

The Images and Data Bases Group at University of São Paulo

Agma Juci Machado Traina, Caetano Traina Júnior, Cristina Dutra de Aguiar Ciferri,
Elaine Parros Machado de Sousa, Jose Fernando Rodrigues Júnior

Computer Science Dept., ICMC-USP, São Carlos-SP, Brazil
{agma, caetano, cdac, parros, junio}@icmc.usp.br

Abstract. Created in the middle of the 1990s, the Images and Data Bases Group (*Grupo de Bases de Dados e de Imagens*) at the Institute of Mathematics and Computer Science of the University of São Paulo at São Carlos – GBdI-ICMC-USP – has since the beginning targeted at the development of core methods and data analysis techniques to handle large databases for scientific applications involving complex data. This paper describes its main research activities, which include applying Fractal Theory concepts to improve analysis and search operations over very large datasets; the development of search algorithms over complex data; indexing structures and optimization techniques for similarity-based queries; the development of data mining and data warehousing techniques aimed at dealing with datasets with large cardinality and dimensionality; the development of information visualization for data stored in relational databases; and techniques to allow data integration from diverse sources including those maintained in complex structures. Derived from those basic research activities, several tools have been developed and released to the community, including tools for computer-assisted medical diagnosis and for climate change analysis, libraries to aid in the development of efficient and efficacious applications handling data in multidimensional and metric spaces, and software libraries to help including similarity queries into existing Database Management Systems. This paper describes the main research activities and the main results achieved, which are the basis for collaborative work with other groups all over the world, aiming at the development of new technologies and applications that employ DBMS's to handle images and several other types of complex data.

Categories and Subject Descriptors: H.2.4 [Database management]: Systems—*Multimedia databases*

Keywords: Fractal Theory, Medical databases, Biological databases, Similarity queries, Data integration and provenance

1. THE EARLY YEARS

The history of the database research in the Institute of Mathematics and Computer Sciences (ICMC) started with the demand on organizing and searching complex data provided from physics experiments from the Physics and Chemistry Institute of São Carlos (IFQSC). The collaboration between both institutes started through several PhD projects developed at IFQSC, where those who later became the professors of ICMC developed their theses. The research on a low field/low cost Resonance Magnetic Tomography device added more demand on putting together images and descriptions.

The research groups started to consolidate inside ICMC around the middle eighties, and in 1988 the Database Laboratory was established under the leadership of Prof. Caetano Traina Jr. Later, Prof. Agma Traina joined the laboratory in 1991 working on Image Processing, and the two faculties created the Images and Data Bases Group (*Grupo de Bases de Dados e de Imagens* - GBdI) in 1992, targeting the development of techniques to store and to retrieve images from databases, at those times a great research issue that was in its very beginning. After then, along the 20 years of its existence, the group grew with the arrival of three other faculties, Profs. Cristina Ciferri, Elaine Sousa and Jose F. Rodrigues Jr., co-authors of this paper.

During the period from 1980 to 1995, the research activities in Database at ICMC were strongly

This work has been supported by FAPESP, CAPES, CNPq, CNRS and Microsoft Research.

Copyright©2010 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

attached to the Physics Institute; its main drive at those times were the research needs from two broad projects: high energy physics and the construction of a low field human body tomography. Both projects involved the storage and retrieval of large amounts of data, and in the case of the tomography, images, many images. Working with both the theoretical and practical aspects of these projects helped shaping the identity of GBdI and its research area: using databases to handle large amounts of scientific data and images, and also its strong roots in the development of database techniques aiming at the Medical field.

Computer scientists working with medical applications require a tight integration with physicians and people from health services. Thus, the period of GBdI formation also saw the start of a prolific partnership with the Image Science and Medical Physics Center (CCIFM) from the Faculty of Medicine of USP at Ribeirão Preto (FMRP), which counts with a school hospital. The CCIFM center is part of the Radiology Department of that school Hospital, and integrates the faculty and the health staff that are in charge of the medical imaging equipment of the hospital, including x-ray, Computerized Tomography (CT), mammography, ultra-sound and other equipments. Besides providing lots of test cases, this partnership provided several research challenges. When new faculties joined the group, after 2004, GBdI expanded its investigation arena to other types of complex data and database applications, such as digital libraries, climate, geographic and biological data, including multi-source data integration and provenance, data warehousing, information visualization, and graph-based analysis, always pursuing the premise of efficiently handling large datasets.

Another seminal partnership that GBdI has maintained with other institutions is the one with the Database Group from the University of Carnegie Mellon (DBG-CMU), in Pittsburgh, PA, USA. While GBdI has been interested in applications involving images from medical domains, DBG-CMU has been interested in broader application fields, involving collaborative networks, television broadcasting news and military applications, but all of them centered on retrieving data from large datasets, either by similarity or through analytical processing. Therefore, theoretical results jointly developed have been applied by each institution to their respective application targets.

Besides those long term partnerships, GBdI has successfully maintained collaboration, both with Brazilian universities, like the Federal University of Uberlândia and the Federal University of Pernambuco, and with universities abroad, like the University of Bourgogne (uB) at Dijon, France, the University of Calgary, Alberta, Canada, and the University of California at Riverside. Those activities have been important to drive our research group in the directions that we have pursued, which can be summarized as: *“To develop and apply database technology in the development of scientific applications requiring large databases, specially technologies suitable for image-based software and multidomain integration.”* Following, we present some of the fundamental concepts that have been explored in our group, as well as we describe the main results already obtained.

2. BASIC RESEARCH ACTIVITIES

There are two main research activities conducted at GBdI, both targeting the development of techniques to improve storage and retrieval of complex data in large databases of scientific applications. The first corresponds to execute similarity queries over complex data, such as multimedia data, multidimensional time series and biological sequences, and the second corresponds to retrieve data represented through complex structures from several sources. We have worked to create techniques that make the use of DBMS's practical when it is required to support scientific activities that handle very large datasets in applications such as medicine, biology, and climate and agricultural. Our techniques allow integrating several data sources and data modalities into a complete database over which data can be retrieved based on the similarities among the stored elements. Figure 1 depicts the main research areas and the relationships among them, as they are being studied at GBdI.

In the following subsections we present the main basic research activities in which GBdI has ac-

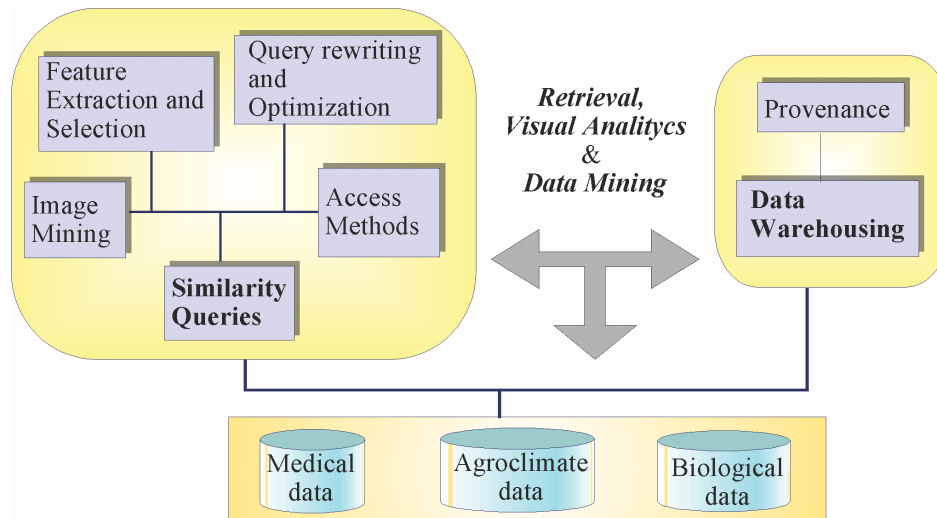


Fig. 1. The relationship among the diverse GBdI research areas.

complicated interesting results. In the next section we present some of the applied results that those research activities have contributed to.

