INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Team AVIZ*

# *Analysis and Visualization*

*Futurs*

THEME COG

# Table of contents

# 1. Team

**Head of Project-Team**
Jean-Daniel Fekete [ Research Director (DR), INRIA, HdR ]

**Project Assistant**
Marie-Carol Lopes

**Research Scientists**
Pierre Dragicevic [ Research Associate (CR2), INRIA ]

**Post-doctoral Fellow**
Nghi Do-Thanh [ Sep. 2006 to Jan. 2008 ]
Niklas Elmqvist [ Jan. 2007 to Jun. 2008 ]
Howard Goodell [ Jul. 2006 to Jun. 2007 ]
Stéphane Huot [ Feb. 2007 to Aug. 2007 ]

**Ph. D. student**
Nathalie Henry [ Sep. 2005 to 2008, Co-Advised by Jean-Daniel Fekete (INRIA) and Peter Eades (Univ. of Sydney, Australia) ]

**Interns**
Félicien François [ Université de Rennes 1 ]
Yoann Lheudé [ Université Paris-Sud ]
Ludovic Hauchemaille [ Université Paris-Sud ]

# 2. Overall Objectives

## 2.1. Objectives

Like many other fields, all the sciences are being transformed by our rapidly-increasing abilities to collect, manage and understand vast amounts of data. A 2003 study estimated that the amount of data produced in the world was increasing by 50% each year [50]. According to SearchEngineWatch[1], the amount of information made available through Internet search engines has grown exponentially for the last decade, and major Web search engines currently index more than 2 billion documents. However, since our brains and sensory capacities have not changed in the meantime, gaining competitive advantage from all this data depends increasingly on the effectiveness with which we support human abilities to perceive, understand, and act on it.

*The focus of the AVIZ project is to design methods and tools that make analyzing large data sets easy and massive data sets possible. Our interests include:*

- *Methods to visualize and smoothly navigate through large datasets;*
- *Efficient analysis methods to reduce huge datasets to visualizable size;*
- *Evaluation methods to assess their effectiveness and usability;*
- *Engineering tools for building visual analytics systems that can access, search, visualize and analyze large datasets with smooth, interactive response.*

[1] http://www.searchenginewatch.com

## 2.2. Research Themes

AVIZ's research on Visual Analytics is organized around four main Research Themes:

Methods to visualize and smoothly navigate through large data sets Large data sets challenge current visualization and analysis methods. Understanding the structure of a graph with one million vertices is not just a matter of displaying the vertices on a screen and connecting them with lines. Current screens only have around two million pixels. Understanding a large graph requires both data reduction to visualize the whole and navigation techniques coupled with suitable representations to see the details. These representations, aggregation functions, navigation and interaction techniques must be chosen as a coordinated whole to be effective and fit the user's mental map.

AVIZ designs new visualization representations and interactions to efficiently navigate and manipulate them.

Efficient analysis methods to reduce huge data sets to visualizable size Designing analysis components with interaction in mind has strong implications for both the algorithms and the processes they use. Some data reduction algorithms are suited to the principle of sampling, then extrapolating, assessing the quality and incrementally enhancing the computation: for example, all the linear reductions such as PCA, Factorial Analysis, and SVM, as well as general MDS and Self Organizing Maps. We investigate the generality of the approach and also explore other methods for cases such as for language processing where sampling severely reduces the quality of the result.

Evaluation methods to assess their effectiveness and usability Designing analysis components with interaction in mind has strong implications for both the algorithms and the processes they use. Some data reduction algorithms are suited to the following process: sampling, then extrapolating, assessing the quality and incrementally enhancing the computation. For example, all the linear reductions methods such as PCA, Factorial Analysis, and SVM, as well as general MDS and Self Organizing Maps can use that process. We investigate the generality of the approach and also explore other methods for cases such as for language processing where sampling severely reduces the quality of the result.

Engineering tools for building visual analytic systems that can access, search, visualize and analyze large data sets with smooth, interactive response.

AVIZ seeks at merging three fields: databases, data analysis and visualization. Part of this merging consists in using common abstractions and interoperable components. This is a long-term challenge, but it is a necessity because generic, loosely-coupled combinations will not achieve interactive performance.

Currently, databases, data analysis and visualization all use the concept of data tables made of tuples and linked by relations. However, databases are storage-oriented and do not describe the data types precisely. Analytical systems describe the data types precisely, but their data storage and computation model are not suited to interactive visualization. Visualization systems use in-memory data tables tailored for fast display and filtering, but their interactions with external analysis programs and databases are often slow.

These themes are presented separately, but they are closely linked: a good multi-scale visualization technique relies on an analysis method to generate the suitable data structure. The effectiveness of the Visual Analytics tool has to be evaluated at several levels (component, system, environment). Finally, to build Visual Analytics systems that manage large data sets, the software infrastructure has to provide the right abstractions and mechanisms. Therefore, each of the four research themes work together. One of the scientific challenges is to fit them all together into a coherent framework supporting the analyst's work process.

## 2.3. Highlights

Hybrid Matrix Representations Nathalie Henry and Jean-Daniel Fekete have published two important articles on network visualization using hybrid representations of matrices and node-link diagrams:

MatLink [28] and NodeTrix [12]. Finding better representations for large and dense networks is important to support exploration tasks for domains such as social network analysis and bioinformatics. Our representations have very good properties and were welcome by the HCI and InfoVis community. MatLink received the best paper award at the Interact 2007 conference (Brian Shackel Award). See section 6.1 for details.

Multi-scale Navigation in Huge Networks AVIZ has reached a new milestone in multi-scale visualization and navigation of large networks with ZAME, the Zoomable Adjacency Matrix Explorer [33] that can reorder and visualize network with millions of vertices and tens of millions of edges. We intend to use it to explore the Wikipedia networks (article to article, article to author, author to author, etc.) and large protein networks in bioinformatics (see section 6.2).

Visualizing Wikipedia for Occasional Users Stéphane Huot and Jean-Daniel Fekete have designed new visualizations called WikipediaViz to improve the trust and transparency of Wikipedia articles.

SVM for Very Large Datasets Thanh-Nghi Do has substantially increased the performance of SVM classification algorithms [20]. His method is applicable to very large dataset in the number of individuals and dimensions (see section 6.4).

# 3. Scientific Foundations

## 3.1. Scientific Foundations

The scientific foundations of Visual Analytics lie primarily in the domains of Information Visualization and Data Mining. Indirectly, it inherits from other established domains such as graphic design, Exploratory Data Analysis (EDA), statistics, Artificial Intelligence (AI), Human-Computer Interaction (HCI), and psychology.

The use of graphic representation to understand abstract data is a goal Visual Analytics shares with Tukey's Exploratory Data Analysis (EDA) [58], graphic designers such as Bertin [40] and Tufte [57], and HCI researchers in the field of Information Visualization [39].

EDA is complementary to classical statistical analysis. Classical statistics starts from a *problem*, gathers *data*, design a *model* and performs an *analysis* to reach a *conclusion* about whether the data follows the model. While EDA also starts with a problem and data, it is most useful *before* we have a model; rather, we perform visual analysis to discover what kind of model might apply to it. However, statistical validation is not always required with EDA; since often the results of visual analysis are sufficiently clear-cut that statistics are unnecessary.

