Processing similarity queries in P2P databases. In *Proceedings of the 5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007)*, pages 1–8, September 2007.

[73] Witold Litwin, Marie-Anna Neimat, and Donovan A. Schneider. LH* – A scalable, distributed data structure. *ACM Transactions on Database Systems*, 21(4):480–525, 1996.

[74] Jakub Lokoč and Tomáš Skopal. On reinsertions in m-tree. In *IEEE 24th International Conference on Data Engineering Workshop (ICDE Workshops 2008)*, pages 410–417, apr. 2008.

[75] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. Multi-probe lsh: efficient indexing for high-dimensional similarity search. In *Proceedings of the 33rd international conference on Very large data bases (VLDB 2007)*, pages 950–961. VLDB Endowment, 2007.

[76] Margarida Mamede. Recursive lists of clusters: A dynamic data structure for range queries in metric spaces. In Pinar Yolum, Tunga Güngör, Fikret Gürgen, and Can Özturan, editors, *Computer and Information Sciences (ISCIS 2005)*, volume 3733 of *Lecture Notes in Computer Science*, pages 843–853. Springer Berlin / Heidelberg, 2005. 10.1007/11569596_86.

[77] Luisa Micó, Jose Oncina, and Rafael C. Carrasco. A fast branch & bound nearest neighbour classifier in metric spaces. *Pattern Recognition Letters*, 17(7):731–739, 1996.

[78] Luisa Micó, Jose Oncina, and Enrique Vidal. An algorithm for finding nearest neighbors in constant average time with a linear space complexity. In *Proceedings of the 11th International Conference on Pattern Recognition (ICPR 1992), The Hague, The Netherlands*, pages 557–560, 1992.

[79] Luisa Micó, Jose Oncina, and Enrique Vidal. A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements. *Pattern Recognition Letters*, 15(1):9–17, 1994.

[80] Peter R. Monge and Noshir S. Contractor. *Theories of Communication Networks*. Oxford University Press, April 2003.

[81] Arnoldo Jose Muller-Molina and Takeshi Shinohara. On the configuration of the similarity search data structure d-index for high dimensional objects. In David Taniar, Osvaldo Gervasi, Beniamino Murgante, Eric Pardede, and Bernady Apduhan, editors, *Computational Science and Its Applications (ICCSA 2010)*, volume 6018 of *Lecture Notes in Computer Science*, pages 443–457. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-12179-1_37.

[82] Gonzalo Navarro. Searching in metric spaces by spatial approximation. In *Proceedings of the 6th International Symposium on String Processing and Information Retrieval (SPIRE 1999), Cancun, Mexico, September 21-24, 1999*, pages 141–148. IEEE Computer Society, 1999.

[83] Gonzalo Navarro. A guided tour to approximate string matching. *ACM Computing Surveys (CSUR 2001)*, 33(1):31–88, 2001.

[84] Gonzalo Navarro. Searching in metric spaces by spatial approximation. *The VLDB Journal*, 11(1):28–46, 2002.

[85] Gonzalo Navarro and Nora Reyes. Dynamic spatial approximation trees. *J. Exp. Algorithmics*, 12:1–68, 2008.

[86] Gonzalo Navarro and Nora Reyes. Dynamic spatial approximation trees for massive data. *International Workshop on Similarity Search and Applications (SISAP 2009)*, pages 81–88, 2009.

[87] Hartmut Noltemeier, Knut Verbarg, and Christian Zirkelbach. A data structure for representing and efficient querying large scenes of geometric objects: MB* Trees. In *Geometric Modelling*, Computing Supplement, pages 211–226. Springer, 1992.

[88] Hartmut Noltemeier, Knut Verbarg, and Christian Zirkelbach. Monotonous Bisector* Trees - a tool for efficient partitioning of complex scenes of geometric objects. In *Data Structures and Efficient Algorithms*, Lecture Notes in Computer Science, pages 186–203. Springer, 1992.

[89] David Novák. *Similarity Search on a Very Large Scale*. PhD thesis, Masaryk University, 2008.

[90] David Novák and Michal Batko. Metric index: An efficient and scalable solution for similarity search. *International Workshop on Similarity Search and Applications (SISAP 2009)*, pages 65–73, 2009.

[91] David Novák, Michal Batko, Vlastislav Dohnal, and Pavel Zezula. Scaling up the image content-based retrieval. In *Second DELOS Conference on Digital Libraries*, pages 1–10, Pisa, Italy, 2007. Information Society Technologies.

[92] David Novák and Pavel Zezula. M-Chord: A scalable distributed similarity search structure. In *Proceedings of First International Conference on Scalable Information Systems (INFOSCALE 2006), Hong Kong, May 30 – June 1, 2006*, pages 1–10, New York, NY, USA, 2006. ACM Press.

