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, Robson Leonardo Cordeiro

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 the development of core databases and data analysis techniques to handle large databases for scientific applications involving complex data. This article describes GBdI's 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 being developed and released for the community in general, including tools for computer-assisted medical diagnosis and climate change analysis, libraries to aid in the development of efficient and effective applications handling data in multidimensional and metric spaces, software libraries to help including similarity queries into existing DBMS, and tools for image mining. This article describes the main research activities and the main results achieved, which are the basis for collaborative work with other groups in all over the world, aiming at the development of new technologies and applications that employ Database Management Systems 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 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 under development at IFQSC, where those ICMC faculty members 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. Prof. Agma Machado 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 retrieve images in databases, at those times a great research issue that was in its very beginning. Today, the faculties of GBdI are the six authors of this article.

During the period from 1980 to 1995, the research activities in databases in ICMC were strongly attached to the Physics Institute, as its main drive at those times was the research needs from two

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.

broad projects: high energy physics and the construction of a low field human body tomograph. Both projects involved the storage and retrieval of large amounts of data, and in the case of the tomograph, 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 School of Medicine of USP at Ribeirão Preto (FMRP). The CCIFM center is part of the Radiology Department of that school Hospital, and integrates the faculty and the health staff whom are in charge of the medical imaging equipment from the hospital, including XRay, Computerized Tomography (CT), mammography, ultra-sound and other equipments. Besides providing lots of test cases (as, for example, a database with the medical exam images obtained from October 2008 to December 2009 totaling 8 Terabytes in 14 Million images, in the many medical modalities and employing the several types of imaging machinery where this third care-level hospital operates), this partnership provided several intriguing research challenges. After Profs. Cristina Ciferri, Elaine Sousa, José Rodrigues Jr. and Robson Cordeiro have joined GBdI since 2004, the group expanded its investigation arena to other types of complex data and databases applications, such as digital libraries, climate, geographic and biological data, including multi-source data integration and provenance, and data warehousing, 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 Carnegie Mellon University (DBG-CMU), in Pittsburgh, PA, USA. Whereas 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, such as the Federal University of Uberlândia, the Federal University of Pernambuco, the Federal University of Sao Carlos and the Federal University of ABC among others, as well as with universities abroad, such as 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 to support the development of scientific applications requiring to handle 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 the development of techniques to execute similarity queries over complex data, such as multimedia data, multidimensional time series and biological sequences, and the second corresponds to the development of techniques to retrieve data represented through complex structures from several sources. All of those techniques always aim at providing the new resources in a relational database engine, and we have worked to create techniques that make the use of database management systems practical when it is required to support scientific activities that handles very large datasets in applications such as medicine, biology, and climate and agricultural analyzes. Our techniques allow

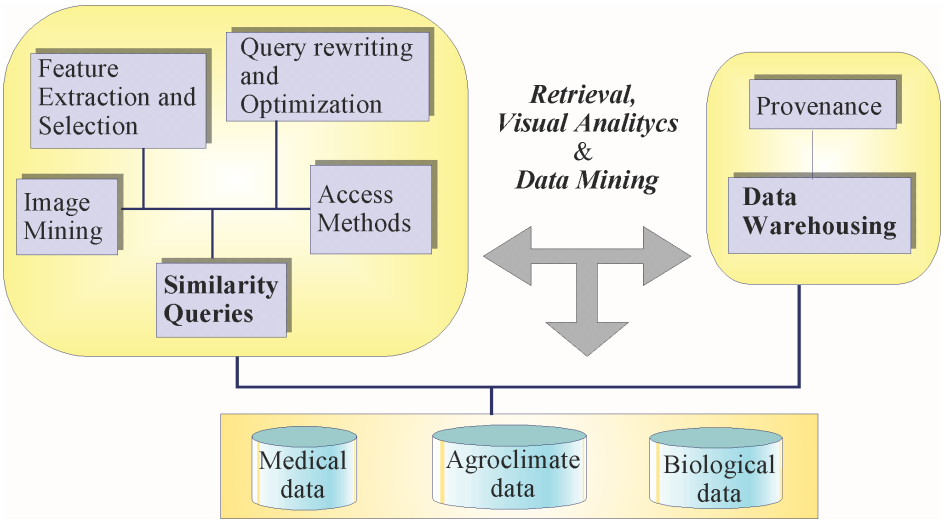

Fig. 1. The relationship among the diverse GBdI research areas.

integrating several data sources with distinct representations into a complete database over which data can be retrieved based on element similarities. Figure 1 depicts the main research areas and the relationships among them, as they are being studied at GBdI.