Visual Analytics relies on a process similar to EDA, but expands its scope to include more sophisticated graphics and areas where considerable automated analysis is required before the visual analysis takes place. This richer data analysis has its roots in the domain of Data Mining, while the advanced graphics and interactive exploration techniques come from the scientific fields of Data Visualization and HCI, as well as the expertise of professions such as cartography and graphic designers who have long worked to create effective methods for graphically conveying information.

The books of the cartographer Bertin and the graphic designer Tufte are full of rules drawn from their experience about how the meaning of data can be best conveyed visually. Their purpose is to find effective visual representation to describe a data set but also (mainly for Bertin) to discover structure in the data by using the right mappings from abstract dimensions in the data to visual ones.

For the last 25 years, the field of Human-Computer Interaction (HCI) has also shown that interacting with visual representations of data in a tight perception-action loop improves the time and level of understanding of data sets. Information Visualization is the branch of HCI that has studied visual representations suitable to understanding and interaction methods suitable to navigating and drilling down on data. The scientific foundations of Information Visualization come from theories about perception, action and interaction.

Several theory of perception are related to information visualization such as the "Gestalt" principles or Gibson's theory of visual perception [46]. However, the only predictive theory related to the perception of visual shapes is Triesman's "preattentive processing" [56].

Information Visualization emerged from HCI when researchers realized that interaction greatly enhanced the perception of visual representations. To be effective, interaction should take place in an interactive loop faster than 100ms. For small data sets, it is not difficult to guarantee that analysis, visualization and interaction steps occur in this time, permitting smooth data analysis and navigation. For larger data sets, more computation should be performed to reduce the data size to a size that may be visualized effectively.

In 2002, we showed that the practical limit of InfoVis was on the order of 1 million items displayed on a screen [3]. Although screen technologies have improved rapidly since then, eventually we will be limited by the physiology of our vision system: about 20 millions receptor cells (rods and cones) on the retina. Another problem will be the limits of human visual attention, as suggested by our 2006 study on change blindness in large and multiple displays [1]. Therefore, visualization alone cannot let us understand very large data sets. Other techniques such as aggregation or sampling must be used to reduce the visual complexity of the data to the scale of human perception.

Abstracting data to reduce its size to what humans can understand is the goal of the Data Mining research domain. It uses data analysis and machine learning techniques. The scientific foundations of these techniques revolve around the idea of finding a good model for the data. Unfortunately, the more sophisticated techniques for finding models are complex, and the algorithms can take a long time to run, making them unsuitable to an interactive environment. Furthermore, some models are too complex for humans to understand; so the results of data mining can be difficult or impossible to understand directly.

Unlike pure Data Mining systems, a Visual Analytics system provides analysis algorithms and processes compatible with human perception and understandable to human cognition. The analysis should provide understandable results quickly, even if they are not ideal. Instead of running to a predefined threshold, algorithms and programs should be designed to allow trading speed for quality and show the tradeoffs interactively. This is not a temporary requirement: it will be with us even when computers are much faster, because good quality algorithms are at least quadratic in time (e.g. hierarchical clustering methods). Visual Analytics systems need different algorithms for different phases of the work that can trade speed for quality in an understandable way.

Designing novel interaction and visualization techniques to explore huge data sets is an important goal and requires solving hard problems, but how can we assess if our techniques and systems provide real improvements? Without this answer, we cannot know if we are heading in the right direction. This is why we have been actively involved in the design of evaluation methods for information visualization [16] [51], [48], [49], [44]. For more complex systems, other methods are required. For these we want to focus on longitudinal evaluation methods, while still trying to improve controlled experiments [29].

# 4. Application Domains

## 4.1. Application Domains

AVIZ develops active collaboration with users from various application domains, making sure it can support their specific needs. By studying similar problems in different domains, we can begin to generalize our results and have confidence that our solutions will work for a variety of applications. Our current application domains include:

- Business Intelligence, in cooperation with EDF.
- Social Network Analysis, in cooperation with France Telecom R&D, Univ. LIAFA, GET/ENST, and the French National Archives;
- Biological research, in cooperation with INRA, the IGM Biological Research Laboratory at Univ. Paris-Sud and Institut Pasteur;
- Digital Libraries, in cooperation with the French National Archives, the Bibilothèque Nationale and ITEM.

#### *4.1.1. Business Intelligence*

Business Intelligence aims at collecting and processing heterogeneous information to orient business decisions in term of product design or commercial offers. Both the quantity of information and the diversity of sites and formats where it can be collected is growing (e.g. Blogs and Social Network websites). We want to address the challenge of offering tools and components to quickly build analysis applications suited to these diverse inputs and the many specific tasks marketing analysts may attempt to do, helping them to quickly carry-out their work and produce understandable synthetic reports. We are working on such applications for EDF (see section 7.3).

#### *4.1.2. Social Network Analysis*

In the social networks domain, we are starting to work on exploratory visualization. Current studies in social networks presuppose that users know the nature of the networks they want to explore and the kinds of of transformations and layouts that will best suit their needs. This is often not true, and tools are very weak at helping users understand the nature of their networks and the transformations they could perform to get meaningful insights. This work began in 2004 with the arrival of Nathalie Henry in the Project. She is co-advised by Jean-Daniel Fekete and Peter Eades from the University of Sydney and NICTA, Australia.

We have been focusing on the use of the matrix representation to explore large graphs, building on our previous work using matrices for constraint-based programming. Matrices present challenging problems both interactively and mathematically. We are designing an interactive system to help users navigate and interact with large matrices. We are also preparing a survey on methods to reorder matrices, whether from graphs from tabular data.

#### *4.1.3. Biological Visualization*

Bioinformatics uses many complex data structures such as phylogenetic trees, genomes made of multi-scale parts (sequences of base pairs, genes, interaction pathways etc.) Biologists navigate through multitudes of these varied and complex structures daily in complex, changeable, data- and insight-driven paths. They also often need to edit these structures to annotate genes and add information about their functions. Visual Analytics is a powerful tool to help them, as we are currently pursuing in the Microbiogenomics project (see section 7.5).

#### *4.1.4. Digital Libraries*

In the digital Library domain, we collaborate with the French National Archives on an exploratory project to visualize and analyze the Evolution of the French Political Organization before and after the Revolution (see section 7.6).

# 5. Software

## 5.1. The InfoVis Toolkit

**Keywords:** *Information Visualization*, *Java*, *Toolkit*.

**Participants:** Jean-Daniel Fekete [correspondant], Howard Goodell, Nathalie Henry, Nghi Do-Thanh, Niklas Elmqvist.

The InfoVis Toolkit [2] is an Interactive Graphics Toolkit written in Java to facilitate the development of Information Visualization applications and components.

The main characteristics of the InfoVis Toolkit are:

Unified data structure The base data structure is a table of columns. Columns contain objects of homogeneous types, such as integers or strings. Trees and Graphs are derived from Tables.

Small memory footprint Using homogeneous columns instead of compound types dramatically reduces the memory required to store large tables, trees or graphs, and usually also the time required to manage them.

Unified set of interactive components Interactive filtering (a.k.a. dynamic queries) can be performed with the same control objects and components regardless of the data structure, simplifying the reuse of existing components and the design of generic ones.

Fast The InfoVis Toolkit can use accelerated graphics provided by Agile2D[2], an implementation of Java2D based on the OpenGL API for hardware accelerated graphics [3]. On machines with hardware acceleration, some visualizations redisplay 100 times faster than with the standard Java2D implementation.

