Visual Analytics

Thorsten May, Martin Knuth

German Abstract

Mit Messgeräten, Sensoren und verschiedensten Computern erzeugen wir heute eine ungeheure Menge an Daten. Das Sammeln und Speichern dieser sehr großen Datenmengen ist kein Selbstzweck: Vielmehr verbergen sich hinter diesen Daten interessante und wichtige Informationen, die mit der steigenden Komplexität der Datenräume viel schwerer auszuwerten sind. Diese Informationen müssen durch geeignete Visualisierungen und Interaktionsmetaphern bestmöglich dargestellt werden. In vielen Szenarien ist die große Datenmenge eine der größten Schwierigkeiten, die zu bewältigen ist. Unsere Abteilung arbeitet schon seit einigen Jahren an Werkzeugen für die parallele Visualisierung großer Datenmengen. Mit dem Visual Analytics Cluster haben wir nun ein Werkzeug, das flexibel genug ist, um die Visualisierungspipeline im Rechnerverbund effizient abzuarbeiten.

Introduction

The increasing power of computer, simulation, and sensor technology allows us to produce and store data at ever growing rates. Though in virtually all application areas, data acquisition has no value in itself: Actually the data hides the important information, which is more and more difficult to come by. Huge investments in time and money must be discarded, because we still lack the possibilities to deal with the databases properly. The complexity of the data and their relationship to the underlying information often allows us to exploit a fraction of the acquired data. This process – the extraction of information from patterns to gain insight – cannot be done without human intervention. Hence, the commonly propagated statement claiming that we live in the information age is only part of the truth. The future is likely to supply us with even more data. As the bandwidth of the human senses and human abilities remain constant, the presentation must provide efficient and effective means to convey, manipulate, and evaluate information. Decision makers, analysts, and engineers are often confronted with vast amounts of complex, seemingly inconsistent or incomplete data, which is derived from a number of heterogeneous sources. The goal of the analysis is a transparent knowledge, which can be applied in the context of future decisions.

The Challenge

Visual Analytics strives to bridge the gap between the machine and the human mind, while taking advantage of their native capabilities. On the technological level, developments in visualization, data processing, and analysis have grown to maturity in the past decade and are integrated into this new field of research. On the other hand, we still rely on interaction metaphors developed 15 years ago and it is questionable if they are able to meet the demands of the ever increasing mass of information to be sought for. Visual Analytics incorporates more »intelligent« means than retrieving and displaying a set of data items, even if the visualization is viable.

In most Visual Analytics application scenarios the large amount of data to be processed is one of the major challenges when developing an efficient tool. For an interactive visual analysis it is necessary to provide an immediate response to any query. One goal is to extract the knowledge from the database the user currently focuses on and display it on the screen. The lack or uncertainty of data in a given scenario adds to the complexity of decision making. Suitable models must be created to cover theses problems. However, such models will be not be a fixed part

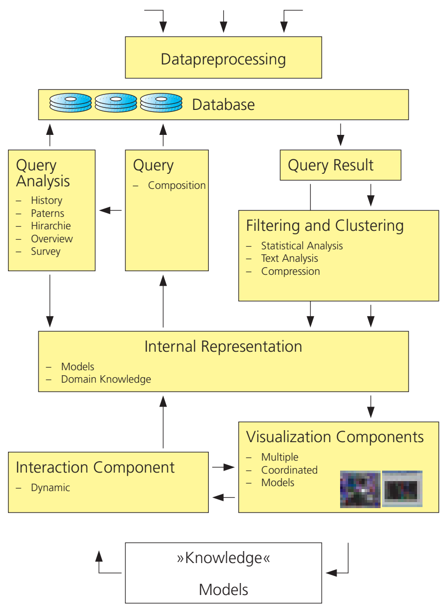

Figure 1: Basic Architectural Concept

within the analysis tool: Experience shows that such models have to be evaluated for every single case – possibly under circumstances completely different from the ones under which they have been established. The new experience gained with new data must be

– Visually represented in a way that matches the user's sketch of the model most closely
– Manipulated to incorporate and test more refined models, based on new data available
– Evaluated to clearly present the influence of data and models to the final results
– Easy to apply to a decision case

These issues do not only apply to the models on the treatment of uncertain information, but to the decision making process as a whole. Its feasibility derives from the direct interaction of the human user with the analysis tool. Human abilities to detect patterns in data which is adequately refined and represented, his flexibility, creativity, and knowledge is augmented by the capacity and computational power of today's computers.