In the subsections following we present the main basic research activities where GBdI has accomplished interesting results. In the next section we present some of the applied results that benefited from those research activities.

2.1 Fractal Theory

A distinctive activity where GBdI has engaged is employing the 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 well beyond linear. As database operations should, ideally, execute within linear complexity cost regarding both time and number of disk accesses, alternatives must be discovered to reduce the complexity of the algorithm. 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 constant, small number of steps. In this way, we have derived fast algorithms to perform classification [Romero et al. 2007], clustering [Cordeiro$^{(1)}$ et al. 2010; Cordeiro et al. 2013] and association rule mining [Ribeiro$^{(1)}$ 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 [Baioco et al. 2007] and multi-dimensional data [Traina et al. 2001] as well as to estimate selectivity of spatial joins [Faloutsos et al. 2000], and to track the behavioral changes of data streams over time [Sousa et al. 2007].

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 on both cardinality and dimensionality of a multi-dimensional dataset. That algorithm was developed by our group and firstly presented at the Brazilian Database Symposium (SBBD) in 2000. It is worth to note that currently it is one of the most cited articles in the SBBD history [Traina Jr. et al. 2000; 2010]. The basic techniques that it first defined have been extended and was successfully applied to datasets from several domains, including climate analysis [Romani et al. 2009] and medical data [Ribeiro$^{(2)}$ 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 Jr. 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 Jr. 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 toward solving several of those issues too, such as developing algorithms for bulk-loading [Vespa et al. 2010], update and delete operations [Bueno et al. 2008], and support for paged similarity queries [Seraphim et al. 2011]. This research line provided also several other metric access methods, such as the DBM-tree [Vieira et al. 2010], the first and yet the only metric access method to explore adequate balancing among depth- and width-navigation over a tree to answer queries that require node overlap resolution; and the Onion-tree [Carelo et al. 2011], the first dynamic main memory indexing for metric data without overlapping among nodes.

Embedding a metric space into a multi-dimensional one was also explored, using the Fractal Theory as the guiding concept to evaluate both the embedding space dimensionality and the most important parameters required to tune the metric access methods. Thus, a number of techniques were developed to free the user from the burden of defining parameters that are otherwise obtainable only after several experimental evaluations [Traina Jr. et al. 2002; Traina Jr. 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 query types were developed to allow a query that specify 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. 2011].

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 efficiency and efficacy improvements – 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 seen to provide fewer improvement opportunities, they both can greatly benefit from the Fractal Theory to automatically tailor the search algorithms parameters 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), 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. 2003], which allows obtaining closely 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. 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 k-nearest neighbor queries for the much cheaper similarity range query [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 comply to. Thus, the ability to " see" the data is an important resource for data miners. Visualization of data derived from scientific activities is an even more valuable resource, as multiple attributes are often correlated in ways that are difficult for the annalist to seize. Therefore, people usually considers 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-dimensionality into a small dimensionality space (two or three dimensions), so that the embedded one can be rendered for visualization, involving data from one or more relations (multi-relational visualization [Barioni et al. 2002]). Several of those techniques where integrated in a tool called FastMapDB, which employs techniques to quickly map the data stored in a database management system into a 3D-space and provides several interactions mechanisms for the users to foster their understanding of the data. In another approach, we have designed a framework that enables the user to create multiple coordinated workspaces over the data, what improves the scalability of analysis over multivariate data analysis [Rodrigues Jr. et al. 2006; Rodrigues Jr. et al. 2007; Rodrigues Jr. et al. 2013], in tasks such as exploring frequent patterns, correlation and tendency prediction [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 the 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. One of the underpinning concepts employed to enable the mining algorithms to handle large amounts of data obviously was the Fractal Theory [Traina Jr. et al. 2005]. 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$^{(2)}$ et al. 2010; Barioni et al. 2008] and correlation clustering [Cordeiro$^{(1)}$ et al. 2010; Cordeiro et al. 2011; Cordeiro et al. 2013], classification [Romero et al. 2007; Pan et al. 2006], dimensionality reduction [Sousa et al. 2007; Silva et al. 2010], and association rule mining [Ribeiro et al. 2005; Ribeiro et al. 2006; Ribeiro$^{(1)}$ et al. 2009]. Our techniques are being applied on medical software to help computer-aided diagnosis [Ribeiro$^{(2)}$ et al. 2009; Ribeiro et al. 2008], on meteorological data for climate change analyzes [Romani et al. 2010], tracking of user intention [Ferreira$^{(1)}$ et al. 2010; Ferreira$^{(2)}$ 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, the data warehouse (DW). Besides being integrated, the data in the warehouse are subject-oriented and historical. Furthermore, aimed at improving the performance of queries target to perform On-line Analytical Processing (OLAP), a data warehouse 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 data warehouse, specially the fact that the

database stores historical information covering a long time horizon, may turn its volume into orders of magnitude larger than the corresponding one from its transactional counterpart.