[93] Rodrigo Paredes and Nora Reyes. List of twin clusters: A data structure for similarity joins in metric spaces. *Similarity Search and Applications, International Workshop on*, pages 131–138, 2008.

[94] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Schenker. A scalable content-addressable network. In *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM 2001), San Diego, California, August 27-31, 2001*, pages 161–172. ACM Press, 2001.

[95] Gerhard Roth and William Scott. Efficient indexing for strongly similar subimage retrieval. In *Proceedings of the Fourth Canadian Conference on Computer and Robot Vision (CRV 2007)*, pages 440–447, Washington, DC, USA, 2007. IEEE Computer Society.

[96] Min-Sung Ryu, Soo-Jun Park, and Chee Sun Won. Image retrieval using subimage matching in photos using mpeg-7 descriptors. In *Proceedings of Asia Information Retrieval Society Conference (AIRS 2005)*, Lecture Notes in Computer Science, pages 366–373. Springer, 2005.

[97] Hanan Samet. *Foundations of Multidimensional and Metric Data Structures*. Computer Graphics and Geometric Modeling. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[98] Jan Sedmidubský, Stanislav Bartoň, Vlastislav Dohnal, and Pavel Zezula. Querying similarity in metric social networks. In *Network-Based Information Systems, First International Conference, NBiS 2007*, number vol. 4658 in Lecture Notes in Computer Science, page 278, Berlin, 2007. Springer.

[99] Jan Sedmidubský, Stanislav Bartoň, Vlastislav Dohnal, and Pavel Zezula. Adaptive approximate similarity searching through metric social networks. In *24th International Conference on Data Engineering (ICDE 2008)*, pages 1424–1426, Los Alamitos CA, 2008. IEEE Computer Society.

[100] Jan Sedmidubský, Vlastislav Dohnal, Stanislav Bartoň, and Pavel Zezula. A self-organized system for content-based search in multimedia. In *IEEE International Symposium on Multimedia (ISM 2008)*, pages 322–327, Los Alamitos, California 90720-1314, 2008. IEEE Computer Society.

[101] Jan Sedmidubský, Vlastislav Dohnal, and Pavel Zezula. Feedback-based performance tuning for self-organizing multimedia retrieval systems. In *International Conference on Advances in Multimedia (MMEDIA 2010)*, pages 102–108, Los Alamitos, CA 90720-1314, 2010. IEEE Computer Society.

[102] Jan Sedmidubský, Vlastislav Dohnal, and Pavel Zezula. On building a self-organizing search system for multimedia retrieval. In *International Workshop on Multimedia and Semantic Technologies (MUST 2010)*, Red Hook, NY 12571, USA, 2010.

[103] Thomas Seidl and Hans-Peter Kriegel. Efficient user-adaptable similarity search in large multimedia databases. In *The VLDB Journal*, pages 506–515, 1997.

[104] Dennis Shasha and James Z. Wang. New techniques for best-match retrieval. *ACM Transactions on Information Systems (TOIS 1990)*, 8(2):140–158, 1990.

[105] Tomáš Skopal. Pivoting M-tree: A metric access method for efficient similarity search. In Václav Snášel, Jaroslav Pokorný, and Karel Richta, editors, *Proceedings of the Dateso 2004 Annual International Workshop on DAtabases, TExts, Specifications and Objects, Desna, Czech Republic, April 14-16, 2004*, CEUR Workshop Proceedings, pages 27–37. CEUR-WS.org, 2004.

[106] Tomáš Skopal, Vlastislav Dohnal, Michal Batko, and Pavel Zezula. Distinct nearest neighbors queries for similarity search in very large multimedia databases. In *11th ACM International Workshop on Web Information and Data Management (WIDM 2009)*, pages 11–14, New York, USA, 2009. ACM.

[107] Tomáš Skopal and Jakub Lokoč. Nm-tree: Flexible approximate similarity search in metric and non-metric spaces. In Sourav Bhowmick, Josef Küng, and Roland Wagner, editors, *Database and Expert Systems Applications*, volume 5181 of *Lecture Notes in Computer Science*, pages 312–325. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-85654-2_30.

[108] Tomáš Skopal, Jaroslav Pokorný, Michal Krátký, and Václav Snášel. Revisiting M-Tree building principles. In Leonid A. Kalinichenko, Rainer Manthey, Bernhard Thalheim, and Uwe Wloka, editors, *Proceedings of the 7th East European Conference on Advances in Databases and Information Systems (ADBIS 2003), Dresden, Germany, September 3-6, 2003*, Lecture Notes in Computer Science. Springer, 2003.