Extensible The InfoVis Toolkit is meant to incorporate new information visualization techniques and is distributed with the full source and a very liberal license. It can be used for student projects, research projects or commercial products.

The InfoVis Toolkit, as of version 0.9, implements nine types of visualization (Fig. 1): Time Series, Scatter Plots, Parallel Coordinates and Matrices for tables, Node-Link diagrams, Icicle trees and Treemaps for trees, Adjacency Matrices and Node-Link diagrams (with several layouts) for graphs.

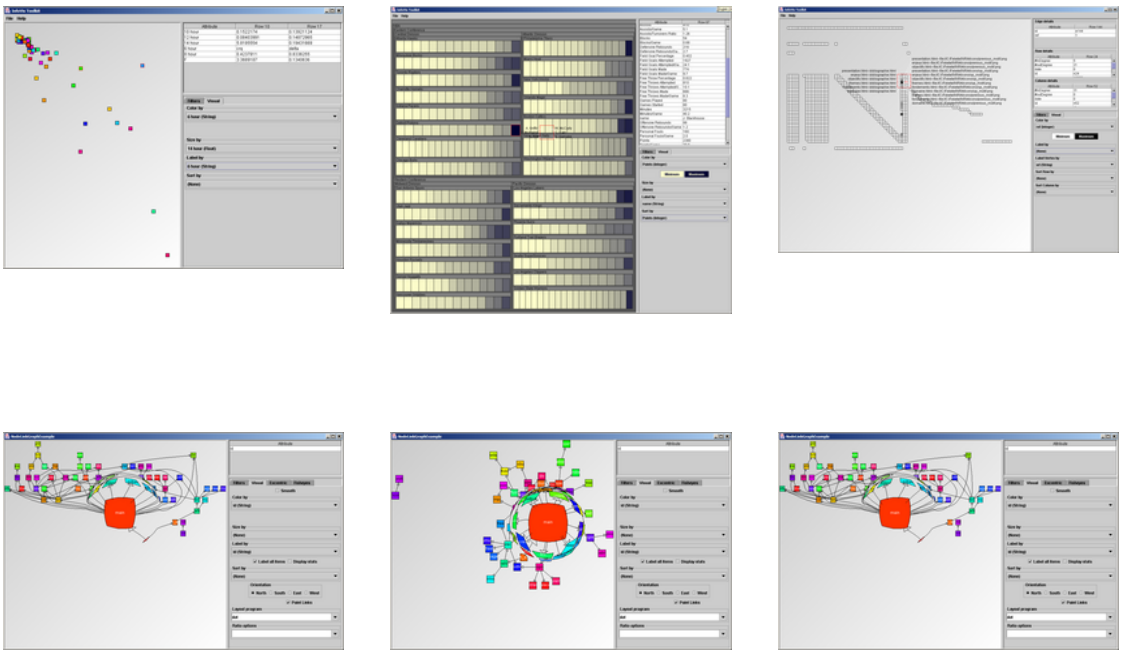

*Figure 1. Several visualizations produced using the Infovis Toolkit*

The InfoVis toolkit is used for teaching the Information Visualization course (Masters level, Univ. of Paris-Sud) and is the basis for all AVIZ contracts. It is available at http://ivtk.sourceforge.net.

[2] http://www.cs.umd.edu/hcil/agile2d

## 5.2. Dataflow Editor for Visual Analytics

**Keywords:** *Dataflow*, *Information Visualization*, *Java*, *JavaBean*, *Toolkit*, *Visual Analytics*.

**Participants:** Jean-Daniel Fekete [correspondant], Nghi Do-Thanh, Howard Goodell, Nathalie Henry.

Building visual analytics application requires combining several analysis modules with visualizations where data sets come from data sources. To simplify this process, we are designing a work-flow visual editor that combines software modules interactively, can run them and can create a stand-alone application.

DeVa relies on the JavaBean component architecture. It allows users to connect JavaBean modules in a generic way using the JBeanStudio system [54]. Visualization modules are built directly from the InfoVis Toolkit which implements the JavaBean architecture natively. To allow high-performance, large-scale processing, we rely on the data model designed for the InfoVis Toolkit. In-memory data tables can be shared between various modules, and views can be defined with local attributes. We are currently building and testing the infrastructure on our tools and on modules created by other research groups in Java or C++ [19].

## 5.3. XML/TEI Eclipse Plugin

**Keywords:** *Information Visualization*, *Java*, *Plugin*, *TEI Eclipse*, *XML.*

**Participants:** Jean-Daniel Fekete [correspondant], Félicien François.

The Millefeuille Platform is a Plugin for the Eclipse programming environment designed to assist historians in encoding their documents. It provides a set of mechanisms found in standard programming environments but not known by historians, including source version control (SVN), project management and asynchronous collaboration tools. The Plugin improves the standard XML editor of the Eclipse Platform with several encoding help for building multiple indexes, verifying the consistency of XML encoding based on high-level properties and dynamically apply a stylesheet to the XML encoded files. Indexes are one kind of cross-document structures that the Plugin can dynamically create from the XML structure.

The Millefeuille Platform is currently used to encode in XML/TEI a sample of the administration of France during 100 years, from just before the revolution to 100 years later. The encoding is done using a generated XML Schema based on TEI P5 (see http://www.tei-c.org/Guidelines/P5/. It is generated with the Roma tool (http://tei.oucs.ox.ac.uk/Roma/).

The Millefeuille Platform incorporates several checking and an experimental visualization module to show an overview of the French administration across time.

The Platform is described in [34]

# 6. New Results

## 6.1. Alternative Visualizations of Social Networks

**Keywords:** *Graph Layout*, *Matrix*, *Social Networks*, *Visualization*.

**Participants:** Nathalie Henry [correspondant], Jean-Daniel Fekete.

Social networks analysis and visualization is becoming more and more important, due to the development of online communities on the Web, but also to the increase of security-related threats such as terrorist attacks and spreading of epidemics.

Visualizing large or dense social networks is simply not possible using current node-link diagram representations. We have shown that the matrix representation was a good alternative to node-link diagrams. However, it has not received as much attention as node-link diagrams in the past and the research community needs to design good navigation and layout methods to improve it.