At GBdI, we have conducted research related to improving analytical query performance over traditional and non-traditional data warehouses. We have investigated algorithms to perform horizontal and vertical fragmentation of data warehouses, and also proposed new data models for warehouses that store images, tackling both how intrinsic features of images should be stored in a data warehouse and the issue of how analytical similarity queries should be efficiently processed over a data warehouse of images [Annibal et al. 2010]. We have also obtained impressive results regarding Geographic data warehouses, which are 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 Spatial On-line Analytical Processing (SOLAP) query processing performance [Siqueira et al. 2009; Mateus et al. 2010]; proposed two efficient indexes for geographic data warehouses [Siqueira et al. 2012], 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 data warehouses using the SB-index [Brito et al. 2011]; developed a spatial benchmark for data warehouses to allow a controlled experimental performance evaluation of environments for spatial data warehouses [Siqueira et al. 2010]; and investigated how vague spatial data affects query performance and storage requirements in geographic data warehouses [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 on 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 provides the identification source and enables tracing the transformations applied to the data over its time life span. We employed data provenance aiming at reproducing the user's decisions, so the user's previous activities on 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 have also 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 another.

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 articles and several other information related to the biological data. The nucleotide sequences are represented as strings of the four nitrogen base types A, C, G and T, while the amino-acids chains are composed of strings that represent the 20 molecules

present in proteins.

At GBdI, we have focused on sequence similarity search over biological databases 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. 1012].

3. APPLIED RESEARCH ACTIVITIES

In this sections we present the most interesting results that extending our basic research activities over a number of application areas has produced.

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 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 later alternative, as we target large sets of, often confidential, images. Therefore, manually tagging a lot of images 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 [Kaster et al. 2011]. in this research area, our collaboration with the Center of Images and Medical Physics has generated important and rewarding results. We developed Computer-assisted diagnosis (CAD) tools to improve the efficacy of the medical exams analysis, thus benefiting the patients diagnosed/treated at the hospital.

The administrative systems of the hospital already employ Oracle, therefore we are developing new applications that store the hospital's image in this DBMS, creating an infrastructure for the health information system integrated to the administrative one. The new applications are based on the ability to execute queries that also include similarity queries, enabling the execution of procedures for similarity retrieval embedded in the database management system [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 developed 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 tools for computer-assisted diagnosis 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 was obtained analyzing blood tests of thalassemia. Thalassemia is a group of genetic blood disorders causing an anemia that begins in early childhood and 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 working on similarity queries, but few of them are studying how to integrate the individual results of the queries, considering that they always are 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 model. Nevertheless, having both resources is important to allow a tighter integration of the current wide spectrum of available database management systems 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; Barioni et al. 2009; Barioni et al. 2010]. SIREN is a software blade running between a conventional database management system 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 database management system to execute the conventional data operations.

SIREN allows the application programs to use queries requiring both unary (similarity range and k-nearest neighbors queries) and binary (range, k-nearest neighbor and k-closet neighbors joins) operators to be expressed, extends SQL to represent the corresponding similarity based-predicates, 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 database management systems, we have developed studies on query cost estimation [Baioco et al. 2007] and query optimization [Arantes et al. 2004; Traina Jr. et al. 2006; Ferreira et al. 2007; Ferreira et al. 2009]. Significant requirements for supporting similarity queries in real database management systems were identified throughout the development of SIREN. A significant milestone in this direction was the development of a technique to perform similarity operations on the results of sub-queries that do not involve the key of the relation [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 database management system 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 analysis 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 Normalized Difference Vegetation Index (NDVI) time series in order to monitor the growing cycles of sugar cane crops throughout the year. The NDVI images were obtained from the AVHRR/NOAA class 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 changes specialists analyzes.