2.1 Fractal Theory

A distinctive activity in which GBdI has engaged is employing Fractal Theory to solve problems of processing large databases of complex data and, in particular, the execution of analysis over them. Data mining tasks are notorious for heavy processing requirements, often relying on algorithms with computational complexities far beyond linear. As database operations should, ideally, execute within linear complexity both on time and on number of disk accesses, alternatives must be discovered to reduce algorithm's complexity. Our group has achieved successful improvements using results from the Fractal Theory to achieve linear processing times to execute several operations, allowing executing each operation in a smaller number of passes. In this way, we have derived fast algorithms to perform classification [Romero et al. 2007], clustering [Cordeiro et al. 2010; Cordeiro et al. 2013] and association rule mining [Ribeiro et al. 2009], to find correlations among attributes in a dataset [Sousa et al. 2002; Sousa et al. 2007], to estimate selectivity of retrieval operations over metric [Baio et al. 2007] and multi-dimensional data [Traina et al. 2001]. Many of those results were developed using a basic algorithm to compute the Correlation Fractal Dimension that was the first one to achieve a linear-time complexity both on cardinality and dimensionality. That algorithm was developed by our group and firstly presented at the Brazilian Database Symposium (SBBD) of 2000. It is worth to note that currently it is one of the most cited papers in the history of SBBD [Traina et al. 2000; 2010]. Those basic techniques have been extended and successfully applied to datasets from several domains, including climate analysis [Romani et al. 2009] and medical data [Ribeiro et al. 2009].

2.2 Similarity Search

Another research topic where GBdI brought well-accepted contributions is the development of access methods for metric data. Our first publication on this subject introduced the now well-known *Slim-tree* access method [Traina et al. 2000], the first metric access method (MAM) able to take into account the overlapping between nodes to reduce the data retrieval complexity [Traina et al. 2002]. The Slim-tree was developed more than 10 years ago and it has been used in several applications all over the world, but it also raised issues that are yet being studied. We contributed to resolve

several of those issues too, such as developing algorithms for bulk-loading [Vespa et al. 2010] (also first published in SBB [Vespa et al. 2007]), update and delete operations [Bueno et al. 2008], and support for paged similarity queries [Seraphim et al. 2011]. This research line also provided several other MAMs, such as the DBM-tree [Vieira et al. 2004; Vieira et al. 2006; 2010], the first and yet the single MAM to explore adequate balancing among deep- and width-navigation over a tree to answer queries that requires node overlap resolution; and the Onion-tree [Carelo et al. 2011], the first dynamic main memory indexing of metric data.

Embedding a metric space into a multi-dimensional one was also explored, using the Fractal Theory as the guiding concept to evaluate both the the embedding space dimensionality and the most important parameters required to tune the MAMs. Thus, a number of techniques were developed to free the user from the burden of defining parameters that are otherwise obtainable only by experimental trial [Traina et al. 2002; Traina et al. 2007].

The execution of similarity searches is highly dependent on two basic kinds of queries: similarity range (Rq) and k -nearest neighbor queries ($kNNq$). However, sometimes using just those two kinds of queries proves to be not enough to meet the user's demands. Thus, we worked on extending both to provide for the user a broader choice of options. For example, new query types were developed to allow a query that specifies more than one query center [Razente et al. 2008], and to allow the system to provide diversification in the similarity query results [Vieira et al. 2011b; 2011a].

There is a dual property relating indexing structures in metric versus multi-dimensional spaces. Although multidimensional spaces provide a much richer set of properties to be explored by the search algorithms than the metric space does – and thus theoretically being able to provide more opportunities for improvement – increasing space dimensionality quickly plummets both, due to the well-known dimensionality curse effect [Katayama and Satoh 2001]. On the other hand, although metric spaces seem to provide fewer improvement opportunities, they both can greatly benefit from the Fractal theory to automatically tailor the parameters of search algorithms to each specific dataset. Indeed, the intrinsic dimensionality measured from a metric or multidimensional dataset is usually very low (usually smaller than six or seven), hence, treating multidimensional dataset as a metric one often brings huge benefits. For example, instead of working with the hundreds of dimensions in datasets of image color histograms, we developed the Metric Histogram distance function [Traina et al. 2002; Traina et al. 2003], which allows obtaining nearly the same results as working on an equivalent dataset having only a half-dozen dimensions.

Non-metric perceptual distance functions can also benefit from the fractal analysis of the datasets [Felipe et al. 2006; 2008]. We have shown that stretching a metric space using the power laws induced from fractals enables postponing the incurring curse of dimensionality to significantly higher dimensionality [Pola et al. 2009]. Therefore, other important contributions of our group were deriving techniques to improve similarity search algorithms, such as showing how to exchange the costly $kNNq$ for the much cheaper Rq [Arantes et al. 2003; Vieira et al. 2007].

2.3 Information Visualization

Automatic tools for data analysis often require fine tuning several parameters, many of them in a non-intuitive manner. Moreover, the most adequate tools for each data analysis task have a tight correlation with specific properties existing in the data being analyzed, such as spatial distribution, existence of clusters and outliers, cardinality of each attribute domain, among others. It is common that the user does not have a clear idea of what properties the data under analysis contains. Thus, the ability to “see” the data is an important resource for data miners. Visualization of the data derived from scientific activities is an even more valuable resource, as multiple attributes are often correlated in ways that are difficult for the analyst to seize. Therefore, it is usual to consider each single attribute at a time, pursuing individual, uni-dimensional measurements. In this scenario, particularly regarding data from scientific experiments or from image processing algorithms, we have developed a number

of information visualization tools. All of them are able to handle data both from metric and from multi-dimensional domains, where there is no intrinsic spatial reference.

In the field of Information Visualization, we proposed techniques based on embedding a metric or a high-dimensional space into a small dimensionality (two or three dimensions), so that the embedded space can be plotted, involving data from one or more relations (Multi-relational visualization [Barioni et al. 2002]). Several of those techniques were integrated in a tool called FastMapDB, which employs techniques to quickly map the data stored in a DBMS in a 3D-space providing several interaction mechanisms for the user to foster her understanding of the data. In another line, we have researched on visualization techniques for multivariate, complex, and network data. Regarding multivariate data, we have designed a framework that enables the user to create multiple coordinated workspaces, so to improve the scalability in tasks such as exploring frequent patterns, correlation and tendency prediction [Rodrigues Jr et al. 2008] [Rodrigues Jr et al. 2005][Rodrigues Jr et al. 2013]. In respect to complex data, we have combined content-based data retrieval with multidimensional projection techniques [Rodrigues Jr et al. 2010; Rodrigues Jr et al. 2009]. Concerning network data, we have developed techniques to deal with very large graph-like data by means of hierarchical partitioning techniques [Rodrigues Jr et al. 2006; Rodrigues Jr et al. 2006; Rodrigues Jr et al. 2013]. Over this research, we have worked on defining the fundamentals of data visualization in the form of a unifying theory [Rodrigues Jr et al. 2007; Rodrigues Jr et al. 2008].

Another interesting visualization tool created to help our activities on developing indexing structures for metric data was the “MAM-View” framework [Vieira et al. 2010]. Most of these related techniques and tools were firstly presented at SBBD and its parallel events [Traina et al. 2001, :FastMapDB] [Paterlini et al. 2005, :MAM-View] [Fedel et al. 2008; Razente et al. 2004].

2.4 Data Mining

Several tools and techniques were developed at GBdI to perform data mining in large datasets. We have provided fast techniques to perform almost all of the most important data mining tasks, both extending existing techniques and creating new ones tailored to handle large datasets, including clustering [Cordeiro et al. 2010; Barioni et al. 2008] and correlation clustering [Cordeiro et al. 2010; Ferreira Cordeiro et al. 2011; Cordeiro et al. 2013], classification [Romero et al. 2007; Pan et al. 2006]. Our techniques are being applied on medical software to help computer-aided diagnosis [Ribeiro et al. 2009; Ribeiro et al. 2008], on meteorological data for climate change analyzes [Romani et al. 2010], tracking of user intention [Ferreira et al. 2010; Ferreira et al. 2010] and to improve similarity retrieval operations through approximate genetic algorithms [Bueno et al. 2007].

2.5 Data Warehousing

Another area of interest of GBdI is data warehousing, whose objective is to offer support for the decision-making process. A data warehousing environment consolidates data from several distributed, autonomous and heterogeneous information sources into a single component. Besides being integrated, the data in the DW are subject-oriented and historical. Furthermore, aiming at improving the performance of queries to perform On-line Analytical Processing (OLAP), a DW contains not only the detailed data consolidated from the information sources, but also several data snippets aggregated from the detailed data, pre-computed to speed up queries. The characteristics of the data in a DW, specially the fact that this database stores historical information covering a long time horizon, may turn its volume orders of magnitude larger than its corresponding transactional counterpart.