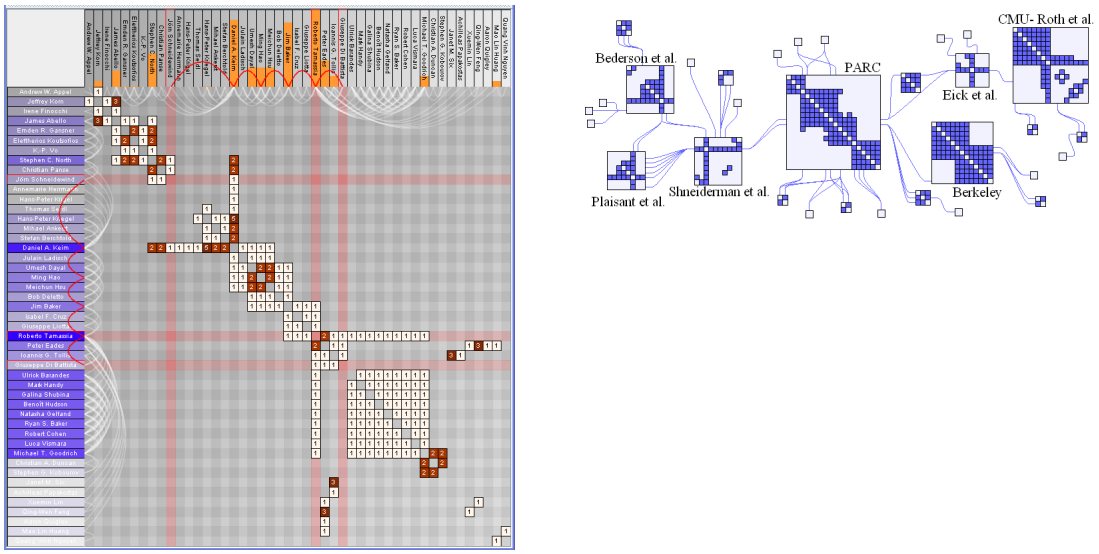


*Figure 2. MatLink and NodeTrix visualizations of a social network*

We have worked in that direction and proposed two enhancements to matrix visualization: hybrid representations using links overlaid on top of a matrix (MatLink, Fig. 2) and node-link representation using small matrices for dense subparts (NodeTrix, Fig. 2).

We have shown that the MatLink representation improved the performance compared to the traditional matrix representation for tasks related to path-finding.

The NodeTrix [12] hybrid representation is meant for small-world networks that are globally sparse but locally dense. NodeTrix mixes node-link diagrams for the global sparse structure and matrices for the local dense structure, effectively using each representation for its strength and avoiding their weaknesses.

We have also worked with international researchers to improve evaluation methods of network visualization systems by proposing a taxonomy of network-related tasks [51].

## 6.2. Multi-scale Navigation on Large Matrices

**Keywords:** *Graph Layout*, *Matrix*, *Multi-Scale Interaction*, *Reordering*, *Visualization*, *Visualization*.

**Participants:** Jean-Daniel Fekete [correspondant], Nathalie Henry, Niklas Elmqvist, Nghi Do-Thanh, Howard Goodell.

The Zoomable Adjacency Matrix Explorer (ZAME) is a visualization tool for exploring graphs at a scale of millions of nodes and edges [33]. ZAME is based on an adjacency matrix graph representation aggregated at multiple scales (Fig. 3). It allows analysts to explore a graph at many levels, zooming and panning with interactive performance from an overview to the most detailed views. Several components work together in the ZAME tool to make this possible. Efficient matrix ordering algorithms group related elements. Individual data cases are aggregated into higher-order meta-representations. Aggregates are arranged into a pyramid hierarchy that allows for on-demand paging to GPU shader programs to support smooth multiscale browsing. Using ZAME, we are able to explore the entire French Wikipedia—over 500,000 articles and 6,000,000 links—with interactive performance on standard consumer-level computer hardware.

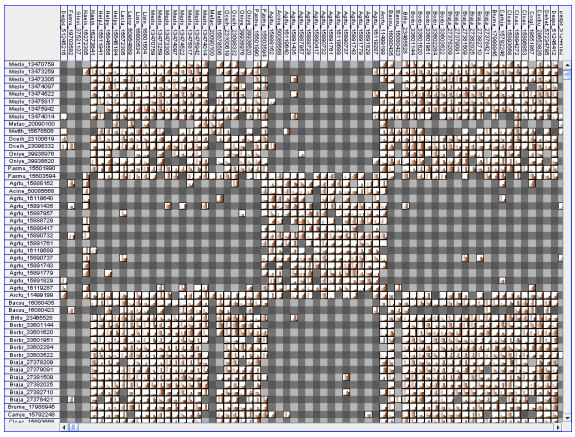

*Figure 3. ZAME aggregated visualization of a 30,000 proteins densely connected graph*

## 6.3. Visualizing Wikipedia for Occasional Users

**Keywords:** *Visualization*, *Wikipedia*.

**Participants:** Jean-Daniel Fekete [correspondant], Stéphane Huot.


*Figure 4. WikipediaViz visualizations on the left of a standard Wikipedia page.*

In seven years, Wikipedia has become one of the top ten most visited web sites in the world, consulted by 36% of US adult internet users. It is made of more than 5 million articles in 250 localized versions. This popularity mainly comes from its availability and coverage: Wikipedia is defined as “the free encyclopedia that anyone can edit”, with “thousands of changes an hour”. This fundamental Wikipedia concept is pointed out as a good way to continuously increase the coverage, accuracy and up-to-dateness of information.

Conversely, this fast changing and volatile content is prone to unverified information, contrary to classical encyclopedia where author selection and peer-reviewing is performed before the information is published. This is considered as the Achilles' heel of Wikipedia as unreliable or incomplete information and vandalism become serious threats to the quality of Wikipedia. Controversies about the relative quality of Wikipedia compared to standard encyclopedia have recently surged and vandalism has become a main concern for Wikipedia administrators. Since more and more people rely on Wikipedia, the cost of unreliable information increases for the society. Helping Wikipedia readers, especially occasional ones, spotting erroneous or bad quality articles is thus becoming increasingly important. This requires some assessment of the quality of articles.

We have designed four visualizations with four thumbnails to keep occasional Wikipedia readers aware of the profile of the article they read (Fig. 4). We asked Wikipedia frequent writers and administrators to give us a list features the considered as important for assessing the quality of the articles. They mainly reported the number of contributors, the evolution of the article across time and the activity of the discussion associated with the article. Our four visualization summarize these features to quickly assess whether the article can be read safely or if it should be double checked. We are currently running experiments to test the effectiveness of these visualization for occasional Wikipedia users.

## 6.4. Large Scale Classification with Support Vector Machine Algorithms

**Keywords:** *Ensemble Methods*, *Incremental Learning*, *Massive Classification*, *Support Vector Machine*.

**Participants:** Thanh-Nghi Do [correspondant], Jean-Daniel Fekete.

Since Support Vector Machine (SVM) learning algorithms were first proposed by Vapnik [59], they have been shown to build accurate models with practical relevance for classification, regression and novelty detection. Successful applications of SVMs have been reported for such varied fields as facial recognition, text categorization and bioinformatics. In particular, SVMs using the idea of kernel substitution have been shown to build good models, and they have become increasingly popular classification tools.

However, in spite of their desirable properties, current SVMs cannot easily deal with very large datasets. A standard SVM algorithm requires solving a quadratic or linear program; so its computational cost is at least $O(m^2)$, where m is the number of training datapoints. Also, the memory requirements of SVM frequently make it intractable. There is a need to scale up these learning algorithms to handle massive datasets.

Effective heuristic methods to improve SVM learning time divide the original quadratic program into series of small problems [41], [52]. Incremental learning methods [42], [43] improve memory performance for massive datasets by updating solutions in a growing training set without needing to load the entire dataset into memory at once. Parallel and distributed algorithms [43] improve learning performance for large datasets by dividing the problem into components that execute on large numbers of networked PCs. Active learning algorithms [55] choose interesting datapoint subsets (active sets) to construct models, instead of using the whole dataset.