The research on climate changes and agro-meteorology have been conducted in the database area with a strong partnership with EMBRAPA, CEPAGRI, and INPE-CPTEC, important research centers regarding meteorology in Brazil, targeting the use of fractal-based techniques to speedup analysis processes and to execute multi-resolution analysis. So far, 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 system. The following software packages have been released by GBdI:

—SIREN – the SQL extender for similarity queries over Postgres 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, etc. 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 metric access methods execution;
—MetricSPlat – a tool that combines visualization and content-based retrieval methodologies;
—VisTree – an analytical environment where multiple workspaces allows 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 building and searching algorithms.
—DBGen – a tool to generate synthetic data aimed at evaluating algorithms. It can provide datasets following many distributions, including fractal ones;
—VisualTPCH – a tool to generate synthetic data for data warehouses;

5. CONCLUSION

Our goal in this article was to summarize the main research issues that the Images and Data Bases Group (*Grupo de Bases de Dados e de Imagens* - GBdI) from the Institute of Mathematics and Computer Sciences at São Carlos-USP has been pursuing so far. We described 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, aimed at extending the relational database management*

systems into practical tools to support scientific activities that handles 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 enables including new complex data types, such as images, audio and time series and their corresponding retrieval operations, extending the Relational Algebra to target similarity-based predicates;

—Develop techniques for indexing, selectivity and cost estimations, optimization and memory management 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 database management systems, 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 toward more "user centered" perspectives. For example, efficient operators exist to perform similarity comparisons inside the database management system and in the analysis tools, but how to algorithmically define what is similarity 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 really range queries and k-nearest neighbors 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 their provenance? Are the visualizations provided about them sufficiently representative of the data and adequate for the users' intent?

As we are seeing people using the tools we have developed, following their expectations and receiving their positive and negative impressions, we conclude that our techniques indeed help them, but we also conclude that much more still need to be done. 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, grouping and further developing fractal-based algorithms to characterize data distribution and operator selectivity; extend the set of similarity operations to closely meet the user's needs, 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 Jr., C. The fractal dimension making similarity queries more efficient. In *Proc. of the Second Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches*. Washington, DC, pp. 12–17, 2003.

Arantes, A. S., Vieira, M. R., Traina Jr., C., 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 Jr., C., Kinoshita, S. K., and Rangayyan, 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 Jr., C. MAMCost: global and local estimates leading to robust cost estimation of similarity queries. In *Proc. of the 19th Intl. Conf. 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 Jr., C., and Marques, P. M. d. A. Fractal analysis of image textures for indexing and retrieval by content. In *Proc. of the 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 Jr., C. Querying multimedia data by similarity in relational DBMS. In *Advanced Database Query Systems: Techniques, Applications and Technologies*. IGI Global, Hershey, NY, USA, pp. 323–359, 2010.

Barioni, M. C. N., Razente, H. L., Traina, A. J. M., and Traina Jr., C. SIREN: a similarity retrieval engine for complex data. In *Proc. of the Demo section of the 32nd Intl. Conference on Very Large Data Bases*. ACM Press, Seoul, South Korea, pp. 1155–1158, 2006.

Barioni, M. C. N., Razente, H. L., Traina, A. J. M., and Traina Jr., C. 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 Jr., C. Seamlessly integrating similarity queries in SQL. *Software: Practice and Experience* 39 (4): 355–384, 2009.

Barioni, M. C. N., Razente, H. L., Traina Jr., C., and Traina, A. J. M. Visually mining on multiple relational tables at once. In *Proc. Advances in Databases and Information Systems*. Vol. 2. Bratislava, Slovakia, pp. 21–30, 2002.

Barioni, M. C. N., Razente, H. L., Traina Jr., C., and Traina, A. J. M. Querying complex objects by similarity in SQL. In *Anais do Simpósio Brasileiro de Banco de Dados*. Vol. 1. SBC, pp. 130–144, 2005.

Brito, J. J., Siqueira, T. L. L., Times, V. C., Ciferri, R. R., and Ciferri, C. D. A. Efficient processing of drill-across queries over geographic data warehouses. In *Proceedings of the 13th International Conference on Data Warehousing and Knowledge Discovery*. Lecture Notes in Computer Science. Toulouse, France, 2011.

Bueno, R., Kaster, D. d. S., Paterlini, A. A., Traina, A. J. M., and Traina Jr., C. Unsupervised scaling of multi-descriptor similarity functions for medical image datasets. In *Proc. of the 22th IEEE Intl. Symposium on Computer-Based Medical Systems*. IEEE Computer Society, Albuquerque, NM, EUA, pp. 8 p., 2009.

