Original Article

Data transformations and representations for computation and visualization

David J. Kasik[a,*]
David Ebert[b]
Guy Lebanon[c]
Haesun Park[c] and
William M. Pottenger[d]

[a]The Boeing Company, PO Box 3707, Seattle, WA 98124, USA.
[b]Purdue University, West Lafayette, IN 47907, USA.
[c]Georgia, Institute of Technology, Atlanta, GA 30332, USA.
[d]Rutgers University, Piscataway, NJ 08854, USA.

*Corresponding author.
E-mail: david.j.kasik@boeing.com

Abstract At the core of successful visual analytics systems are computational techniques that transform data into concise, human comprehensible visual representations. The general process often requires multiple transformation steps before a final visual representation is generated. This article characterizes the complex raw data to be analyzed and then describes two different sets of transformations and representations. The first set transforms the raw data into more concise representations that improve the performance of sophisticated computational methods. The second transforms internal representations into visual representations that provide the most benefit to an interactive user. The end result is a computing system that enhances an end user's analytic process with effective visual representations and interactive techniques. While progress has been made on improving data transformations and representations, there is substantial room for improvement.

Information Visualization (2009) **8,** 275 – 285. doi:10.1057/ivs.2009.27

Keywords: algorithms; visual metaphors; data characteristics; visual representations; synthesis

Introduction

Visual analytics systems must integrate a number of different computing capabilities. In many ways, a visual analytics system is similar to other complex systems that people use daily. When abstracted, systems have user interface, algorithmic and data components. When dissected more completely, systems differ in terms of the tasks that a user must perform to transform data into more meaningful forms.

Because of certain data characteristics, a wide range of algorithmic approaches are needed to transform the raw data into increasingly concise representations that are then transformed into visual representations that users examine to obtain insight.

This article examines specific types of raw data and the types of computational and visualization transformations and representations that improve a user's analytic ability. There are two different types of transformations and representations. The first is used to identify higher-order characteristics in the data, such as relationships, trends, summaries, clusters and synopses. The second is responsible for transforming data into the visual representations that help the user navigate the overall data space. Both types of transformations and representations must cope with scale and complexity.

Raw Data Characteristics

Computers store, move and analyze data that, on initial examination, are a simple collection of bits. Collections of bits are organized into different units (files, directories, databases, network packets and so on).

This article is a product of a workshop on the Future of Visual Analytics, held in Washington, DC on 4 March 2009. Workshop attendees included representatives from the visual analytics research community across government, industry and academia. The goal of the workshop, and the resulting article, was to reflect on the first 5 years of the visual analytics enterprise and propose research challenges for the next 5 years. The article incorporates input from workshop attendees as well as from its authors.

Received: 27 May 2009
Revised: 8 July 2009
Accepted: 8 July 2009

These collections of bits form primitive data types[1] that include text, numbers, still images, audio and video. Combinations of primitive data forms can be:

- Structured (for example, relational tables, geometry). Often contains numeric values. Some fields may contain relatively small amounts of free-form text.
- Semi-structured (for example, e-mail that contains header data, attachments and text; network packets headers and payloads; scientific data resulting from simulations).
- Unstructured (for example, a collection of text).

Many of the challenges[1], especially in dealing with textual data, still exist. This article examines a number of algorithmic approaches organized around key data characteristics. The characteristics apply to all primitive data forms rather than algorithms that apply to specific data types. The article adds the notion of inserting a user-in-the-algorithmic-loop to help guide the raw data transformation process. In addition, it introduces a set of transformations needed to produce effective visual representations. Transforming data into an effective visual representation is fundamentally different from transforming incoming raw data.

When defining approaches for data transformations and representations, algorithm designers must consider that visual analytics systems are interactive in nature, which makes algorithms that are sufficiently fast enough to interactive performance critical. Interactive users expect a response for a simple task in a few seconds or less and are more patient when they realize that the computer is performing a complex computation. Even so, tasks that take more than a few minutes can lead to user frustration and reduced productivity.