We propose methods to build boosting of incremental LS-SVM algorithms for classifying very large datasets on standard personal computers. Most of our work is based on LS-SVM classifiers proposed by Suykens and Vandewalle [53]. They replace standard SVM optimization inequality constraints with equalities in least squares error; so the training task only requires solving a system of linear equations instead of a quadratic program. This makes training time very short. We have extended LS-SVM in three ways:

1. We developed a row-incremental algorithm for classifying massive datasets (billions of points) of dimensionality up to $10^4$.
2. Using a Tikhonov regularization term and the Sherman-Morrison-Woodbury formula [47], we developed a column-incremental LS-SVM algorithm for very-high-dimensional datasets with small training datapoints, such as bioinformatics microarrays.
3. Applying boosting techniques like Adaboost [45] and arcx4 to these incremental LS-SVM algorithms, we developed efficient classifiers for massive, very high-dimensional datasets.

We also applied these ideas to build boosting of other efficient SVM algorithms proposed by Mangasarian and colleagues: Lagrangian SVM (LSVM), Proximal SVM (PSVM) and Newton SVM (NSVM) in the same way, because they have similar properties to LS-SVM. Boosting based on these algorithms is interesting and useful for classification on very large datasets. Some performances in terms of learning time and accuracy are evaluated on UCI, Forest cover type, KDD cup 1999, Reuters-21578 and RCV1-binary datasets. The results showed that our boosting of LS-SVM algorithms are usually much faster and/or more accurate for classification tasks compared with the highly efficient standard SVM algorithm LibSVM and with two recent algorithms, SVM-perf and CB-SVM. An example of the effectiveness of the new algorithms is their performance on the 1999 KDD cup dataset. They performed a binary classification of 5 million datapoints in a 41-dimensional input space within 3 minutes on a standard PC when the fastest method (CB-SVM) required 30 minutes and LibSVM ran out of memory.

# 7. Contracts and Grants with Industry

## 7.1. ReActivity

**Participants:** Jean-Daniel Fekete [correspondant], Niklas Elmqvist, Nathalie Henry.

This project belongs to the joint INRIA-Microsoft Research Laboratory and is a collaboration of the VIBE Group at Microsoft Research in Redmond, the in|situ| and AVIZ INRIA groups. It is a three-year project started in 2007, focused on analyzing researchers' activities to help them reflect on these activities, analyze them or communicate them more effectively. The project has to deal with logging, storing, summarizing, visualizing and interacting with activity data to solve interesting problems in science.

Both VIBE and INRIA are faced with difficult problems in term of data capture, management, retrieval, effective visualization of stored data, effective aggregation, higher-level summarization (inferring the high-level user activity from the captured low-level user activity) and reflective presentation of that information. The teams are collaborating in designing Information Visualization infrastructures capable of managing large amounts of information and interacting with it. The ReActivity project involves logging, visualizing and interacting with logged data. It is split into three phases: collecting the logs in a consistent, extensible and robust way, mining the logs to extract higher-level information and visualizing the information for understanding, interaction and sharing. It addresses these issues for simple desktop-based information initially and then increase the scope of the project by aggregating information from outside sources.

## 7.2. TARANIS: Technologies for the Appraisal of Risks through Animation and Simulation

**Participants:** Pierre Dragicevic [correspondant], Jean-Daniel Fekete, Niklas Elmqvist.

The TARANIS 2-year ANR project (program "Concepts, systèmes et outils pour la sécurité globale") stated in 2006, supported by MASA, ESRI-France and INRIA. It aims at creating a new training system for crisis managers, based on innovative simulation tools, allowing trainers to easily recreate complex crisis situations. Simulation tools give the trainer unprecedented control over the training session while making the virtual crisis reactive to the trainees actions and providing an unlimited variety of extreme crisis situations, a challenge that even very expensive ground exercises cannot meet. TARANIS is a "Global Security" project conducted by the company "Mathématique Appliquée S.A." to design crisis simulation environment.

More information can be found at http://www.masa-sci.com/taranis.htm

## 7.3. Classification and Visualization for Business Intelligence

**Participants:** Jean-Daniel Fekete [correspondant], Nghi Do-Thanh.

SEVEN is a 3-year Business Intelligence project funded by ANR (program RNTL) started in 2006 and conducted by EDF, the main European electricity supplier, with INRIA, LIMSI (Univ. Paris-Sud), and the CEREMADE (Univ. Paris-IX Dauphine). Its goal is to develop a Visual Analytics software platform to understand market segments for EDF. The platform is made of modules that analyze textual documents or numerical data and integrate them to find profiles of clients. This profiling leads to understanding the main concerns of market segments and plan price offerings targeted to these segments.

The partners are experts in language processing (LIMSI), data analysis (Dauphine) and Information Visualization (INRIA and LIMSI).

## 7.4. Analysis and visualization of the Auto-organization process of online social communities

**Participants:** Jean-Daniel Fekete [correspondant], Niklas Elmqvist, Howard Goodell, Stéphane Huot, Nathalie Henry.

Autograph is a 2-year ANR project (program RNRT), started in 2006. The aim of Autograph is to develop tools and services for governance of large cooperative organizations on Internet. This exploratory project intends to draw through research on several online communities (Debian, Wikipedia, international activists, Blogs, SIMS). In order to study the organizational properties of these collectives, the graph theory gives new directions for sociologists, linguists, computer scientists and mathematicians who want to describe social, semantic and computer networks and analyze their structures. The aim of the project is to develop new visualization services, enabling the actors in these communities to "see" the universe in which they cooperate to help them make decisions about the life of their communities. Cartographic and dynamic representations will be given, enabling an exploration of the structure of the links and the thematic universe of the exchanges. All these results will be developed in a tight relationship with the user communities.

Partnership: France Telecom, ENST, LIAFA (Univ. Paris-6), LIMSI (Univ. Paris-Sud), INRIA, LRI (Univ. Paris-Sud), FING

More information can be found at http://autograph.fing.org/.

## 7.5. Integrated Resources for Microbial Genomics

**Participants:** Jean-Daniel Fekete [correspondant], Stéphane Huot, Howard Goodell.

Microbiogenomics is a 3-year ANR project (program "Masses de données" ) stated in 2006. The project is designed to address the challenges raised by the ongoing deluge of genomic data. It plans at designing an integrating resources for microbial genomics. The objective is to gather the largest amount of relevant data and to make it available for a number of data mining approaches, despite its heterogeneity. A graphic interface will be designed for efficient and simple but still expressive queries, letting users extract relevant pieces of knowledge through a visual interactive system. This will make cross-fertilization between domains possible, and allow detailed analysis of a wide range of available genomic data.

Partnership: IGM (Univ. Paris-Sud), LRI (Univ. Paris-Sud), MIG (INRA).

## 7.6. Analysis and Visualization of the History of the French Central Institutions

**Participants:** Jean-Daniel Fekete [correspondant], Félicien François.

Millefeuille is a 2-year CNRS/ACI project (program "histoire des savoirs") stated in 2006. The Millefeuille project ("Archeology of Administrative Knowledge") is led by the French "École nationale des Chartes" with the French National Archives, INRIA, Univ. Sorbonne-Paris I and Univ. Paris X, as partners. It is aimed at analyzing the evolution of the organizational structure before and after the French revolution. The structure of French institutions is represented as a hierarchical or mostly hierarchical structure that evolves with time. This organization is used as a backbone for further analysis, such as the structural evolution of the organization or the path taken by administrative forms in the structure. From these perspectives, the project will study how various structural changes have affected the administrative practices and visualize their evolution through time.