Bueno, R., Kaster, D. d. S., Traina, A. J. M., and Traina Jr., C. A new approach for optimization of dynamic metric access methods using an algorithm of effective deletion. In *Proc. of the 20th Intl. Conf. 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 Jr., C. 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. d. A., Ciferri, R. R., Traina Jr., C., and Traina, A. J. M. The Onion-tree: Quick indexing of complex data in the main memory. In *Proc. of the Conference on Advances in Databases and Information Systems*. Springer, Riga, Latvia, 2009.

Carelo, C. C. M., Pola, I. R. V., Ciferri, R. R., Traina, A. J. M., Traina Jr., C., and Ciferri, C. D. 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 *Proc. of the 12th Intl. Conference on Data Warehousing and Knowledge Discovery*. Lecture Notes in Computer Science, vol. 6263. Springer, pp. 297–308, 2010.

Chino, D. Y. T., Louza, F. A., Traina, A. J. M., Ciferri, C. D. A., and Traina Jr., C. Time series indexing taking advantage of the Generalized Suffix Tree. *Journal of Information and Data Management* 3 (1): 101–109, 1012.

Chino, D. Y. T., Romani, L. A. S., Traina Jr., C., 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 Jr., C. nsP-index: A robust and persistent index for nucleotide sequences. In *Proc. of the 12th East European Conference*. Tampere University of Technology, pp. 28–41, 2008.

Cordeiro, R. L. F., Traina, A. J. M., Faloutsos, C., and Traina Jr., C. Halite: Fast and scalable multi-resolution local-correlation clustering. *IEEE Trans. on Knowledge and Data Engineering* 25 (2): 387–401, 2013.

Cordeiro, R. L. F., Traina, A. J. M., Traina Jr., C., López, J., Kang, U., and Faloutsos, C. Clustering very large multi-dimensional datasets with MapReduce. In *Proc. of the ACM Intl. Conf. On Knoledge Discovery and Datamining*. ACM Press, San Diego, CA, USA, 2011.

Cordeiro$^{(1)}$, R. L. F., Traina, A. J. M., Faloutsos, C., and Traina Jr., C. Finding clusters in subspaces of very large, multi-dimensional datasets. In *Proc. of the 26th IEEE Intl. Conf. on Data Engineering*. IEEE Computer Society, Long Beach, California, pp. 625–636, 2010.

Cordeiro$^{(2)}$, R. L. F., Guo, F., Haverkamp, D. S., Horne, J. H., Hughes, E. K., Kim, G., Traina, A. J. M., Traina Jr., C., and Faloutsos, C. QMAS: Querying, Mining And Summarization of multi-modal databases. In

Proc. of the 10th IEEE Intl. Conf. on Data Mining. IEEE Computer Society, Sydney, NSW , Australia, pp. 785 – 790, 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 *Proc. of the Demos Session of the Brazilian Symposium on Databases*. SBC, Campinas, São Paulo, Brazil, pp. 31–36, 2008.

Faloutsos, C., Seeger, B., Traina, A. J. M., and Traina Jr., C. Spatial join selectivity using power laws. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*. ACM Press, Dallas, TX, pp. 177–188, 2000.

Fedel, G. d. S., Razente, H. L., Traina Jr., C., and Barioni, M. C. N. Uma ferramenta para visualização em SGBDRs com uma implementação interativa do algoritmo para detecção de agrupamentos k-medoid. In *Anais da 5ª 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 Jr., C. Retrieval by content of medical images using texture for tissue identification. In *Proc. of the 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 Jr., C. 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., Traina, A. J. M., Dias, I., Chbeir, R., and Traina Jr., C. Identifying algebraic properties to support optimization of unary similarity queries. In *Proc. of the The III Alberto Mendelzon International Workshop on Foundations of Data Management*. CEUR Workshop Proceedings Series (http://ceur-ws.org), Arequipa, Peru., pp. 10 pags., 2009.

Ferreira, M. R. P., Traina Jr., C., and Traina, A. J. M. An efficient framework for similarity query optimization. In *Proc. of the 15th ACM International Symposium on Advances in Geographic Information Systems*. Seattle, Washington, pp. 396–39, 2007.

Ferreira(1), M. R. P., Ribeiro, M. X., Traina, A. J. M., Chbeir, R., and Traina Jr., C. Adding knowledge extracted by association rules into similarity queries. *Journal of Information and Data Management* 1 (3): 391–406, 2010.