[109] Ion Stoica, Robert Morris, David Liben-Nowell, David R. Karger, Frans M. Kaashoek, Frank Dabek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup protocol for internet applications. *IEEE/ACM Transactions on Networking*, 11(1):17–32, 2003.

[110] Jeffrey K. Uhlmann. Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters*, 40(4):175–179, 1991.

[111] Enrique Vidal. An algorithm for finding nearest neighbors in (approximately) constant average time. *Pattern Recognition Letters*, 4(3):145–157, 1986.

[112] Enrique Vidal. New formulation and improvements of the nearest-neighbour approximating and eliminating search algorithm (AESA). *Pattern Recognition Letters*, 15(1):1–7, 1994.

[113] Stanley Wasserman, Katherine Faust, and Dawn Iacobucci. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, November 1994.

[114] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, June 1998.

[115] Peter N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the 4th Annual ACM Symposium on Discrete Algorithms (SODA 1993), Austin, Texas, USA, January 25-27, 1993*, pages 311–321. ACM Press, 1993.

[116] Peter N. Yianilos. Excluded middle vantage point forests for nearest neighbor search. In *Proceedings of the 6th DIMACS Implementation Challenge: Near Neighbor Searches (ALENEX 1999), Baltimore, Maryland, USA, January 15-16, 1999*, 1999.

[117] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity Search: The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer, 2006.

[118] Pavel Zezula, Michal Batko, and Vlastislav Dohnal. *Encyclopedia of Database Systems*, chapter Indexing Metric Spaces, pages 1–4. Database Management and Information Retrieval. Springer-Verlag, New York, 2009.

[119] Pavel Zezula, Michal Batko, Vlastislav Dohnal, David Novák, and Jan Sedmidubský. Similarity search in large collections of biometric data. In *NATO RTO Modelling and Simulation Group Symposium*, pages 1–13, Brussels, 2009.

[120] Pavel Zezula, Vlastislav Dohnal, and Michal Batko. *Encyclopedia of Computer Science and Engineering*, chapter File Organizations, pages 1219–1227. Wiley-Interscience, Hoboken, NJ, USA, 2009.

[121] Pavel Zezula, Vlastislav Dohnal, and David Novák. *Global Data Management*, chapter Towards Scalability of Similarity Searching, pages 277–300. IOS Press, Amsterdam, The Netherlands, 2006.

Part II

Collection of Articles

Chapter 9

Books

1. Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity Search: The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer, 2006.

Due to copyrights pending, the contents of papers are not included.

Chapter 10

Book Chapters

This chapter contains the list of book chapters including in this habilitation thesis. The individual contributing articles that follow after are titled as the original book chapters and not as the book they are published in.

1. Stanislav Bartoň, Vlastislav Dohnal, Jan Sedmidubský, and Pavel Zezula. *Computational Social Network Analysis: Trends, Tools and Research Advances*, chapter Towards Self-organizing Search Systems, pages 49–80. Computer Communications and Networks. Springer, New York, NY, USA, 2010.

2. Pavel Zezula, Michal Batko, and Vlastislav Dohnal. *Encyclopedia of Database Systems*, chapter Indexing Metric Spaces, pages 1–4. Database Management and Information Retrieval. Springer-Verlag, New York, 2009.

3. Pavel Zezula, Vlastislav Dohnal, and Michal Batko. *Encyclopedia of Computer Science and Engineering*, chapter File Organizations, pages 1219–1227. Wiley-Interscience, Hoboken, NJ, USA, 2009.

4. Vlastislav Dohnal, Claudio Gennaro, and Pavel Zezula. *Computational Intelligence in Medical Informatics*, chapter Efficiency and Scalability Issues in Metric Access Methods, pages 235–264. Springer-Verlag Berlin Heidelberg, Berlin, Germany, 1 edition, 2008.

5. Pavel Zezula, Vlastislav Dohnal, and David Novák. *Global Data Management*, chapter Towards Scalability of Similarity Searching, pages 277–300. IOS Press, Amsterdam, The Netherlands, 2006.

Due to copyrights pending, the contents of papers are not included.

Chapter 11

Journals

1. Vlastislav Dohnal and Pavel Zezula. Real-life performance of metric searching. *SIGSPATIAL Special*, 2(2):28–31, 2010.

2. Vlastislav Dohnal and Pavel Zezula. Similarity searching in structured and unstructured p2p networks. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 400–416, 2009.

Due to copyrights pending, the contents of papers are not included.

Chapter 12

Conference Papers

1. Tomáš Homola, Vlastislav Dohnal, and Pavel Zezula. Proximity-based order-respecting intersection for searching in image databases. In *8th International Workshop on Adaptive Multimedia Retrieval, AMR'2010*, Linz, 2010.