In addition to the raw performance needed to support interactive analytics, characteristics of the data itself affect the transformations and representations for both computation and visualization. The key characteristics are:

Massive data. The amount of data that may be pertinent to a specific analysis task is potentially unlimited. Even though the vast majority of data may be trivially rejected, the data volume can easily range from megabytes to petabytes. Some analysts must make decisions based on a relatively small amount of data (for example, a safety engineer looking at commercial aircraft incident reports), while others require terabytes (for example, an administrator looking at event logs for network intrusions). Massive data sets must often be transformed into a smaller number of dimensions or aggregated to allow users to cope with the scale.

Geospatial and temporal data. Significant amounts of data have location and temporal dependencies. Both geospatial and temporal data are dynamic, and understanding the evolution of value changes is often important. Examples include:

- A snapshot of a given data set (for example, a large set of documents) freezes geo-location and time at a specific point.
- A series of snapshots (for example, transaction-based systems, the web) that evolves over time.
- Streaming data (for example, real-time sensors, network data) are collected continuously, which increases data volume.

Geospatial data gives an analyst critical understanding of the physical location of specific event occurrences. When coupled with temporal data, significant patterns of activity may emerge using implicit methods (for example, kernel methods) and explicit methods (for example, feature combinations, supervised learning).

Imperfect data. The data, regardless of volume, often contain noisy, missing, erroneous, incomplete or deliberately misleading values. Text data are particularly difficult. The values in a given text field or document range from cleanly edited to quick-and-dirty entries. Shorthand and abbreviations are often present, especially in data that are pertinent to a specific domain. For example, consider the variation in language among medical records, airplane incident reports and cell phone text messages. Different natural languages pose a problem because text can be entered carelessly or erroneously by either native or non-native speakers. Analysts often gain insight from data anomalies, and the analyst is responsible for determining whether the unusual data are informative or extraneous. A significant amount of work is needed in this area, although some techniques, such as those based on vector norm formulations, are available.

Heterogeneous data. Analysts must often gain significant insight from multiple data sources. In some cases, integrating the schemas may be possible. Even if schema integration is possible, multiple data sources increase the raw data volume and increase the probability that specific fields or values will have conflicting meanings. Furthermore, the methods needed to assemble the data in a heterogeneous data environment generally differ from one another on a data store-by-data store basis. Extracting, translating and loading (ETL) the data into the visual analytics system may take longer than the analysis itself. The long duration for ETL may even cause currency problems with the data and negatively affect temporal trend analyses. Two promising approaches are to use a joint probabilistic model for different attributes and to carefully build a dissimilarity metric.

User-in-the-loop. The data transformations and representations that apply to basic analysis tasks are different from those that produce a visual display. In addition, data volume differs from visual volume. For example, consider a set of data that captures network traffic. The transformations and representations that produce various

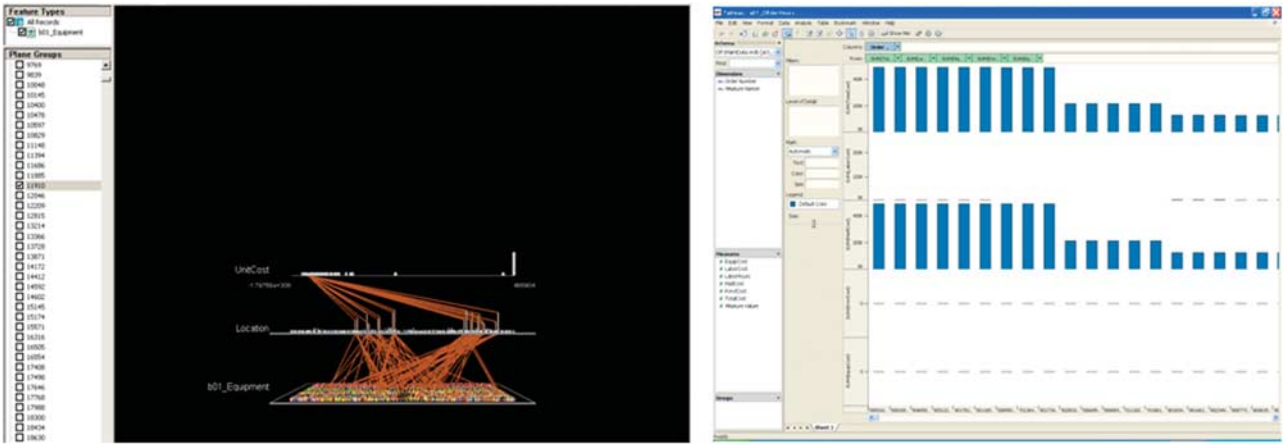

Figure 1: Complex vs. simple displays.