At GBdI, we have conducted research related to improving OLAP query performance over traditional and non-traditional DWs. We have investigated algorithms to perform horizontal and vertical fragmentation of data warehouses, and also proposed new data warehouse models for DWs that store images, tackling both how intrinsic features of images should be stored in a DW and the issue of how

OLAP similarity queries should be efficiently processed [Annibal et al. 2010]. We have also obtained impressive results regarding Geographic DWs, which are DWs extended to store spatial objects represented by geometries, such as points, lines and polygons. In this subject, we have investigated the effects of spatial data redundancy in SOLAP (spatial OLAP) query processing performance [Siqueira et al. 2009; Mateus et al. 2010]; proposed two efficient indexes for geographic DWs [Siqueira et al. 2011], called the Spatial Bitmap Index (SB-index) and the Hierarchical Spatial Bitmap Index (HSB-index); analyzed how to efficiently process drill-across queries over geographic DWs using the SB-index [Brito et al. 2011]; developed a spatial DW benchmark to allow a controlled experimental performance evaluation of environments for spatial DWs [Siqueira et al. 2010]; and investigated how vague spatial data affect query performance and storage requirements in geographic DWs [Siqueira et al. 2011]. As applied research on data warehousing, we developed the VisualTPCH tool, which offers a graphic interface to assist in the generation of synthetic data for data warehouse algorithm development and tuning, using the TPC-H benchmark [Domingues et al. 2008] as a basis.

2.6 Data Integration and Provenance

The research done at GBdI regarding data integration targets integrating data represented at the finest resolution, the so called instance level integration, empowered with data provenance. At the instance level, data integration aims at solving inconsistencies in data obtained from heterogeneous sources, which may contain information about the same real world entity but using distinct values or even distinct attributes to represent the same real world concept. Therefore, data cleaning refers to the process of solving attribute representation and value conflicts. Regarding data provenance, it is the related metadata that permits the identification source to track the transformations applied to the data over its life span. We employed data provenance aiming at reproducing the user's decisions, so the user's previous activities cleaning data can be tracked and automatically reproduced in future integration processes.

As a result, we proposed PrInt, a data provenance model that reproduces user's decisions in applications requiring data integration, where back propagation of updates on heterogeneous sources are not possible. [Tomazela et al. 2010]. We also have developed a model for XML data fusion, which allows the integration administrator to define rules for data cleaning, to solve value conflicts that have been detected during previous integration processes, and to automatically solve new conflicts arising in subsequent integration processes [Cecchin et al. 2010]. Further, we have developed a tool aimed at helping users to identify and solve inconsistencies among heterogeneous data [Tomazela et al. 2008]. It performs similarity search over two sets of heterogeneous entities, aligning pairs of the most similar entities, thus allowing the user to migrate data from one entity to the other.

2.7 Biological Databases

Biological databases (BDB) store biological data related to living organisms, such as biological sequences of nucleotides and amino-acids, genetic maps, gene annotation, three-dimensional structures of proteins, as well as references to related scientific papers and several other information related to biological data. The nucleotide sequences are represented as strings of the four nitrogen base types A, C, G and T, while the amino-acid chains are composed of strings that represent the 20 molecules present in proteins.

At GBdI, we have focused on sequence similarity search over BDBs using indexes, considering two different areas of interest. The first one refers to the development of access methods based on compression techniques and on matrices to index the sequences. We have developed the nsP-index [Ciferri et al. 2008] and the Proteinus-index [Louza et al. 2011], which aim at efficiently querying nucleotide and protein sequences, respectively. The second area of interest focuses on investigating access methods based on suffix trees. Here, we created Perseus, a technique to handle suffix trees that exceed the main memory capacity, allowing an efficient indexing of nucleotide sequences [Carelo et al.

2009]. We have also developed a visual tool to assist in the manipulation of suffix trees, allowing the interactive execution and visualization of the processes of suffix tree's creation and of the corresponding search algorithms [Louza et al. 2010; Chino et al. 2012].

3. APPLIED RESEARCH ACTIVITIES

In this section we present the most interesting results produced from our basic research activities over a number of application areas.

3.1 Medical software

When processing an image dataset, there are two ways to retrieve images that meet the criteria expressed in a user query: through tags manually attached to each image, or by executing image processing algorithms that automatically extract features to represent each image. When comparing pairs of images, either tags or features are employed in place of the real images. At GBdI, we are pursuing the latter alternative, as we target at large sets of, often confidential, images. Also, manually tagging such amount of images, which come from a highly specialized and technical domain, is not a suitable option. In particular, we are interested in image datasets of medical exams, where huge amounts of images are generated. As an example, the number of images from medical exams collected in the Clinical Hospital of Ribeirão Preto (HCFMRP) – USP, from October 2008 to November 2009 exceeds fourteen million images, filling eight Terabytes of disk [Kaster et al. 2011]. In this research area, our collaboration with the Center of Images and Medical Physics (CCIFM) has generated important and rewarding results. We developed computer-assisted diagnosis (CAD) tools to improve the efficacy of the medical exam analysis, thus benefiting every person treated at the hospital.

As the administrative systems of the hospital already employ an Oracle DBMS, we are developing new applications that store all of the hospital's images to create an infrastructure for health data integrated into the administrative information. Those new applications are based on the ability to perform queries executing the required procedures for similarity retrieval [Barioni et al. 2010; Bueno et al. 2009; Kaster et al. 2009; Kaster et al. 2010]. As part of that applied research, we have also developed: specialized feature extractors to handle images of medical exams, such as tissue identification using image texture [Felipe et al. 2003; Balan et al. 2005; Balan et al. 2012]; explored specialist's relevance feedback to reduce the semantic gap in content-based image retrieval [Traina et al. 2009; Azevedo-Marques et al. 2008]; and are working on feature extraction and selection algorithms to aid in decision making over several kinds of medical image modalities [Traina et al. 2010].

Images from medical exams are analyzed by physicians and radiologists using CAD tools specialized for each medical specialty. We have developed tools to integrate image similarity search on patient data, tools for mammography analyses [Azevedo-Marques et al. 2008; Rosa et al. 2002] and several techniques and algorithms to enable similarity search for this particular application [Ribeiro et al. 2007]. Another interesting result obtained performing similarity queries over non-image-based exams were obtained analyzing blood tests of thalassemia. Thalassemia is a group of genetic blood disorders that causes anemia in the early childhood and that lasts throughout life, caused by faulty synthesis of part of the hemoglobin molecule. An early identification of the problem can help improving the patients' quality of life, but it requires collecting newborns' blood for five distinct tests. Our analysis revealed that just two exams, with carefully chosen attributes provide statistically significant evidences of the disease, so the blood samples can be significantly reduced [Valêncio et al. 2004].

3.2 Query optimization

Several research groups all over the world are performing research on similarity queries, but few of them are studying how to integrate the individual results of the queries, considering that they

are always executed over a single predicate. In fact, currently there is no standard support for similarity queries, neither in SQL nor in more theoretical approaches based on the relational theory. Nevertheless, having both resources is important to allow a tighter integration of the current wide spectrum of available DBMS and relational tools with the new algorithms and techniques developed for similarity queries. That integration is specially important because similarity-based operations are computationally complex, but they can be significantly accelerated using relational-based optimization and buffer management techniques. To target that integration and to provide a real workbench for applied research on similarity queries using real data and operational environments, we developed SIREN (SImilarity Retrieval ENgine) [Barioni et al. 2005; 2009; Barioni et al. 2010]. Siren is a software wrapper running between a conventional DBMS and the application programs. It intercepts every SQL command sent from the application and executes the similarity-related operations internally, whereas it uses the underlying DBMS to execute the conventional data operations.

SIREN allows both unary (Rq and $kNNq$) and binary (range, k -nearest neighbor and k -closest neighbors joins) similarity queries to be expressed in SQL, and allows storing, indexing and retrieving complex objects such as images, audio files, and others [Barioni et al. 2006]. Siren is a good workbench for both basic and applied research on how to include similarity queries as a new predicate type. Aiming at seamlessly integrating similarity queries to those already existing in DBMS, we have developed studies on query cost estimation [Baio et al. 2007] and query optimization [Arantes et al. 2004; Traina et al. 2006; Ferreira et al. 2007; Ferreira et al. 2009]. Significant requirements for supporting similarity queries in real DBMS, such as working on the results of sub-queries that do not involve the key of the relation, were developed using SIREN [Razente et al. 2008; Razente et al. 2007].