2. Tomáš Homola, Vlastislav Dohnal, and Pavel Zezula. Sub-image searching through intersection of local descriptors. In *3rd International Conference on Similarity Search and Applications (SISAP 2010)*, pages 127–128, New York, 2010. ACM Press.

3. Jan Sedmidubský, Vlastislav Dohnal, and Pavel Zezula. Feedback-based performance tuning for self-organizing multimedia retrieval systems. In *International Conference on Advances in Multimedia (MMEDIA 2010)*, pages 102–108, Los Alamitos, CA 90720-1314, 2010. IEEE Computer Society.

4. Jan Sedmidubský, Vlastislav Dohnal, and Pavel Zezula. On building a self-organizing search system for multimedia retrieval. In *International Workshop on Multimedia and Semantic Technologies (MUST 2010)*, Red Hook, NY 12571, USA, 2010.

5. Michal Batko, Vlastislav Dohnal, David Novák, and Jan Sedmidubský. Mufin: A multi-feature indexing network. In *Proceedings of the Second International Workshop on Similarity Search and Applications (SISAP 2009)*, pages 158–159, Washington, DC, USA, 2009. IEEE Computer Society.

6. Vlastislav Dohnal and Jan Sedmidubský. Query routing mechanisms in self-organizing search systems. In *2nd International Workshop on Similarity Search and Applications (SISAP 2009)*, pages 132–139, Los Alamitos, CA 90720-1314, 2009. IEEE Computer Society.

7. Tomáš Skopal, Vlastislav Dohnal, Michal Batko, and Pavel Zezula. Distinct nearest neighbors queries for similarity search in very large multimedia databases. In *11th ACM International Workshop on Web Information and Data Management (WIDM 2009)*, pages 11–14, New York, USA, 2009. ACM.

8. Pavel Zezula, Michal Batko, Vlastislav Dohnal, David Novák, and Jan Sedmidubský. Similarity search in large collections of biometric data. In *NATO RTO Modelling and Simulation Group Symposium*, pages 1–13, Brussels, 2009.

9. Vlastislav Dohnal, Jan Sedmidubský, Pavel Zezula, and David Novák. Similarity searching: Towards bulk-loading peer-to-peer networks. In *1st International Workshop on Similarity Search and Applications (SISAP 2008)*, pages 87–94, Los Alamitos CA, Washington, Tokyo, 2008. IEEE Computer Society.

10. Jan Sedmidubský, Stanislav Bartoň, Vlastislav Dohnal, and Pavel Zezula. Adaptive approximate similarity searching through metric social networks. In *24th International Conference on Data Engineering (ICDE 2008)*, pages 1424–1426, Los Alamitos CA, 2008. IEEE Computer Society.

11. Jan Sedmidubský, Vlastislav Dohnal, Stanislav Bartoň, and Pavel Zezula. A self-organized system for content-based search in multimedia. In *IEEE International Symposium on Multimedia (ISM 2008)*, pages 322–327, Los Alamitos, California 90720-1314, 2008. IEEE Computer Society.

12. Stanislav Bartoň, Vlastislav Dohnal, Jan Sedmidubský, and Pavel Zezula. Building self-organized image retrieval network. In *Proceeding of the 2008 ACM workshop on Large-Scale distributed systems for information retrieval (LSDS-IR'08)*, pages 51–58, USA, 2008. ACM.

13. Jan Sedmidubský, Stanislav Bartoň, Vlastislav Dohnal, and Pavel Zezula. Querying similarity in metric social networks. In *Network-Based Information Systems, First International Conference, NBiS 2007*, number vol. 4658 in Lecture Notes in Computer Science, page 278, Berlin, 2007. Springer.

14. Stanislav Bartoň, Vlastislav Dohnal, Jan Sedmidubský, and Pavel Zezula. Gauging the evolution of metric social network. In *5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007) held at 33rd International Conference on Very Large Data Bases (VLDB 2007)*, pages 1–12, Vienna, 2007. VLDB Endowment.

15. David Novák, Michal Batko, Vlastislav Dohnal, and Pavel Zezula. Scaling up the image content-based retrieval. In *Second DELOS Conference on Digital Libraries*, pages 1–10, Pisa, Italy, 2007. Information Society Technologies.

16. Michal Batko, Vlastislav Dohnal, and Pavel Zezula. M-grid: Similarity searching in grids. In *Proceedings of International Workshop on Information Retrieval in Peer-to-Peer Networks, ACM CIKM 2006*, pages 17–24, Arlington, 2006. ACM Press.