Ferreira(2), M. R. P., Ponciano da Silva, M., Traina, A. J. M., Traina Jr., C., Amo, S. A. d., Pereira, F. S., and Chbeir, R. Integrating user preference to similarity queries over medical images datasets. In *Proc. of the Intl. Symposium on Computer-Based Medical Systems*. IEEE Computer Society, Perth, Australia, pp. 1–6, 2010.

Kaster, D. d. S., Bueno, R., Bugatti, P. H., Traina, A. J. M., and Traina Jr., C. Incorporating metric access methods for similarity searching on Oracle database. In *Proc. of the 24º Simpósio Brasileiro de Bases de Dados, Applications and Experiences Track*. Vol. 1. SBC, Fortaleza, CE, pp. 15 p., 2009.

Kaster, D. d. S., Bugatti, P. H., Ponciano da Silva, M., Traina, A. J. M., Marques, P. M. d. A., and Traina Jr., C. MedFMI-SiR: a powerful DBMS solution for large-scale medical image retrieval. In *Proc. of the 2nd Intl. Conference on Information Technology in Bio- and Medical Informatics*. Lecture Notes in Computer Science (LNCS). Springer Verlag, Toulouse, France, pp. 15 pags., 2011.

Kaster, D. d. S., Bugatti, P. H., Traina, A. J. M., and Traina Jr., C. 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 Intl. Conf. on Data Engineering*. 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 *Anais da Seção de Demos do Simpósio Brasileiro de Banco de Dados*. 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 *Proc. of the 2011 Brazilian Symposium on Bioinformatics*. Lecture Notes in Bioinformatics. Springer, Brasília, DF, Brazil, 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 Jr., C., Sousa, E. P. M., 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 *Proc. of the 12th ACM Intl. Conf. on Knowledge Discovery and Data Mining*. ACM Press, Philadelphia, PA, pp. 693–698, 2006.

Paterlini, A. A., de Faria, R. F. T., Razente, H. L., Traina Jr., C., and Traina, A. J. M. FastMapDB: uma ferramenta para visualização em SGBDRs com suporte à filtragem e seleção visual dos dados. In *Anais da 2ª Sessão de Demos em Banco de Dados, junto ao 20º Simpósio Brasileiro de banco de Dados*. Uberlândia, MG, pp. 1–6, 2005.

Pola, I. R. V., Traina, A. J. M., and Traina Jr., C. Easing the dimensionality curse by stretching metric spaces. In *Proc. of the 21st Intl. Conf. on Scientific and Statistical Database Management*. 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 Jr., C. A novel optimization approach to efficiently process aggregate similarity queries in MAM. In *Proc. of the ACM 17th Intl. Conf. on Information and Knowledge Management*. ACM Press, Napa Valley, CA, pp. 193–202, 2008.

Razente, H. L., Barioni, M. C. N., Traina, A. J. M., and Traina Jr., C. Consultas por similaridade agregada em métodos de realimentação de relevância para consultas por conteúdo de imagens. In *Anais da 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 Jr., C. Visual analysis of feature selection for data mining processes. In *Proc. of the Simpósio Brasileiro de Banco de Dados*. 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 Jr., C. Mining statistical association rules to select the most relevant medical image features. In *Proc. of the First International Workshop on Mining Complex Data, in conjunction with ICDM'05*. IEEE Computer Society, Houston, TX, pp. 91–98, 2005.

Ribeiro, M. X., Marques, J., Traina, A. J. M., and Traina Jr., C. Statistical association rules and relevance feedback: Powerful allies to improve the retrieval of medical images. In *Proc. of the 19th IEEE Intl. Symposium on Computer-Based Medical Systems*. Vol. 1. IEEE Computer Society, Salt Lake City, Utah, USA., pp. 887–892, 2006.

Ribeiro, M. X., Traina, A. J. M., Balan, A. G. R., Traina Jr., C., and Marques, P. M. d. A. SuGAR: association-rules based framework to support diagnosis of abnormalities in mammograms. In *Proc. of the 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 Jr., C., and Marques, P. M. d. A. An efficient association rule-based method to support medical image diagnosis. *IEEE Trans. on Multimedia* 10 (2): 277 – 285, 2008.

Ribeiro[(1)], M. X., Balan, A. G. R., Felipe, J. C., Traina, A. J. M., and Traina Jr., C. Mining statistical association rules to select the most relevant medical image features. In *Mining Complex Data*. Studies in Computational Intelligence, vol. 165/2009. Springer Berlin / Heidelberg, Berlin / Heidelberg, pp. 113–131, 2009.