# 8. Other Grants and Activities

## 8.1. National actions

- Jean-Daniel Fekete is member of the Scientific Committee of the French ANR for the *Data Masses* program
- Jean-Daniel Fekete is co-responsible of the Working Group: Tools and Formalisms for HCI (ALF) with Eric Lecolinet
- Jean-Daniel Fekete is a member of the directing committee of the French GDR I3

## 8.2. International actions

- *EDGE: Evaluation methods, Design Guidelines and Environments for Virtual Reality and Information Visualization Techniques*. This project is a French-Brazilian collaboration supported by INRIA and CNPq (36 months, 2005-2008). The partners are MErLIn (INRIA), the CS Institute of the Federal University of Rio Grande do Sul and the CS Department of PUC-Rio University. Members of In Situ involved: Nicolas Roussel (coordinator of the French side) and Jean-Daniel Fekete.
- *Navigation and Visualization of Large Social Networks*. Nathalie Henry is preparing a joint PhD (co-tutelle) with the University of Sidney, Information Visualization Research Group (Australia). Members of In Situ involved: Nathalie Henry and Jean-Daniel Fekete (advisor).
- *Evaluation of Information Visualization*. Jean-Daniel Fekete and Catherine Plaisant of the University of Maryland are gathering resources to improve the evaluation techniques used in the domain of Information Visualization. They have initiated an international contest, taking place every year during the IEEE Symposium on Information Visualization. They gather and maintain the benchmarks and results on an open web site at http://www.cs.umd.edu/hcil/InfovisRepository.

# 9. Dissemination

## 9.1. Keynote addresses and Invited Lectures

- Co-organizer of a Dagstuhl Seminar on Information Visualization (Seminar 07221, 28/05/2007-01/06/2007): Jean-Daniel Fekete
- Invited Lecture at the Human-Computer Interaction Laboratory, University of Maryland (31/05/2007): Nathalie Henry
- Invited Lecture at University of Toronto, Canada, 24/04/2007: Jean-Daniel Fekete
- Invited Lecture at Microsoft Research, Redmond, USA: Jean-Daniel Fekete (27/04/2007), Nathalie Henry (02/09/2007)
- Wikipédia Colloque 2007, Cité des sciences et de l'industrie, Paris, (20/10/2007): Jean-Daniel Fekete
- 5th Cytoscape Symposium, Amsterdam, (8/11/2007): Jean-Daniel Fekete

## 9.2. Journal editorial board

- Associate Editor of the International Journal of Human-Computer Study (IJHCS): Jean-Daniel Fekete

## 9.3. Journal reviewing

- Information Visualization Journal, Palgrave Macmillan: Jean-Daniel Fekete, Niklas Elmqvist
- Document Numérique, Hermès, France: Jean-Daniel Fekete
- Revue de l'Interaction Homme-Machine (RIHM), Cepadues, France: Jean-Daniel Fekete
- Journal of Graph Algorithms and Applications: Jean-Daniel Fekete
- ACM Transactions on Applied Perception: Jean-Daniel Fekete
- ACM Transaction on Human-Computer Interaction: Nathalie Henry
- IEEE Transactions on Visualization and Computer Graphics: Jean-Daniel Fekete
- Empirical Software Engineering, Springer: Niklas Elmqvist
- International Journal of Human-Computer Studies, Elsevier: Niklas Elmqvist
- Journal of Visual Languages and Computing, Elsevier: Niklas Elmqvist

## 9.4. Conference organization

- ACM CHI 2008: Human Factors in Computing Systems, Florence, Italy: Jean-Daniel Fekete (Program Committee member)
- IEEE Symposium on Information Visualization 2007: Jean-Daniel Fekete (Publicity chair)
- IEEE Pacific Visualization Symposium 2008: Jean-Daniel Fekete (Program Committee member)

## 9.5. Workshop organization

- Information Visualization Software Infrastructure Workshop at the Visualization Summit 2007 in Zürich: Jean-Daniel Fekete, in collaboration with Katy Börner and Bruce Herr (Indiana Univ.)
- Information Visualization - Human-Centered Issues in Visual Representation, Interaction, and Evaluation Jean-Daniel Fekete, Andreas Kerren, Chris North and John Stasko Dagstuhl Seminar 07221, Dagstuhl, Germany, May 28-Jun 1 2007
- Atelier Visualization et Extraction de Connaissances, associé à EGC 2008 INRIA Sophia Antipolis Nice, 29 Janvier 2008: Thanh-Nghi Do, en collaboration avec François Poulet (IRISA Rennes) et Bénédicte Le Grand (LIP6)

## 9.6. Conference reviewing

- ACM CHI 2007: Jean-Daniel Fekete, Niklas Elmqvist
- ACM UIST 2007: Pierre Dragicevic, Jean-Daniel Fekete, Nathalie Henry
- Conférence Francophone d'Interaction Homme-Machine (IHM) 2007: Jean-Daniel Fekete
- IEEE Conference on Information Visualization 2007: Jean-Daniel Fekete, Niklas Elmqvist
- IEEE Symposium on Visual Analytics Science and Technology 2007: Jean-Daniel Fekete, Niklas Elmqvist
- VIEW (Visual Information Expert Workshop) 2007: Jean-Daniel Fekete, Thanh-Nghi Do
- Asia-Pacific Symposium on Visualisation 2007: Jean-Daniel Fekete
- Eurographics Eurovis 2007: Jean-Daniel Fekete
- EGC (Extraction et Gestion de Connaissances) 2008: Thanh-Nghi Do
- Atelier QDC, EGC 2008: Thanh-Nghi Do
- EGC (Extraction et Gestion de Connaissances) 2007: Thanh-Nghi Do
- Atelier Visualisation et ECD, EGC 2007: Thanh-Nghi Do
- Eurographics 2008: Niklas Elmqvist, Nathalie Henry

- Graphics Interface 2007: Niklas Elmqvist
- IEEE PacificVis 2008: Niklas Elmqvist, Jean-Daniel Fekete
- IEEE Conference on Virtual Reality 2007: Niklas Elmqvist
- ACM Symposium on Virtual Reality Software and Technology: Niklas Elmqvist

## 9.7. Scientific associations

- AFIHM (French speaking HCI asssociation): Jean-Daniel Fekete, Executive Committee members

## 9.8. Evaluation committees and invited expertise

- MDD program (ANR, National Research Agency): Jean-Daniel Fekete, member of the evaluation committee since 2005
- LIP6, Paris: Jean-Daniel Fekete, member of the evaluation committee

## 9.9. PhD defenses

- Davis Da Costa (Univ. de Tours), Ph.D. Thesis “Visualisation et fouille interactive de données à base de points d’intérêts”: Jean-Daniel Fekete, jury member
- Pierre Salom (Univ. de Bordeaux), Ph.D. Thesis “Visualisation interactive de données volumiques texturées pour la détection supervisée de failles en imagerie sismique”: Jean-Daniel Fekete, reviewer

# 10. Bibliography

## Major publications by the team in recent years

[1] A. BEZERIANOS, P. DRAGICEVIC, R. BALAKRISHNAN. *Mnemonic rendering: an image-based approach for exposing hidden changes in dynamic displays*, in "UIST ’06: Proceedings of the 19th annual ACM symposium on User interface software and technology, New York, NY, USA", ACM, 2006, p. 159–168.