3.3 Agroclimate Data

Studies about Climate Changes is an important research target that provides tough problems for the database community due to, among other reasons, the huge amount of multi-dimensional data being collected and generated by simulation processes. As an example, simulation of future climate scenarios can produce billions of records including over forty climate parameters, generated by processes that take weeks to be computed in large super-computers but that take years to be analyzed. Storing and retrieving such multi-dimensional data can easily lead the most resilient DBMS to a sluggish behavior.

We have developed Fractal-based analysis techniques to analyze multiple climate-related time series, in order to identify intrinsic temporal patterns and trend changes. Treating multiple time series as a single multidimensional data stream and combining fractal-based analysis with clustering, allow us to integrate different climate variables in a single analytical process. This approach is distinct from the usual analysis that are being conducted by the other research groups in general, because it requires dealing with much more data at the same time. However, it has been shown that in this way it is possible to discover and anticipate more elusive general behavior changes over time, which can conduct to further improvements in the simulation models [Nunes et al. 2011]. The CLIPSMiner tool discretizes time series and uses concepts from the Fractal Theory to look for patterns of correlation changes in the data streams [Romani et al. 2010]. We have proposed also a methodology based on data clustering to analyze NDVI time series in order to monitor the growing cycles of sugar cane crops throughout the years. The NDVI images were obtained from AVHRR/NOAA satellites that have low spectral resolution, but can be used to monitor agricultural crops grown in large areas such as sugar cane [Romani et al. 2011]. In order to make it available to the specialists in agro-meteorology and climatology, our group developed the SatImageExplorer tool [Chino et al. 2010], which builds maps of spectral correlation changes from sets of images regarding specific regions and parameters. It also integrates clustering and correlation analysis and visualization resources, providing a complete workstation software for the climate change specialists to analyze.

The research on climate changes and agro-meteorology has been conducted in the database area with a strong partnership with EMBRAPA, CEPAGRI, and CPTEC, important research centers regarding

meteorology in Brazil, targeting the use of fractal-based techniques to speed up analysis processes and to execute multi-resolution analysis. Preliminary but promising results were published in several forums [Romani et al. 2010; Romani et al. 2010; Traina et al. 2010; Romani et al. 2010; Romani et al. 2013].

4. PUBLIC MATERIAL

As a research group supported by public funds, GBdI releases all the tools that it develops as free software, available at its site (<http://www.gbdi.icmc.usp.br/>). Every software piece is distributed as source code (mostly written in C or C++) and (where it applies) as executable tools for Linux and Windows operating systems. The following software packages have been released by GBdI:

- SIREN – the SQL extender for similarity queries over PostgreSQL and Oracle;
- Arboretum – a library of metric access methods (developed both in GBdI and abroad), including the Slim-tree, DBM-Tree, DF-Tree, M-Tree, VP-Tree, GH-Tree, among others. Related libraries of image feature extractors and distance functions are also available;
- FastMapDB – a tool to visualize metric and multi-dimensional datasets;
- MAM-View – a tool to visualize MAM execution;
- MetricSplat – a tool that combines visualization and content-based retrieval methodologies;
- VisTree – an analytical environment where multiple workspaces allow joining data in parallel and in sequential visual queries;
- GMine – a system for simple and hierarchical graph analysis performed visually and interactively;
- VSTree – a tool to execute and visualize the suffix tree’s construction and search algorithms.
- DBGen – a tool to generate synthetic data aiming at evaluating algorithms. It can provide datasets following many distributions, including fractal ones;
- VisualTPCH tool – a tool to generate synthetic data for data warehouses.

5. CONCLUSION

The goal of this paper was to summarize the main research issues that the Images and Data Bases Group (*Grupo de Bases de Dados e de Imagens* - GBdI) with the Institute of Mathematics and Computer Sciences (ICMC) – USP at São Carlos, have been pursuing so far. It describes several of the achievements obtained in its broad objective of *creating techniques to improve storage and retrieval of complex data in large databases of scientific applications, aiming at transforming relational DBMSs into practical tools to support scientific activities that handle very large sets of complex data in applications such as medicine, biology, and meteorology*. In fact, using the Fractal Theory and the Metric Spaces Theory as conceptual bases, we have succeeded to:

- Create technologies and algorithms that enable including new complex data types, such as images, audio and time series and its corresponding retrieval operations, extending the Relational Algebra to target similarity-based predicates;
- Develop techniques for indexing, selectivity and cost estimation, optimization and memory management, so to include similarity queries as first-class comparison operators in SQL;
- Develop techniques to ease the exploration of the complex, big data that are being stored in DBMSs, providing techniques and tools for information visualization, data warehousing and data mining.

Now, as several of our original objectives are being fulfilled and effective and efficient tools are available, we are conducting our studies towards more “user-centered” perspectives. For example, efficient operators exist to perform similarity comparisons inside the DBMS and in the analysis tools, but how to algorithmically define what is similar for the specialists, i.e., for a physician, a biologist,

a meteorologist? How to define the distance functions, how to define the precise features that must be extracted and compared from images of medical exams or from satellite imagery? Are range queries and k -nearest neighbors queries really enough, or at least suitable, to meet all the queries the users want to ask? Are the big databases collected from several sources homogeneously represented and interpreted regardless of its provenance? Are the visualizations provided about them sufficiently representative of the data and adequate for the users' intent?

As we see people using the tools we developed, following their expectations and receiving their positive and negative impressions, we conclude that the techniques we developed are, indeed, helpful; nevertheless, we also note that much more remains to be developed. Therefore, we are working now on techniques that target the previous questions. Among others, we are working on techniques to: extract relevant features from images of medical exams of specific modalities and from satellite images; reduce the amount of data to be analyzed using feature selection, clustering, and grouping; developing fractal-based algorithms to characterize data distribution and operator selectivity; developing a more user-tailored extended set of similarity operations, such as including variety (or excluding excessive similarity) in the similarity answers, working on operations that include multiple similarity functions, and extending the unary operators to process joins and aggregate operations; and developing a unifying model for complex data visualization. All of those subjects are very motivating and can lead to further research and development activities that will certainly benefit the whole society.

REFERENCES

- ANNIBAL, L. P., FELIPE, J. C., CIFERRI, C. D. A., AND CIFERRI, R. R. iCube: A similarity-based data cube for medical images. In *Proceedings of the 23rd IEEE International Symposium on Computer-Based Medical Systems*. IEEE Computer Society, Perth, Australia, pp. 321–326, 2010.
- ARANTES, A. S., VIEIRA, M. R., TRAINA, A. J. M., AND TRAINA, CAETANO, J. The fractal dimension making similarity queries more efficient. In *Second Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches (in conjunction with 9th ACM International Conference on Knowledge Discovery and Data Mining)*. ACM Press, Washington, DC, pp. 12–17, 2003.
- ARANTES, A. S., VIEIRA, M. R., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Efficient algorithms to execute complex similarity queries in rdbms. *Journal of The Brazilian Computer Society* 3 (9): 5–24, 2004.
- AZEVEDO-MARQUES, P. M. D., ROSA, N. A., TRAINA, A. J. M., TRAINA, CAETANO, J., KINOSHITA, S. K., AND RANGAYAN, R. M. Reducing the semantic gap in content-based image retrieval in mammography with relevance feedback and inclusion of expert knowledge. *International Journal of Computer Assisted Radiology and Surgery* 3 (1-2): 123–130, 2008.
- BAIOCO, G. B., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Mamcost: Global and local estimates leading to robust cost estimation of similarity queries. In *19th International Conference on Scientific and Statistical Database Management*. ACM Press, Banff, Canada, pp. 6–16, 2007.
- BALAN, A. G. R., TRAINA, A. J. M., RIBEIRO, M. X., AZEVEDO-MARQUES, P., AND JR, C. T. Smart histogram analysis applied to the skull-stripping problem in t1-weighted mri. *Computers in Biology and Medicine* 42 (5): 509–522, 2012.
- BALAN, A. G. R., TRAINA, A. J. M., TRAINA, CAETANO, J., AND MARQUES, P. M. D. A. Fractal analysis of image textures for indexing and retrieval by content. In *18th IEEE Intl. Symposium on Computer-Based Medical Systems*. IEEE Computer Society, Dublin, Ireland, pp. 581 – 586, 2005.
- BARIONI, M. C. N., KASTER, D. D. S., RAZENTE, H. L., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Querying multimedia data by similarity in relational dbms. In *Advanced Database Query Systems: Techniques, Applications and Technologies*, L. Yan and Z. Ma (Eds.). IGI Global, Hershey, NY, USA, pp. 323–359, 2010.
- BARIONI, M. C. N., RAZENTE, H. L., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Siren: A similarity retrieval engine for complex data. In *Demo section of the 32nd Intl Conference on Very Large Data Bases*, U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim (Eds.). ACM Press, Seoul, South Korea, pp. 1155–1158, 2006.
- BARIONI, M. C. N., RAZENTE, H. L., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Accelerating k-medoid-based algorithms through metric access methods. *Journal of Systems and Software* 81 (3): 343–355, 2008.
- BARIONI, M. C. N., RAZENTE, H. L., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Seamlessly integrating similarity queries in sql. *Software: Practice and Experience* 39 (4): 355–384, 2009.
- BARIONI, M. C. N., RAZENTE, H. L., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Visually mining on multiple relational tables at once. In *Advances in Databases and Information Systems*, Y. Manolopoulos and P. Návrat (Eds.). Vol. 2. Bratislava, Slovakia, pp. 21–30, 2002.