Ribeiro[(2)], M. X., Bugatti, P. H., Traina, A. J. M., Traina Jr., C., 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* Special issue on Knowledge Discovery in Medicine: Vol. 68 (12): 1370–1382, 2009.

Rodrigues Jr., J. F., Tong, H., yu Pan, J., Traina, A. J. M., Traina Jr., C., and Faloutsos, C. Large graph analysis in the GMine system. *IEEE Trans. on Knowledge and Data Engineering* 25 (1): 106–118, 2013.

Rodrigues Jr., J. F., Traina, A. J. M., Faloutsos, C., and Traina Jr., C. Supergraph visualization. In *Proc. of the IEEE International Symposium on Multimedia*. 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 Jr., C. 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 Jr., C. Mining frequent patterns for visual interpretation in a multi-view environment. In *Visual Data Mining: Theory, Techniques and Tools*. LNCS State of the Art Surveys, vol. 4404. Springer-Verlag Press, Heidelberg, Germany, pp. 50–69, 2008.

Romani, L., Traina, A. J. M., Traina Jr., C., Chbeir, R., Avila, A. M. H., and Jr, J. Z. A new time series mining approach applied to multitemporal remote sensing imagery. *IEEE Trans. on Geoscience and Remote Sensing* 51 (1): 140–150, 2013.

Romani, L. A., Sousa, E. P. M., Ribeiro, M. X., Zullo Jr., J., Traina Jr., C., and Traina, A. J. M. Employing fractal dimension to analyze climate and remote sensing data streams. In *Proc. of the First SIAM Workshop on Multimedia Data Mining*. Vol. 1. SIAM, Sparks, Nevada, pp. 5–16, 2009.

Romani, L. A., Sousa, E. P. M., Ribeiro, M. X., Ávila, A. A. M. H. d., Zullo Jr., J., Traina Jr., C., 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*. Springer Verlag, 2010.

Romani, L. A., Traina, A. J. M., Traina Jr., C., 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 *Proc. of the 10th IEEE International Geoscience and Remote Sensing Symposium*. 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 Jr., C., Sousa, E. P. M., and Traina, A. J. M. Clustering analysis applied to NDVI/NOAA multitemporal images to improve the monitoring process of sugarcane crops. In *Proc. of the Sixth International Workshop on the Analysis of Multi-temporal Remote Sensing Images*. Trento, Italy, pp. 1–4, 2011.

Romani, L. A. S., Ávila, A. M. H. d., Zullo Jr., J., Traina Jr., C., 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 Jr., C. Sistema de recuperação de imagens similares em um hospital universitário. In *Anais do VIII Congresso Brasileiro de Informática em Saúde*. Natal, RN-Brazil, 2002.

Seraphim, E., Seraphim, T. F. P., Moreira, E. M., Ricotta, F. C., and Traina Jr., C. Paged similarity queries. *Information Sciences* 181 (13): 2600–2607, 2011.

Silva, S. F. d., Ribeiro, M. X., Neto, J. d. E. S. B., Traina Jr., C., and Traina, A. J. M. Improving ranking quality in medical image retrieval: the Fc-Family of genetic feature selection evaluation functions. *Decision Support Systems*, 2010.

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* 16 (1): 165–205, 2012.

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 *Proc. of the 12th Intl. Conference on Data Warehousing and Knowledge Discovery*. Lecture Notes in Computer Science, vol. 6263. Springer, 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 *Proc. of the Advancing Geoinformation Science for a Changing World*. Vol. 1. pp. 379–397, 2011.

Sousa, E. P. M., Traina, A. J. M., Traina Jr., C., and Faloutsos, C. Measuring evolving data streams' behavior through their intrinsic dimension. *New Generation Computing* 25 (1): 33–60, 2007.

Sousa, E. P. M., Traina Jr., C., Traina, A. J. M., and Faloutsos, C. How to use fractal dimension to find correlations between attributes. In *Proc. of the First Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches in conjunction with 8th ACM Intl. Conf. on Knowledge Discovery and Data Mining*. ACM Press, Edmonton, Alberta, Canada, pp. 26–30, 2002.

Sousa, E. P. M., Traina Jr., C., 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 Jr., C. Reconciliando dados de cunho acadêmico. In *Anais do Simpóxio Brasileiro de Banco de Dados*. 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 *Proc. of the 19th ACM Conference on Information and Knowledge Management*. ACM, pp. 1349–1352, 2010.