[2] J.-D. FEKETE. *The InfoVis Toolkit*, in "Proceedings of the 10th IEEE Symposium on Information Visualization (InfoVis 04), Austin, TX", IEEE Press, October 2004, p. 167-174, http://www.lri.fr/~fekete/ps/ivtk-04.pdf.

[3] J.-D. FEKETE, C. PLAISANT. *Interactive Information Visualization of a Million Items*, in "Proc. IEEE Symposium on Information Visualization 2002 (InfoVis 2002), Boston, USA", IEEE Press, October 2002, p. 117-124.

[4] M. GHONIEM, J.-D. FEKETE, P. CASTAGLIOLA. *Readability of Graphs Using Node-Link and Matrix-Based Representations: Controlled Experiment and Statistical Analysis*, in "Information Visualization Journal", vol. 4, n$^{o}$ 2, 2005, p. 114–135.

## Year Publications

### Articles in refereed journals and book chapters

[5] T.-N. DO, J.-D. FEKETE. *V4Miner pour la fouille de données*, in "numéro spécial de la revue RIA, Revue d’Intelligence Artificielle", à paraître, 2008.

[6] T.-N. DO, N.-K. PHAM, F. POULET. *Exploration interactive de résultats d'arbre de décision*, in "Revue des Nouvelles Technologies de l'Information (RNTI-E-9) — Série Extraction et Gestion des Connaissances", vol. 2, 2007, p. 157-168.

[7] T.-N. DO, F. POULET. *Classification de grands ensembles de données avec un nouvel algorithme de SVM*, in "Revue des Nouvelles Technologies de l'Information (RNTI-E-9) — Série Extraction et Gestion des Connaissances", (Best paper of EGC'07), vol. 2, 2007, p. 739-750.

[8] T.-N. DO, F. POULET. *Interval Data Mining with Kernel-based Algorithms and Visualization*, Z. R. D. A. ZIGHED, H. HACID (editors), to appear, Idea Group Inc., 2008.

[9] N. ELMQVIST, P. TSIGAS. *CiteWiz: A Tool for the Visualization of Scientific Citation Networks*, in "Information Visualization", to appear, 2007.

[10] N. ELMQVIST, P. TSIGAS. *View-Projection Animation for 3D Occlusion Management*, in "Computers and Graphics", to appear, 2007.

[11] N. ELMQVIST, M. E. TUDOREANU. *Occlusion Management in Immersive and Desktop 3D Virtual Environments: Theory and Evaluation*, in "International Journal of Virtual Reality", vol. 6, 2007, p. 21–32.

[12] N. HENRY, J.-D. FEKETE, M. J. MCGUFFIN. *NodeTrix: a Hybrid Visualization of Social Networks*, in "IEEE Transactions on Visualization and Computer Graphics", Brian Shackel Award, vol. 13, n$^{o}$ 6, 2007, p. 1302-1309.

[13] N. HENRY, H. GOODELL, N. ELMQVIST, J.-D. FEKETE. *20 Years of four HCI conferences: A Visual Exploration*, in "International Journal of Human-Computer Interaction — Reflections on Human-Computer Interaction: A special issue in honor of Ben Shneiderman's 60th birthday", to appear, 2008.

[14] A. KERREN, J. T. STASKO, J.-D. FEKETE, C. NORTH. *Workshop report: information visualization-human-centered issues in visual representation, interaction, and evaluation*, in "Information Visualization", October 2007, p. 1473–8724.

[15] N.-K. PHAM, T.-N. DO, F. POULET, A. MORIN. *Tree-view pour l'exploration interactive des arbres de décision*, in "numéro spécial de la revue RIA, Revue d'Intelligence Artificielle", à paraître, 2008.

[16] C. PLAISANT, J.-D. FEKETE, G. GRINSTEIN. *Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository*, in "IEEE Transactions on Visualization and Computer Graphics", To appear, vol. 14, n$^{o}$ 1, 2008, p. 120–134, http://doi.ieeecomputersociety.org/10.1109/TVCG.2007.70412.

[17] F. POULET, T.-N. DO. *Interactive Decision Tree Construction for Interval and Taxonomical data*, in "Visual Data Mining: Theory, Techniques and Tools for Visual Analytics", to appear, vol. 4404, 2007.

## Publications in Conferences and Workshops

[18] C. APPERT, J.-D. FEKETE. *Naviguer dans des grands arbres avec ControlTree*, in "Proceedings of IHM 2007, 19ème conférence francophone sur l'Interaction Homme-Machine", ACM Press, International Conference Proceedings Series, November 2007, p. 139–142.

[19] T.-N. DO, J.-D. FEKETE. *Flot visuel de données*, in "Acte du 5ème Atelier Visualisation et extraction de connaissances, EGC'07", January 2007, p. 21-30.

[20] T.-N. DO, J.-D. FEKETE. *Large Scale Classification with Support Vector Machine Algorithms*, in "Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA'07)", to appear, IEEE Press, December 2007.

[21] T.-N. DO, F. POULET, J.-D. FEKETE. *Massive Data Mining via Boosting of Least Squares SVM Algorithm*, in "Proceedings of the 5th IEEE International Conference on Computer Sciences: Research & Innovation — Vision for the Future (RIVF'07)", IEEE Press, March 2007, p. 47-52.

[22] N. ELMQVIST, U. ASSARSSON, P. TSIGAS. *Employing Dynamic Transparency for 3D Occlusion Management: Design Issues and Evaluation*, in "Human-Computer Interaction – INTERACT 2007", C. BARANAUSKAS, P. PALANQUE, J. ABASCAL, S. D. J. BARBOSA (editors), LNCS, vol. 4662, Springer, 2007, p. 532–545.

[23] N. ELMQVIST, J. STASKO, P. TSIGAS. *DataMeadow: A Visual Canvas for Analysis of Large-Scale Multivariate Data*, in "Proceedings of the IEEE Symposium on Visual Analytics Science and Technology 2007", IEEE Press, 2007, p. 187–194.

[24] N. ELMQVIST, P. TSIGAS. *A Taxonomy of 3D Occlusion Management Techniques*, in "Proceedings of the IEEE Conference on Virtual Reality", IEEE Press, 2007, p. 51–58.

[25] N. ELMQVIST, P. TSIGAS. *TrustNeighborhoods: Visualizing Trust in Distributed File Systems*, in "Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization 2007", IEEE Press, 2007, p. 107–114.

[26] N. ELMQVIST, M. E. TUDOREANU, P. TSIGAS. *Tour Generation for Exploration of 3D Virtual Environments*, in "Proceedings of the ACM Symposium on Virtual Reality Software and Technology 2007", to appear, ACM Press, 2007.

[27] T. GROSSMAN, P. DRAGICEVIC, R. BALAKRISHNAN. *Strategies for accelerating on-line learning of hotkeys*, in "CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA", ACM, 2007, p. 1591–1600.

[28] N. HENRY, J.-D. FEKETE. *MatLink: Enhanced Matrix Visualization for Analyzing Social Networks*, in "Human-Computer Interaction – INTERACT 2007", C. BARANAUSKAS, P. PALANQUE, J. ABASCAL, S. D. J. BARBOSA (editors), LNCS, (Brian Shackel Award), vol. 4663, Springer, 2007, p. 288–302.

[29] W. MACKAY, C. APPERT, M. BEAUDOUIN-LAFON, O. CHAPUIS, Y. DU, J.-D. FEKETE, Y. GUIARD. *TouchStone: Exploratory Design of Experiments*, in "Proceedings of ACM CHI 2007 Conference on Human Factors and Computing Systems", ACM Press, April 2007, p. 1425-1434, http://doi.acm.org/10.1145/1240624.1240840.