- BARIONI, M. C. N., RAZENTE, H. L., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Querying complex objects by similarity in sql. In *Simpósio Brasileiro de Banco de Dados*, C. A. Heuser and S. A. de Amo (Eds.). Vol. 1. SBC, Uberlândia, MG, pp. 130–144, 2005.
- BRITO, J. J., SIQUEIRA, T. L. L., TIMES, V. C., CIFERRI, R. R., AND DE CIFERRI, C. D. Efficient processing of drill-across queries over geographic data warehouses. In *Proceedings of the 13th international conference on Data warehousing and knowledge discovery*. DaWaK'11. Springer-Verlag, Berlin, Heidelberg, pp. 152–166, 2011.
- BUENO, R., KASTER, D. D. S., PATERLINI, A. A., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Unsupervised scaling of multi-descriptor similarity functions for medical image datasets. In *22th IEEE Intl. Symposium on Computer-Based Medical Systems*. IEEE Computer Society, Albuquerque, NM, EUA, pp. 1–8, 2009.
- BUENO, R., KASTER, D. D. S., TRAINA, A. J. M., AND TRAINA, CAETANO, J. A new approach for optimization of dynamic metric access methods using an algorithm of effective deletion. In *20th International Conference on Scientific and Statistical Database Management*. Vol. 5069/2008. Springer Berlin / Heidelberg, Hong Kong S.A.R., China, pp. 366–383, 2008.
- BUENO, R., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Genetic algorithms for approximate similarity queries. *Data and Knowledge Engineering* 62 (3): 459–482, 2007.
- CARELO, C. C. M., POLA, I. R. V., CIFERRI, C. D. A., CIFERRI, R. R., TRAINA, CAETANO, J., AND TRAINA, A. J. M. The onion-tree: Quick indexing of complex data in the main memory. In *Advances in Databases and Information Systems*. Springer, Riga, Latvia, pp. 235–252, 2009.
- CARELO, C. C. M., POLA, I. R. V., CIFERRI, R. R., TRAINA, A. J. M., TRAINA, CAETANO, J., AND CIFERRI, C. D. A. Slicing the metric space to provide quick indexing of complex data in the main memory. *Information Systems Journal* 36 (1): 79–98, 2011.
- CECCHIN, F., CIFERRI, C. D. A., AND HARA, C. S. XML data fusion. In *Proceedings of the 12th International Conference on Data Warehousing and Knowledge Discovery*, T. B. Pedersen, M. K. Mohania, and A. M. Tjoa (Eds.). Lecture Notes in Computer Science, vol. 6263. Springer, Bilbao, Spain, pp. 297–308, 2010.
- CHINO, D. Y. T., LOUZA, F. A., TRAINA, A. J. M., CIFERRI, C. D. A., AND JR., C. T. Time series indexing taking advantage of the generalized suffix tree. *Journal of Information and Data Management* 3 (1): 101–109, 2012.
- CHINO, D. Y. T., ROMANI, L. A. S., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Construindo séries temporais de imagens de satélite para sumarização de dados climáticos e monitoramento de safras agrícolas. *Revista Eletrônica de Iniciação Científica* 10 (3): 1–20, 2010.
- CIFERRI, R. R., CIFERRI, C. D. A., CARÉLO, C. C. M., AND TRAINA, CAETANO, J. nsP-index: A robust and persistent index for nucleotide sequences. In *Proceedings of the 12th East European Conference*, P. Atzeni, A. Caplinskias, and H. Jaakkola (Eds.). Tampere University of Technology, Pori, Finland, pp. 28–41, 2008.
- CORDEIRO, R. L. F., GUO, F., HAVERKAMP, D. S., HORNE, J. H., HUGHES, E. K., KIM, G., TRAINA, A. J. M., TRAINA, CAETANO, J., AND FALOUTSOS, C. Qmas: Querying, mining and summarization of multi-modal databases. In *10th IEEE International Conference on Data Mining*, C. Zhang, D. Gunopulos, G. Webb, and B. Liu (Eds.). IEEE Computer Society, Sydney, NSW, Australia, pp. 785 – 790, 2010.
- CORDEIRO, R. L. F., TRAINA, A. J. M., FALOUTSOS, C., AND JR., C. T. Halite: Fast and scalable multi-resolution local-correlation clustering. *IEEE Transactions on Knowledge and Data Engineering* 25 (2): 387–401, 2013.
- CORDEIRO, R. L. F., TRAINA, A. J. M., FALOUTSOS, C., AND TRAINA, CAETANO, J. Finding clusters in subspaces of very large, multi-dimensional datasets. In *26th IEEE International Conference on Data Engineering*, U. Dayal and V. J. Tsotras (Eds.). IEEE Computer Society, Long Beach, California, pp. 625–636, 2010.
- DOMINGUES, G. R., CIFERRI, C. D. A., AND CIFERRI, R. R. VisualTPCH: Uma ferramenta para a geração de dados sintéticos para data warehouse. In *Proceedings of the Demos Session of the Brazilian Symposium on Databases*. SBC, Campinas, São Paulo, Brazil, pp. 31–36, 2008.
- FEDEL, G. D. S., RAZENTE, H. L., TRAINA, CAETANO, J., AND BARIONI, M. C. N. Uma ferramenta para visualização em sgbdrs com uma implementação interativa do algoritmo para detecção de agrupamentos k-medoid. In *3ª Sessão de Demos em Banco de Dados, ocorrido junto ao 23º Simpósio Brasileiro de Bases de Dados*. Vol. 1. SBC, Campinas, SP, pp. 7–12, 2008.
- FELIPE, J. C., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Retrieval by content of medical images using texture for tissue identification. In *16th IEEE Symposium on Computer-based Medical Systems*. IEEE Computer Society, New York, pp. 175–180, 2003.
- FELIPE, J. C., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Perceptual distance functions for similarity retrieval of medical images. In *5th Intl. Conference on Image and Video Retrieval (CIVR 2006)*. Lecture Notes in Computer Science (LNCS), vol. 4071. Springer Verlag, Tempe, AZ, USA, pp. 432–442, 2006.
- FELIPE, J. C., TRAINA, A. J. M., AND TRAINA, CAETANO, J. A new family of distance functions for perceptual similarity retrieval of medical images. *Journal of Digital Imaging* 22 (2): 183–201, 2008.
- FERREIRA, M. R. P., PONCIANO DA SILVA, M., TRAINA, A. J. M., TRAINA, CAETANO, J., AMO, S. A. D., PEREIRA, F. S., AND CHBEIR, R. Integrating user preference to similarity queries over medical images datasets. In *International Symposium on Computer-Based Medical Systems*. IEEE Computer Society, Perth, Australia, pp. 1–6, 2010.