summaries are fundamentally different from those that produce images of the network traffic.

The visuals themselves can vary significantly as shown in Figure 1. The image on the left offers a visual technique to show relationships between specific values in a large table. The image on the right uses a traditional histogram to show the numeric order of specific fields in a relational table. Different transformations produced the different visual representations.

Adding a user-in-the-loop can help direct the analysis when the user has specific domain knowledge. Because of the breadth of visual analytics applicability, determining generalized methods for data and visual transformations and representations is challenging. Domain knowledge often leads to simplifying assumptions and customizations that improve both computational and visualization transformation accuracy and performance.

Transforming and Representing Data for Computation

Successful data transformation and representation methods combine mathematical, statistical and linguistic analysis with hardware and software techniques to handle massive data, geospatial and temporal data, imperfect data, heterogeneous data, and users-in-the-loop. The combination poses significant research challenges. This section discusses approaches and shortcomings in those approaches that require additional work.

Massive data

A major challenge arises from the sheer volume of data.[2] The size and complexity of the data sets appearing now and in the future are an impediment to the full exploitation of visual analytics. This section focuses on the computational and algorithmic methods needed to distill information from ever-expanding data streams. The need for scalable data representation and transformation methods forces the development of new paradigms that will enable major improvements in decision-making processes through better methods for understanding and predicting outcomes in complex situations and scenarios.

Achieving interactive performance adds further complexity. Complex operations on large data sets today often require minutes and even hours to perform. Long response times render such operations ineffective in highly interactive environments. Furthermore, data sets are becoming more massive and complex over time, necessitating development of scalable algorithms that are implementable on parallel computers.

Massiveness of data also refers to its high dimensionality. While humans are excellent at finding patterns in a 2D- or 3D space, they have difficulty processing massive amounts of data in higher dimensions. Dimension reduction generally is achieved through *feature extraction* that creates new coordinate spaces through linear and nonlinear transformations or *feature selection* that identifies an important subset of features from the original high-dimensional data set. Dimension reduction is often used to improve efficiency in computational cost and storage complexity, noise reduction, or noise removal. It can produce improved accuracy and is essential for 2D and 3D visualization of data. Dimension reduction methods may differ when applied to visual analytics rather than automated analysis.

For data sets for which there is no *a priori* knowledge, dimension reduction methods such as Principal Component Analysis (PCA)[3] and Latent Semantic Indexing (LSI)[4] provide theoretically well-justified projections of high-dimensional data onto lower-dimensional spaces. Both PCA and LSI are based on the Singular Value Decomposition (SVD).[5] SVD, on which many methods are based, is a powerful mathematical tool in understanding the space spanned by the data represented in a vector space. It provides a method to capture the rank, orthonormal bases and characteristics of the noise space associated with the space spanned by the data. SVD has been used extensively in numerous science and engineering problems, including signal, image and text processing. When additional information concerning characteristics of the data such as its cluster structure or the fact that data values are always non-negative is available, dimension reduction methods that reveal this fact can achieve better results. Two examples are Linear Discriminant Analysis (LDA)[6] for clustered data and non-negative matrix factorization[7,8] for non-negative data. If the inter-data relationship is not linear, nonlinear extensions such as Kernel PCA and Kernelized Discriminant Analysis may be used.