[30] N.-K. PHAM, T.-N. DO, F. POULET, A. MORIN. *Interactive Exploration of Decision Tree Results*, in "Proceeding of the 12th International Conference on Applied Stochastic Models and Data Analysis (ASMDA'07)", 2007.

[31] L. TREMBLAY, P. DRAGICEVIC, M. MCGUFFIN. *Ballistic and Current-control Phases of Aiming and Throwing*, in "12th International Conference of the Association of Researchers in Physical and Sporting Activities (ACAPS '07)", 2007.

[32] S. ZHAO, P. DRAGICEVIC, M. CHIGNELL, R. BALAKRISHNAN, P. BAUDISCH. *Earpod: eyes-free menu selection using touch input and reactive audio feedback*, in "CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA", ACM, 2007, p. 1395–1404.

### Internal Reports

[33] J.-D. FEKETE, N. ELMQVIST, T.-N. DO, H. GOODELL, N. HENRY. *Navigating Wikipedia with the Zoomable Adjacency Matrix Explorer*, Technical report, n$^{o}$ RR-6163, INRIA Research Report (Paris), 2007, http://hal.inria.fr/inria-00141168/en.

### Miscellaneous

[34] F. FRANÇOIS. *Conception et développement d'une plateforme d'encodage et de valorisation de documents historiques*, October 2007, Rapport de stage D2I de l'IFSIC, Université de Rennes 1.

[35] L. HAUCHEMAILLE. *Utiliser la visualisation pour aider l'utilisateur à mieux comprendre son flot de communication email : une extension pour Columba*, June 2007, Rapport de stage IUT d'Orsay.

[36] A. KERREN, J. T. STASKO, J.-D. FEKETE, C. NORTH. *07221 Abstracts Collection – Information Visualization - Human-Centered Issues in Visual Representation, Interaction, and Evaluation*, in "Information Visualization - Human-Centered Issues in Visual Representation, Interaction, and Evaluation", J.-D. FEKETE, A. KERREN, C. NORTH, J. T. STASKO (editors), Dagstuhl Seminar Proceedings, n$^{o}$ 07221, Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2007, http://drops.dagstuhl.de/opus/volltexte/2007/1136.

[37] A. KERREN, J. T. STASKO, J.-D. FEKETE, C. NORTH. *07221 Executive Summary - Information Visualization - Human-Centered Issues in Visual Representation, Interaction, and Evaluation*, in "Information Visualization - Human-Centered Issues in Visual Representation, Interaction, and Evaluation", J.-D. FEKETE, A. KERREN, C. NORTH, J. T. STASKO (editors), Dagstuhl Seminar Proceedings, n$^{o}$ 07221, Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2007, http://drops.dagstuhl.de/opus/volltexte/2007/1135.

[38] Y. LHEUDÉ. *Intégration d'outils d'analyse de données dans un flot visuel de données*, June 2007, Rapport de stage IUT d'Orsay.

## References in notes

[39] S. K. CARD, J. D. MACKINLAY, B. SHNEIDERMAN (editors). *Readings in information visualization: using vision to think*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[40] J. BERTIN. *Sémiologie graphique : Les diagrammes - Les réseaux - Les cartes*, Les réimpressions, Editions de l'Ecole des Hautes Etudes en Sciences, Paris, France, 1967.

[41] B. E. BOSER, I. M. GUYON, V. N. VAPNIK. *A training algorithm for optimal margin classifiers*, in "COLT '92: Proceedings of the fifth annual workshop on Computational learning theory, New York, NY, USA", ACM, 1992, p. 144–152.

[42] G. CAUWENBERGHS, T. POGGIO. *Incremental and Decremental Support Vector Machine Learning*, in "NIPS", 2000, p. 409-415, http://citeseer.ist.psu.edu/cauwenberghs00incremental.html.

[43] T.-N. DO, F. POULET. *Classifying one Billion Data with a New Distributed SVM Algorithm*, in "proc. of RIVF'06, 4th IEEE International Conference on Computer Science, Research, Innovation and Vision for the Future, Ho Chi Minh, Vietnam", 2006, p. 59–66.

[44] J.-D. FEKETE, C. PLAISANT. *Les leçons tirées des deux compétitions de visualisation d'information*, in "Proceedings of IHM 2004, Namur, Belgium", International Conference Proceedings Series, ACM Press, September 2004, p. 7-12.

[45] Y. FREUND, R. E. SCHAPIRE. *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, in "EuroCOLT '95: Proceedings of the Second European Conference on Computational Learning Theory, London, UK", Springer-Verlag, 1995, p. 23–37.

[46] J. J. GIBSON. *The Ecological Approach to Visual Perception*, Lawrence Erlbaum Associates, New Jersey, USA, 1979.

[47] G. H. GOLUB, C. F. V. LOAN. *Matrix computations (3rd ed.)*, Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[48] Y. GUIARD, Y. DU, J.-D. FEKETE, M. BEAUDOUIN-LAFON, C. APPERT, O. CHAPUIS. *Shakespeare's Complete Works as a Benchmark for Evaluating Multiscale Document-Navigation Techniques*, in "Proceedings of BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV'06), Venice, Italy", ACM Press, May 2006, p. 65-70.

[49] N. HENRY, J.-D. FEKETE. *Evaluating Visual Table Data Understanding*, in "Proceedings of BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV'06), Venice, Italy", ACM Press, May 2006, 6 pages, to be publised.

[50] P. LYMAN, H. R. VARIAN. *How Much Information*, 2003, http://www.sims.berkeley.edu/how-much-info-2003.

[51] C. PLAISANT, B. LEE, C. SIMS PARR, J.-D. FEKETE, N. HENRY. *Task Taxonomy for Graph Visualization*, in "Proceedings of BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV'06), Venice, Italy", 6 pages, to be published, ACM Press, May 2006.

[52] J. C. PLATT. *Fast training of support vector machines using sequential minimal optimization*, MIT Press, Cambridge, MA, USA, 1999, p. 185–208.

[53] J. A. K. SUYKENS, J. VANDEWALLE. *Least Squares Support Vector Machine Classifiers*, in "Neural Process. Lett.", vol. 9, n$^{o}$ 3, 1999, p. 293–300.

[54] M. TAKATSUKA. *A component-oriented software authoring system for exploratory visualization*, in "Future Generation Computer Systems: Journal of Grid Computing: Theory, Methods and Applications", vol. 21, n$^{o}$ 7, July 2005, p. 1213–1222.

[55] S. TONG, D. KOLLER. *Support Vector Machine Active Learning with Applications to Text Classification*, in "J. Mach. Learn. Res.", vol. 2, 2002, p. 45–66.

[56] A. TRIESMAN. *Preattentive Processing in Vision*, in "Computer Vision, Graphics, and Image Processing", vol. 31, n$^{O}$ 2, August 1985, p. 156-177.

[57] E. TUFTE. *The Visual Display of Quantitative Information*, Graphics Press, 1983.

[58] J. W. TUKEY. *Exploratory Data Analysis*, Addison-Wesley, 1977.

[59] V. N. VAPNIK. *The nature of statistical learning theory*, Springer-Verlag New York, Inc., New York, NY, USA, 1995.