- FERREIRA, M. R. P., RIBEIRO, M. X., TRAINA, A. J. M., CHBEIR, R., AND TRAINA, CAETANO, J. Adding knowledge extracted by association rules into similarity queries. *Journal of Information and Data Management* 1 (3): 391–406, 2010.
- FERREIRA, M. R. P., TRAINA, A. J. M., DIAS, I., CHBEIR, R., AND TRAINA, CAETANO, J. Identifying algebraic properties to support optimization of unary similarity queries. In *The III Alberto Mendelzon International Workshop on Foundations of Data Management*, M. Arenas and L. Bertossi (Eds.). CEUR Workshop Proceedings (<http://ceur-ws.org>), Arequipa, Peru., pp. 10 pages., 2009.
- FERREIRA, M. R. P., TRAINA, CAETANO, J., AND TRAINA, A. J. M. An efficient framework for similarity query optimization. In *15th ACM International Symposium on Advances in Geographic Information Systems (ACM GIS 2007)*. Seattle, Washington, pp. 396–39, 2007.
- FERREIRA CORDEIRO, R. L., TRAINA, JUNIOR, C., MACHADO TRAINA, A. J., LÓPEZ, J., KANG, U., AND FALOUTSOS, C. Clustering very large multi-dimensional datasets with mapreduce. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '11. ACM, New York, NY, USA, pp. 690–698, 2011.
- KASTER, D. D. S., BUENO, R., BUGATTI, P. H., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Incorporating metric access methods for similarity searching on oracle database. In *24º Simpósio Brasileiro de Bases de Dados, Applications and Experiences Track*. Vol. 1. SBC, Fortaleza, CE, pp. 196–210, 2009.
- KASTER, D. D. S., BUGATTI, P. H., PONCIANO DA SILVA, M., TRAINA, A. J. M., MARQUES, P. M. D. A., AND TRAINA, CAETANO, J. Medfmi-sir: A powerful dbms solution for large-scale medical image retrieval. In *2nd International Conference on Information Technology in Bio- and Medical Informatics (ITBAM 2011)*, C. Böhm (Ed.). Lecture Notes in Computer Science (LNCS). Springer Verlag, Toulouse, France, pp. 16–30, 2011.
- KASTER, D. D. S., BUGATTI, P. H., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Fmi-sir: A flexible and efficient module for similarity searching on oracle database. *Journal of Information and Data Management* 1 (2): 229–244, 2010.
- KATAYAMA, N. AND SATOH, S. Distinctiveness-sensitive nearest-neighbor search for efficient similarity retrieval of multimedia information. In *IEEE International Conference on Data Engineering*, D. Georgakopoulos and A. Buchmann (Eds.). IEEE Computer Society, Heidelberg, Germany, pp. 493–502, 2001.
- LOUZA, F. A., BRITO JUNIOR, J. S., CIFERRI, R. R., AND CIFERRI, C. D. A. VSTree: Uma ferramenta de execução e visualização de algoritmos de construção e busca em árvores de sufixo. In *Proceedings of the Demos Session of the Brazilian Symposium on Databases*. SBC, pp. 7–12, 2010.
- LOUZA, F. A., CIFERRI, R. R., AND CIFERRI, C. D. A. Efficiently querying protein sequences with the Proteinus index. In *Proceedings of the 2011 Brazilian Symposium on Bioinformatics. To appear*. Lecture Notes in Bioinformatics. Springer, Brasília, DF, Brazil, pp. 58–65, 2011.
- MATEUS, R. C., SIQUEIRA, T. L. L., TIMES, V. C., CIFERRI, R. R., AND CIFERRI, C. D. A. How does the spatial data redundancy affect query performance in geographic data warehouses. *Journal of Information and Data Management* 1 (3): 519–534, 2010.
- NUNES, S., ROMANI, L. A. S., ÁVILA, A. A. M. H. D., TRAINA, CAETANO, J., SOUSA, E. P. M. D., AND TRAINA, A. J. M. Fractal-based analysis to identify trend changes in multiple climate time series. *Journal of Information and Data Management* 2 (1): 51–58, 2011.
- PAN, J.-Y., BALAN, A. G. R., TRAINA, A. J. M., FALOUTSOS, C., AND XING, E. P. Automatic mining of fruit fly embryo images. In *12th ACM International Conference on Knowledge Discovery and Data Mining (KDD2006)*. ACM Press, Philadelphia, PA, pp. 693–698, 2006.
- PATERLINI, A. A., DE FARIA, R. F. T., RAZENTE, H. L., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Fastmapdb: Uma ferramenta para visualização em sgbdns com suporte à filtragem e seleção visual dos dados. In *2ª Sessão de Demos em Banco de Dados - junto ao 20 Simpósio Brasileiro de banco de Dados*, n. Brayner and C. F. Dorneles (Eds.). Uberlândia, MG, pp. 1–6, 2005.
- POLA, I. R. V., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Easing the dimensionality curse by stretching metric spaces. In *21st International Conference on Scientific and Statistical Database Management*, M. Winslett (Ed.). Lecture Notes in Computer Science, vol. 5566. Springer, New Orleans, LA, pp. 417–434, 2009.
- RAZENTE, H. L., BARIONI, M. C. N., TRAINA, A. J. M., FALOUTSOS, C., AND TRAINA, CAETANO, J. A novel optimization approach to efficiently process aggregate similarity queries in mam. In *ACM 17th International Conference on Information and Knowledge Management*, J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K.-S. Choi, and A. Chowdhury (Eds.). ACM Press, Napa Valley, CA, pp. 193–202, 2008.
- RAZENTE, H. L., BARIONI, M. C. N., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Consultas por similaridade agregada em métodos de realimentação de relevância para consultas por conteúdo de imagens. In *Sessão de Pôsteres do 22º Simpósio Brasileiro de Bases de Dados*. João Pessoa, PB, pp. 3–6, 2007.
- RAZENTE, H. L., CHINO, F. J. T., BARIONI, M. C. N., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Visual analysis of feature selection for data mining processes. In *Simpósio Brasileiro de Banco de Dados*, S. Lifschitz (Ed.). Vol. 1. SBC, Brasília, DF, pp. 163–177, 2004.