To reveal nonlinear structure in the data, many promising methods such as manifold learning have been developed. In manifold learning, the goal is to find a lower-dimensional (typically nonlinear) representation of the data given in a high-dimensional space. A rich literature exists in this area, and the most widely used methods include multi-dimensional scaling ISOMAP, locally linear embedding, Laplacian eigenmap, and local tangent space alignment. Typically, in these manifold learning methods, the dimension-reducing nonlinear transformations are not explicitly available. In other linear and nonlinear dimension-reducing transformations such as PCA, LDA and their kernel counterparts, transformations are explicitly computed and therefore make representation of unseen data points in the same lower-dimensional space possible. Development of an effective and general asymptotic theory for manifold learning in terms of differential operators on manifolds can yield new algorithms for nonlinear dimension reduction and address many practical questions.

To make linear and nonlinear dimension reduction methods more effective in handling massive data, the basic characteristics of the dimension reduction methods for 2D or 3D representation of high-dimensional data sets must be understood. In many dimension reduction methods, the optimal reduced dimension, that is, the smallest acceptable reduced dimension with respect to the specific criterion of a dimension reduction method, is either unknown or much larger than 2 or 3. One may simply choose the leading two or three dimensions, but this may result in loss of information. This loss hinders understanding because the true characteristics of the data sets (for example, cluster structure, relationships, or anomalies) are hidden. Substantial research effort needs to be made for progress in this direction, although there are some promising preliminary results.[9]

Feature selection is another way to achieve dimension reduction. Unlike feature extraction, feature selection specifically *selects* a small number of relevant features. Feature selection algorithms typically perform feature ranking or subset selection. Feature ranking methods determine relevant features by a certain scoring metric and can be computationally expensive when the data dimension is very high. When feature selection is used to derive a 2D or 3D representation of the data, the results may not convey much information because too much data are discarded. The applicability of feature selection as a dimension reduction technique has not been extensively explored in visual analytics. Promising new methods can be expected to arise from the development of a comprehensive theory of automatic feature selection by sparse recovery. Such methods combine concepts from learning theory and can yield insights into new algorithms (for example, boosting, kernel machines).

One example where the data sets are represented in high-dimensional space is text. Text documents are originally represented as a sequence of words over a finite vocabulary V. This representation is problematic because documents of different lengths cannot be easily compared to one another. Instead, the first step in text analysis is to convert the documents into numeric vectors of fixed dimensionality. One option, leading to vector representation of dimensionality |V|, is to construct vectors whose components are the relative word frequency or normalized word counts in the document. A slight variation represents a document as a binary vector of dimensionality |V| whose components represent presence or absence of words. Higher-dimensional representations may be constructed by keeping track of appearances of short phrases, called *n*-grams.

Promising new methods can be expected to arise from the development of a fundamental comprehensive theory of automatic feature selection by sparse recovery. Such methods link together many ideas from learning theory and can yield insights into new algorithms such as boosting and kernel machines.

Many powerful new algorithms for dimension reduction pose even more difficult optimization problems than arise in current methods, leading to the need to solve very large-scale, semi-definite programming problems. Recent research has focused on the design of dimension-reduction methods that incorporate interpretability constraints such as sparsity and non-negativity. The resulting algorithms increase one's understanding of the transformations and further facilitate visual representation of very high-dimension data. In addition, incorporating expert opinion and necessary constraints in the problem formulation of dimension reduction is expected to produce more insightful representations of data.

Geospatial and temporal data

Complex geospatial and temporal data provide a wealth of information on complex phenomena that varies over time and/or place. Such data streams are called spatio-temporal multi-dimensional data (STMD). Geospatial and temporal data include dynamically changing location and/or time stamps as part of its metadata. STMD can be readily found in many real-world critical sources today, including

- sensor networks;
- event logging;

- human-activity logs that are becoming increasingly digital (for example, online news);
- less formal digital socializing (for example, web logs, RSS feeds, and sites like Twitter.com).

These applications and others like them reveal complex, time-series data that must be manually monitored for near real-time analytic results. It is possible to apply traditional algorithms to these data, but doing so typically pushes analytic results beyond near real-time application. Near real-time results can be accomplished through techniques such as sampling and aggregation. Such techniques often remove or further mask the important underlying semantic information analysts seek to discover. New computational transformations are needed to leverage such data in a near real-time visual analytics environment.