Traina, A. J. M., Romani, L. A., Cordeiro, R. L. F., Sousa, E. P. M., Ribeiro, M. X., Ávila, A. A. M. H. d., Zullo Jr., J., Rodrigues Jr., J. F., and Traina Jr., C. How to find relevant patterns in climate data: an efficient and effective framework to mine climate time series and remote sensing images. In *Proc. SIAM Annual Meeting 2010*. SIAM, Pittsburgh, PA, pp. 6 pags., 2010.

Traina, A. J. M., Traina Jr., C., 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*. Springer-Verlag, pp. 181–209, 2010.

Traina, A. J. M., Traina Jr., C., 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*. SBC, Rio de Janeiro, RJ, pp. 95–109, 2001.

Traina, A. J. M., Traina Jr., C., 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* 6 (2): 157–185, 2003.

Traina, A. J. M., Traina Jr., C., 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* 4 (1): 47–67, 2009.

Traina, A. J. M., Traina Jr., C., Papadimitriou, S., and Faloutsos, C. Tri-Plots: scalable tools for multidimensional data mining. In *Proc. of the ACM Intl. Conf. on Knowledge Discovery and Data Mining*. ACM Press, San Francisco, CA, pp. 184–193, 2001.

Traina Jr., C., Santos Filho, R. F., Traina, A. J. M., Vieira, M. R., and Faloutsos, C. The OMNI-Family 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 Jr., C., Sousa, E. P. M., and Traina, A. J. M. Using fractals in data mining. In *New Generation of Data Mining Applications*. Vol. 1. Wiley/IEEE Press, pp. 30p., 2005.

Traina Jr., C., Traina, A. J. M., Faloutsos, C., and Seeger, B. Fast indexing and visualization of metric datasets using Slim-trees. *IEEE Trans. on Knowledge and Data Engineering* 14 (2): 244–260, 2002.

Traina Jr., C., Traina, A. J. M., Santos Filho, R. F., and Faloutsos, C. How to improve the pruning ability of dynamic metric access methods. In *Proc. of the Intl. Conf. on Information and Knowledge Management*. ACM Press, McLean, VA, USA, pp. 219–226, 2002.

Traina Jr., C., Traina, A. J. M., Seeger, B., and Faloutsos, C. Slim-Trees: high performance metric trees minimizing overlap between nodes. In *Intl. Conf. on Extending Database Technology*. Lecture Notes in Computer Science, vol. 1777. Springer Verlag, Konstanz, Germany, pp. 51–65, 2000.

Traina Jr., C., 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 *Proc. of the ACM 15th Intl. Conf. on Information and Knowledge Management*. ACM Press, Arlington - VA, USA, pp. 4–13, 2006.

Traina Jr., C., Traina, A. J. M., Wu, L., and Faloutsos, C. Fast feature selection using fractal dimension. In *Brazilian Symposium on Databases*. SBC, João Pessoa, PB, pp. 158–171, 2000.

Traina Jr., C., 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 Jr., C., and Traina, A. J. M. Knowledge extraction using visualization of hemoglobin parameters to identify thalassemia. In *Proc. of the 17th IEEE Symposium on Computer-Based Medical Systems*. Vol. 1. Bethesda, Maryland, pp. 523–528, 2004.

Vespa, T. G., Traina, A. J. M., and Traina Jr., C. Efficient bulk-loading on dynamic metric access methods. *Information Systems Journal* 35 (5): 557–569, 2010.

Vieira, M. R., Chino, F. J. T., Traina Jr., C., and Traina, A. J. M. A visual framework to understand similarity queries and explore data in metric access methods. *Intl. Journal on Business Intelligence and Data Mining* 5 (4): 370–387, 2010.

Vieira, M. R., Razente, H. L., Barioni, M. C. N., Hadjieleftheriou, M., Srivastava, D., Traina Jr., C., and Tsotras, V. J. On query result diversification. In *Proc. of the 27th IEEE Intl. Conference on Data Engineering*. Hannover, Germany, pp. 1163–1174, 2011.

Vieira, M. R., Traina Jr., C., 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 Jr., C., Traina, A. J. M., Arantes, A. S., and Faloutsos, C. Estimating suitable query radii to boost knearest neighbor queries. In *Proc. of the 19th Intl. Conf. on Scientific and Statistical Database Management*. ACM Press, Banff, Canada, pp. 1–10, 2007.