- RIBEIRO, M. X., BALAN, A. G. R., FELIPE, J. C., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Mining statistical association rules to select the most relevant medical image features. In *Mining Complex Data*, D. A. Zighed, S. Tsumoto, Z. W. Ras, and H. Hacid (Eds.). Studies in Computational Intelligence, vol. 165/2009. Springer Berlin / Heidelberg, Berlin / Heidelberg, pp. 113–131, 2009.
- RIBEIRO, M. X., BUGATTI, P. H., TRAINA, A. J. M., TRAINA, CAETANO, J., MARQUES, P. M. D. A., AND ROSA, N. A. Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques. *Data and Knowledge Engineering* 68 - Special issue on Knowledge Discovery in Medicine (12): 1370–1382, 2009.
- RIBEIRO, M. X., TRAINA, A. J. M., BALAN, A. G. R., TRAINA, CAETANO, J., AND MARQUES, P. M. D. A. Sugar: Association-rules based framework to support diagnosis of abnormalities in mammograms. In *20th IEEE Intl Symposium on Computer-Based Medical Systems*. Vol. 1. IEEE Computer Society, Maribor, Slovenia, pp. 47–52, 2007.
- RIBEIRO, M. X., TRAINA, A. J. M., TRAINA, CAETANO, J., AND MARQUES, P. M. D. A. An efficient association rule-based method to support medical image diagnosis. *IEEE Transactions on Multimedia* 10 (2): 277 – 285, 2008.
- RODRIGUES JR, J. F., BALAN, A. G. R., TRAINA, A. J. M., AND TRAINA, CAETANO, J. The visual expression process: Bridging vision and data visualization. In *9th International Symposium on Smart Graphics (SG 2008)*. Vol. 5166/2008. Springer Berlin / Heidelberg, Rennes, France, pp. 207–215, 2008.
- RODRIGUES JR, J. F., CIRILO, C. E., ZAINA, L. A. M., AND PRADO, A. F. Hierarchical visual filtering, pragmatic and epistemic actions for database visualization. In *Proceedings of the ACM Symposium on Applied Computing*. ACM Press, pp. to appear, 2013.
- RODRIGUES JR, J. F., ROMANI, L. A., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Combining visual analytics and content based data retrieval technology for efficient data analysis. In *14th International Conference Information Visualization - IV'10*, E. Banissi, S. Bertschi, R. A. Burkhard, J. Counsell, M. Dastbaz, M. J. Eppler, C. Forsell, G. G. Grinstein, J. Johansson, M. Jern, F. Khosrowshahi, F. T. Marchese, C. Maple, R. Laing, U. Cvek, M. Trutschl, M. Sarfraz, L. J. Stuart, A. Ursyn, and T. G. Wyeld (Eds.). London, England, pp. 61–67, 2010.
- RODRIGUES JR, J. F., ROMANI, L. A. S., ZAINA, L., AND CIFERRI, R. Metricsplat - a platform for quick development, testing and visualization of content-based retrieval techniques. In *Simposio Brasileiro de Bancos de Dados - SBBD2009*. pp. 1–6, 2009.
- RODRIGUES JR, J. F., TONG, H., TRAINA, A. J. M., FALOUTSOS, C., AND LESKOVEC, J. Gmine: A system for scalable, interactive graph visualization and mining. In *Demo section of the 32nd Intl Conference on Very Large Data Bases*, U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim (Eds.). ACM Press, Seoul, South Korea, pp. 1195–1198, 2006.
- RODRIGUES JR, J. F., TONG, H., YU PAN, J., TRAINA, A. J. M., JR., C. T., AND FALOUTSOS, C. Large graph analysis in the gmine system. *IEEE Transactions on Knowledge and Data Engineering* 25 (1): 106–118, 2013.
- RODRIGUES JR, J. F., TRAINA, A. J. M., FALOUTSOS, C., AND TRAINA, CAETANO, J. Supergraph visualization. In *IEEE International Symposium on Multimedia (ISM2006)*. IEEE Computer Society, San Diego - CA, pp. 227–234, 2006.
- RODRIGUES JR, J. F., TRAINA, A. J. M., OLIVEIRA, M. C. F. D., AND TRAINA, CAETANO, J. The spatial/perceptual design space: a new comprehension for data visualization. *Information Visualization Journal* 6 (4): 261–279, 2007.
- RODRIGUES JR, J. F., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Visualization tree, multiple linked analytical decisions. In *5th International Symposium on Smart Graphics, SG 2005*, A. Butz, B. Fisher, A. Krüger, and P. Olivier (Eds.). Lecture Notes in Computer Science, vol. 3638. Springer-Verlag Press, Frauenwörth Cloister, Germany, pp. 65 – 76, 2005.
- RODRIGUES JR, J. F., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Mining frequent patterns for visual interpretation in a multi-view environment. In *Visual Data Mining: Theory, Techniques and Tools*, S. J. Simoff, M. H. Böhlen, and A. Mazeika (Eds.). LNCS State of the Art Surveys, vol. 4404. Springer-Verlag Press, Heidelberg, Germany, pp. 50–69, 2008.
- ROMANI, L., TRAINA, A. J. M., JR., C. T., CHBEIR, R., AVILA, A. M. H., AND JR, J. Z. A new time series mining approach applied to multitemporal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* 51 (1): 140–150, 2013.
- ROMANI, L. A., SOUSA, E. P. M. D., RIBEIRO, M. X., ÁVILA, A. A. M. H. D., ZULLO JR., J., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Mining climate and remote sensing time series to improve monitoring of sugar cane fields. In *Computational Methods Applied to Agricultural Research: Advances and Applications*, H. A. d. Prado, A. J. B. Luiz, and H. C. Filho (Eds.). Springer Verlag, 2010.
- ROMANI, L. A., SOUSA, E. P. M. D., RIBEIRO, M. X., ZULLO JR., J., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Employing fractal dimension to analyze climate and remote sensing data streams. In *First SIAM SDM Workshop on Multimedia Data Mining (MDM/SDM 2009)*. Vol. 1. SIAM, Sparks, Nevada, pp. 5–16, 2009.
- ROMANI, L. A., TRAINA, A. J. M., TRAINA, CAETANO, J., CHBEIR, R., ÁVILA, A. A. M. H. D., AND ZULLO JR., J. New dtw-based method to similarity search in sugar cane regions represented by climate and remote sensing time

- series. In *10th IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2010)*, pp. 4 pags, 25-30 de julho de 2010, Honolulu, USA. Honolulu, USA, pp. 25-30, 2010.
- ROMANI, L. A. S., GONÇALVES, R. R. V., AMARAL, B. F., CHINO, D. Y. T., ZULLO JR., J., TRAINA, CAETANO, J., SOUSA, E. P. M. D., AND TRAINA, A. J. M. Clustering analysis applied to ndvi/noaa multitemporal images to improve the monitoring process of sugarcane crops. In *Sixth International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp-2011)*. Trento, Italy, pp. 1-4, 2011.
- ROMANI, L. A. S., ÁVILA, A. M. H. D., ZULLO JR., J., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Mining relevant and extreme patterns on climate time series with clipsminer. *Journal of Information and Data Management* 1 (2): 245-260, 2010.
- ROMERO, R. A. F., VICENTINI, J. F., OLIVEIRA, P. R., AND TRAINA, A. J. M. Investigating the potential of art neural network models for indexing and information retrieval. *International Journal of Intelligent Systems* 22 (4): 219-336, 2007.
- ROSA, N. A., SANTOS FILHO, R. F., BUENO, J. M., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Sistema de recuperação de imagens similares em um hospital universitário. In *VIII Congresso Brasileiro de Informática em Saúde*. Natal, RN-Brazil, 2002.
- SERAPHIM, E., SERAPHIM, T. F. P., MOREIRA, E. M., RICOTTA, F. C., AND JR., C. T. Paged similarity queries. *Information Sciences* 181 (13): 2600-2607, 2011.
- SIQUEIRA, T. L. L., CIFERRI, C. D. A., TIMES, V. C., AND CIFERRI, R. R. The SB-index and the HSB-index: efficient indices for spatial data warehouses. *Geoinformatica*, 2011.
- SIQUEIRA, T. L. L., CIFERRI, C. D. A., TIMES, V. C., OLIVEIRA, A. G., AND CIFERRI, R. R. The impact of spatial data redundancy on SOLAP query performance. *Journal of the Brazilian Computer Society* 15 (2): 19-34, 2009.
- SIQUEIRA, T. L. L., CIFERRI, R. R., TIMES, V. C., AND CIFERRI, C. D. A. Benchmarking spatial data warehouses. In *Proceedings of the 12th International Conference on Data Warehousing and Knowledge Discovery*, T. B. Pedersen, M. K. Mohania, and A. M. Tjoa (Eds.). Lecture Notes in Computer Science, vol. 6263. Springer, Bilbao, Spain, pp. 40-51, 2010.
- SIQUEIRA, T. L. L., MATEUS, R. C., CIFERRI, R. R., TIMES, V. C., AND CIFERRI, C. D. A. Querying vague spatial information in geographic data warehouses. In *Advancing Geoinformation Science for a Changing World*, S. Geertman, W. Reinhardt, and F. Toppen (Eds.). Lecture Notes in Geoinformation and Cartography, vol. 1. Springer Berlin Heidelberg, Utrecht, The Netherlands, pp. 379-397, 2011.
- SOUSA, E. P. M. D., TRAINA, CAETANO, J., TRAINA, A. J. M., AND FALOUTSOS, C. How to use fractal dimension to find correlations between attributes. In *First Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches (in conjunction with 8th ACM International Conference on Knowledge Discovery and Data Mining)*. ACM Press, Edmonton, Alberta, Canada, pp. 26-30, 2002.
- SOUSA, E. P. M. D., TRAINA, CAETANO, J., TRAINA, A. J. M., WU, L., AND FALOUTSOS, C. A fast and effective method to find correlations among attributes in databases. *Data Mining and Knowledge Discovery* 14 (3): 367-407, 2007.
- TOMAZELA, B., CIFERRI, C. D. A., AND TRAINA, CAETANO, J. Reconciliando dados de cunho acadêmico. In *Proceedings of the Brazilian Symposium on Databases*, S. de Amo (Ed.). SBC, Campinas, São Paulo, Brazil, pp. 283-297, 2008.
- TOMAZELA, B., HARA, C. S., CIFERRI, R. R., AND CIFERRI, C. D. A. PrInt: a provenance model to support integration processes. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, J. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, and A. An (Eds.). ACM, Toronto, Ontario, Canada, pp. 1349-1352, 2010.
- TRAINA, A. J. M., ROMANI, L. A., CORDEIRO, R. L. F., SOUSA, E. P. M. D., RIBEIRO, M. X., ÁVILA, A. A. M. H. D., ZULLO JR., J., RODRIGUES JR., J. F., AND TRAINA, CAETANO, J. How to find relevant patterns in climate data: an efficient and effective framework to mine climate time series and remote sensing images. In *SIAM Annual Meeting 2010*, B. L. Keyfitz and L. N. Trefethen (Eds.). SIAM, Pittsburgh, PA, pp. 124-125, 2010.
- TRAINA, A. J. M., TRAINA, CAETANO, J., BALAN, A. G. R., RIBEIRO, M. X., BUGATTI, P. H., WATANABE, C. Y. V., AND AZEVEDO-MARQUES, P. M. D. Feature extraction and selection for decision making over medical images. In *Biomedical Image Processing - Methods and Applications*, T. M. Deserno (Ed.). Springer-Verlag, pp. 181-209, 2010.
- TRAINA, A. J. M., TRAINA, CAETANO, J., BARIONI, M. C. N., BOTELHO, E., AND BUENO, R. Visualização de dados em sistemas de bancos de dados relacionais. In *Simpósio Brasileiro de Bancos de Dados*, M. L. d. Q. Mattoso and G. Xexéo (Eds.). SBC, Rio de Janeiro, RJ, pp. 95-109, 2001.
- TRAINA, A. J. M., TRAINA, CAETANO, J., BUENO, J. M., CHINO, F. J. T., AND MARQUES, P. M. D. A. Efficient content-based image retrieval through metric histograms. *World Wide Web Journal (WWWJ)* 6 (2): 157-185, 2003.
- TRAINA, A. J. M., TRAINA, CAETANO, J., BUENO, J. M., AND MARQUES, P. M. D. A. The metric histogram: A new and efficient approach for content-based image retrieval. In *Sixth IFIP Working Conference on Visual Database Systems*, X. Zhou and P. Pu (Eds.). IFIP Conference Proceedings, vol. 216. Kluwer Academic Publishers, Brisbane, Australia, pp. 297-311, 2002.