Information visualization reproduces the internal structure of the data and the domain of the user and combines them into an interactive, responsive visualization. Visual Analytics still goes beyond that point: While the information visualization methods can be used to extract the knowledge to start hypotheses, Visual Analytics strives to visually represent the data and the knowledge (rules, models) about the data, to foster, to evaluate, and to refine insights about the information. The results of creating a decision support tool do not only include new knowledge or the decision rules derived from this knowledge, but also the path which leads to it.

Application Areas

By nature, the application areas for the field of Visual Analytics can be any area which at least partly incorporates the automated acquisition of data. While, in the past, basic research sometimes focused on applications of direct interest and use for the developer, we can expect more and more applications to come from a large number of other domains, including, for example, biotechnology, business & financial market analysis, or GIS. The wide range of possible applications is a challenge in itself, because the visual representations and interaction metaphors have to be tailored for virtually every costumer. Consequently one of our first efforts is to provide an architecture base which can be re-used for different scenarios. A conceptual design of the architecture is depicted in figure 1. It is derived from the commonly known simulation loop. User interactions trigger queries to the database which, in turn, are transformed into a visual representation of the data. The internal representation of the models form the core of the loop. This is where the different representations are matched to each other. To provide an effective analysis, it must be guaranteed that the loop can be processed in real-time. To provide the computational power necessary to cope with huge databases, we built up the Visual Analytics lab with a 32 node cluster.

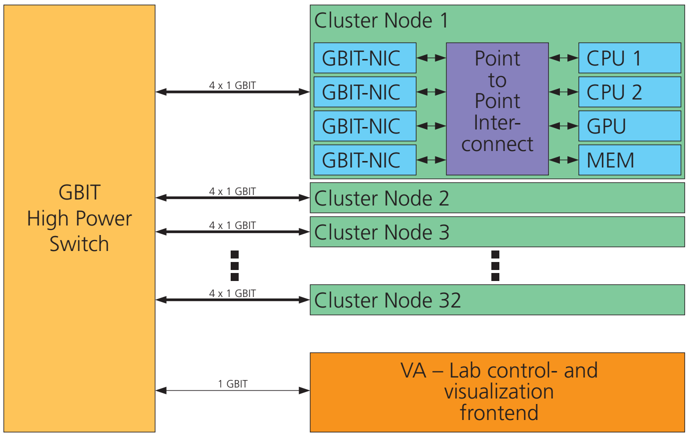

Figure 2: In the center, a high performance switch is used to couple the Nodes and the VA-Laboratory, which is used for controlling and visualization of the results.

Laboratory

Cluster-based systems are an effective data processing platform to keep response latency at a minimum. However, cluster systems often have a need for specialized communication interconnects. This makes it difficult to construct them with »out of the shelf« computers and increases the purchase cost dramatically. Additionally, existing cluster-based software is only one component of a Visual Analytics system. In a VA system several of these components are needed, which greatly increases communication effort. Since such an interaction component is not completely developed yet, a main issue for constructing the VA laboratory was the construction of a cost effective cluster system using standard hardware which remains open for a wide range of applications.

The cluster consists of 32 dual processor workstations each equipped with a high end graphics accelerator (GPU) each. This enables the cluster to perform all Visual Analytics tasks inside the cluster system, including the last step – the visualization. Alternatively the GPU can be used to act as a co-processor. A distributed database can be coupled with filter functions or high performance visualization on the same system. To be cost effective in construction, a standard GBit network interconnection was used. To get the desired connection speed a channel bonding of 4 x 1 GBit circuits is established on every node. Hardware which normally is used for servers, guarantees the maximum bandwidth of 64 GBit/sec. The internal chipset is able to handle the extremely high amount of communication within such a system. This allows a collaboration of 4 network cards, 2 processors, GPU, and memory at the same time.

Point of Contact

Dr.-Ing. Jörn Kohlhammer
Fraunhofer IGD, Darmstadt, Germany
E-mail:
joern.kohlhammer@igd.fraunhofer.de