Kernel methods[10] have been applied as an implicit data transformation for STMD. A kernel function can be viewed as an implicit (nonlinear) mapping of data objects from the original input space to a high-dimensional feature space. The application of learning methods subsequently takes place in this feature space. The strength of kernel methods lies in their ability to expose hidden dependencies between input features relevant to the learning task. This in turn leads to simplification of the problem and improved performance of simple (for example, hyperplane-based) learning methods. However, applying a kernel-based data transformation causes latent relationships among input features to be distributed over a (sometimes infinite) number of dimensions of the feature space. A kernel only allows the computation of a certain aggregate quantity (the scalar product) in the feature space. Therefore, it is not possible to analyze the relations exposed by the kernel mapping between input features. Even though a variety of kernel functions have been developed, these methods are only appropriate for homogeneous data where the (dis)similarity between objects can be estimated as some average of (dis)similarities across all features. Finally, kernel methods are critically dependent on domain experts for construction of appropriate kernel functions. Extending kernel methods to overcome their shortcomings as applied to STMD is a significant research challenge.

In contrast to implicit STMD transformations, explicit transformation approaches can explicitly access the feature space and apply visualization and learning methods (such as Winnow[11] or Association Rule Mining[12]) that cannot be formulated in terms of vector products only.

Explicit transformations can be applied to other problematic data forms because explicit data transformations allow increased expressivity of features. One popular example is feature combination, which may be used for expansion of the base set of features in natural language.[13] This work demonstrated that such feature spaces allow for robust learning, whereas implicit kernel expansion of the feature space may lead to degradation in generalization performance if the dimensionality of the space is not controlled.

Another example of an explicit STMD transformation[14] builds a graph-based data representation[15], which considers a given data set as a bipartite graph. This approach increases the performance of supervised learning algorithms while leaving the data space's dimensionality unchanged. The latter aspect mitigates the exponential growth in dimensionality inherent in feature combination approaches. Vertices of one partition of the graph correspond to data instances. Vertices of the other partition correspond to features. Two vertices u and v are connected by an edge (u,v) if feature v has non-zero value in instance u. Unlike approaches that assume data instances are independent, this approach leverages higher-order co-occurrence relations between feature values across different instances and enables virtually any learning method to take advantage of this rich connectivity. Developing an unsupervised analogue will add further value to this approach.

Imperfect data

Effectiveness and accuracy of a solution should not be compromised in the name of achieving high efficiency whether dealing with massive or small volumes of data. The fact that most real-life data sets are noisy, corrupt and have missing values presents a challenge. In some cases, data may have been tampered with to be deliberately misleading. In addition, measures of accuracy are not always known because of the high complexity of the solution process in visual analytics.

Methods for representing the noise level in data may guide the analyst to ensure proper utilization of noisy data. Ideally, methods for noise reduction and noise removal can be applied. However, extreme caution must be taken because many existing practices are rather heuristic and often lack theoretical justification. Manually entered data, in contrast to physical data that comes from sensors, radio frequency identification devices, and the like, contain noise characteristics that cannot be well defined.

An even more difficult situation arises when the data set contains completely missing components. Many analysis algorithms assume complete knowledge of the data points. Use of such algorithms in the presence of missing values requires imputation methods. Effective information representation often comes from mathematical modeling of the problem and is constrained and driven by interactive visualization and analytical methods.

The choice of representation of noisy data should be guided by close collaboration with domain experts and an understanding of the users' needs so that they can be formulated in the model. Often these turn into large scale *constrained* optimization, matrix computation and graph theoretic problems. Robust algorithms that produce solutions that are insensitive to perturbations in input or conditions are needed, as are stable algorithms that reliably produce accurate solutions.

Another important challenge arises when there is the possibility of intentional disinformation or deception. In this case, the transformation and subsequent visualization should reflect the provenance and trustworthiness of the data. Data provenance[16] refers to the origin of the data and its movement and transformation from the point of origin to the visualization system. Source trustworthiness refers to the probability that the information source includes disinformation. Data trustworthiness refers to the probability that the received information was subjected to deception somewhere along the provenance path.