- TRAINA, A. J. M., TRAINA, CAETANO, J., CIFERRI, C. D. D. A., RIBEIRO, M. X., AND MARQUES, P. M. D. A. How to cope with the performance gap in content-based image retrieval systems. *International Journal of Healthcare Information Systems and Informatics (IJHISI)* 4 (1): 47–67, 2009.
- TRAINA, A. J. M., TRAINA, CAETANO, J., PAPADIMITRIOU, S., AND FALOUTSOS, C. Tri-plots: Scalable tools for multidimensional data mining. In *ACM International Conference on Knowledge Discovery and Data Mining*, F. Provost and S. Ramakrishnan (Eds.). ACM Press, San Francisco, CA, pp. 184–193, 2001.
- TRAINA, CAETANO, J., SANTOS FILHO, R. F., TRAINA, A. J. M., VIEIRA, M. R., AND FALOUTSOS, C. The omnifamily of all-purpose access methods: A simple and effective way to make similarity search more efficient. *The International Journal on Very Large Databases* 16 (4): 483–505, 2007.
- TRAINA, CAETANO, J., TRAINA, A. J. M., FALOUTSOS, C., AND SEEGER, B. Fast indexing and visualization of metric datasets using slim-trees. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 14 (2): 244–260, 2002.
- TRAINA, CAETANO, J., TRAINA, A. J. M., SANTOS FILHO, R. F., AND FALOUTSOS, C. How to improve the pruning ability of dynamic metric access methods. In *International Conference on Information and Knowledge Management*. ACM Press, McLean, VA, USA, pp. 219–226, 2002.
- TRAINA, CAETANO, J., TRAINA, A. J. M., SEEGER, B., AND FALOUTSOS, C. Slim-trees: High performance metric trees minimizing overlap between nodes. In *International Conference on Extending Database Technology (EDBT)*, C. Zaniolo, P. C. Lockemann, M. H. Scholl, and T. Grust (Eds.). Lecture Notes in Computer Science, vol. 1777. Springer Verlag, Konstanz, Germany, pp. 51–65, 2000.
- TRAINA, CAETANO, J., TRAINA, A. J. M., VIEIRA, M. R., ARANTES, A. S., AND FALOUTSOS, C. Efficient processing of complex similarity queries in rdbms through query rewriting. In *ACM 15th International Conference on Information and Knowledge Management*, P. S. Yu, V. J. Tsotras, E. A. Fox, and B. Liu (Eds.). ACM Press, Arlington - VA, USA, pp. 4–13, 2006.
- TRAINA, CAETANO, J., TRAINA, A. J. M., WU, L., AND FALOUTSOS, C. Fast feature selection using fractal dimension. In *Brazilian Symposium on Databases*, C. M. B. Medeiros and K. Becker (Eds.). SBC, João Pessoa, PB, pp. 158–171, 2000.
- TRAINA, CAETANO, J., TRAINA, A. J. M., WU, L., AND FALOUTSOS, C. Fast feature selection using fractal dimension - ten years later. *Journal of Information and Data Management* 1 (1): 17–20, 2010.
- VALÊNCIO, C. R., TRONCO, M. N., BONINI-DOMINGOS, A. C., BONINI-DOMINGOS, C. R., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Knowledge extraction using visualization of hemoglobin parameters to identify thalassemia. In *17th IEEE Symposium on Computer-Based Medical Systems*. Vol. 1. IEEE Computer Society, Bethesda, Maryland, pp. 523–528, 2004.
- VESPA, T. G., TRAINA, A. J. M., AND TRAINA, CAETANO, J. Efficient bulk-loading on dynamic metric access methods. *Information Systems Journal* 35 (5): 557–569, 2010.
- VESPA, T. G., TRAINA, CAETANO, J., AND TRAINA, A. J. M. Bulk-loading dynamic metric access methods. In *22º Simpósio Brasileiro de Bases de Dados*. Vol. 1. SBC, João Pessoa, PB, pp. 160–173, 2007.
- VIEIRA, M. R., CHINO, F. J. T., TRAINA, CAETANO, J., AND TRAINA, A. J. M. A visual framework to understand similarity queries and explore data in metric access methods. *Int. J. Business Intelligence and Data Mining* 5 (4): 370–387, 2010.
- VIEIRA, M. R., RAZENTE, H. L., BARIONI, M. C. N., HADJIELEFTHARIOU, M., SRIVASTAVA, D., TRAINA, CAETANO, J., AND TSOTRAS, V. J. Divdb: A system for diversifying query results", 37th IEEE international conference on very large databases. In *37th International Conference on Very Large Databases - Demo Section*, P. Larson, J. Patel, and M. Yoshikawa (Eds.). VLDB End., Seattle, WA, USA, pp. 1395–1398, 2011a.
- VIEIRA, M. R., RAZENTE, H. L., BARIONI, M. C. N., HADJIELEFTHARIOU, M., SRIVASTAVA, D., TRAINA, CAETANO, J., AND TSOTRAS, V. J. On query result diversification. In *27th IEEE International Conference on Data Engineering*. Hannover, Germany, pp. 1163–1174, 2011b.
- VIEIRA, M. R., TRAINA, CAETANO, J., CHINO, F. J. T., AND TRAINA, A. J. M. Dbm-tree: Trading height-balancing for performance in metric access methods. *Journal of The Brazilian Computer Society* 11 (3): 37–52, 2006.
- VIEIRA, M. R., TRAINA, CAETANO, J., CHINO, F. J. T., AND TRAINA, A. J. M. Dbm-tree: A dynamic metric access method sensitive to local density data. *Journal of Information and Data Management* 1 (1): 111–128, 2010.
- VIEIRA, M. R., TRAINA, CAETANO, J., TRAINA, A. J. M., ARANTES, A. S., AND FALOUTSOS, C. Estimating suitable query radii to boost knearest neighbor queries. In *19th International Conference on Scientific and Statistical Database Management*. ACM Press, Banff, Canada, pp. 1–10, 2007.
- VIEIRA, M. R., TRAINA, CAETANO, J., TRAINA, A. J. M., AND CHINO, F. J. T. Dbm-tree: A dynamic metric access method sensitive to local density data. In *Brazilian Symposium on Databases*, S. Lifschitz (Ed.). Vol. 1. SBC, Brasília, DF, pp. 33–47, 2004.