The trustworthiness of the source may be determined from historical data or human judgment. The trustworthiness of the received data may be computed from the provenance path and the trustworthiness of the sources along the path.

There are some similarities between imperfect or noisy data and deception. In the former, noisy data may be removed or modified before selecting the computational and visualization transformations. In the latter, the potential for deception and the trustworthiness of the different information sources are important factors that need to be considered. The suspected data may be removed or modified before deriving the optimal transformation. However, the data, their provenance, and trustworthiness need to be transformed and visualized along with the more reliable data.

For anomaly cleaning and detection, formulations based on various vector norms, especially the L1 norm, can be extended to achieve practical robust methods. Extensions to streaming, dynamic data and specific data types (for example, text, images) and data of mixed type need to be considered. Transforming imperfect data remains a continuing challenge in terms of reliable and robust results for visual analytics.

Heterogeneous data

Heterogeneous data occur in a number of different forms, which include:

- Nominal attributes that possess different sets of possible values. For example, medical records contain attributes with substantially different ranges of values.
- A combination of numeric and nominal values. For example, medical records may contain numeric attributes such as weight, height and age, along with nominal attributes such as ethnicity, symptom appearance and family history.
- Multiple attributes possessing different noise characteristics. For example, sensor network observations form a vector of measurements, where each component has a different noise model.
- A combination of quantitative and qualitative information. This is the case when quantitative physical measurements are combined with qualitative human judgment that takes the form of text.
- Attributes from multiple, merged databases. Joining databases for analysis is a difficult task that becomes even harder when similar attributes have different meanings.

Heterogeneity causes substantial difficulties in developing data transformation and dimensionality reduction techniques. Many techniques assume, either implicitly or explicitly, that the attributes are normally distributed. For example, PCA implicitly assumes a normal distribution because it is based on maximum likelihood estimation applied to a normal distribution. A similar observation applies to the k-means and Gaussian mixture clustering models. It is not immediately clear why the normal distribution is an appropriate assumption in cases of heterogeneous data. It is certainly a questionable assumption for nominal values.

A promising direction for deriving transformations for heterogeneous data is to first obtain a joint probabilistic model for the heterogeneous attributes. Probabilistic models for heterogeneous data include loglinear models and undirected graphical models[17,18] and Bayesian networks.[19] Once the model parameters are estimated using a technique like maximum likelihood, an appropriate transformation may be obtained by considering the model parameters. This approach can also be used to extend standard methods such as PCA. Examples include probabilistic PCA and exponential family PCA.

An alternative approach is to forgo the modeling process and to rely instead on a carefully constructed distance or dissimilarity measure. Such a measure may be used to derive an appropriate transformation in conjunction with multi-dimensional scaling.[20] Avoiding the need to construct a model for heterogeneous data and obtain the maximum likelihood parameters is a substantial advantage. A disadvantage is that the quality of the obtained transformation is in direct relation to the quality of the distance or dissimilarity measure. Constructing a sensible distance or dissimilarity for heterogeneous data may be a very challenging task. The use of domain knowledge or interactive feedback is likely to play a key role in designing effective distance or dissimilarity measures for heterogeneous data in visual analytics systems.

User-in-the-loop

The goal of a visual analytics system is not to perform analysis automatically but to facilitate it. A user-in-the-loop is therefore a central and critical element of visual analytics systems and must be in constant consideration throughout the design and implementation of such a system.

All of the above techniques take on an additional burden when placed in the context of a person. Humans have limited faculties (physical, mental and otherwise) that must be addressed by viable solutions if they are to be used in the context of visual analytics. For example, while the winner of the InfoVis 2003 Contest[21] could computationally compare two trees of 100 000 elements each, it also provided several interface methods to support a human's understanding and navigation. Data with high-order dimensionality must be reduced to two or three dimensions just to be displayed without losing key information after dimension reduction is performed.

People add a social dimension to visual analytics. Many organizations that perform large-scale analysis work in teams that may or may not be co-located. Some organizations may address distributed analysis over an organizational private network. Still other organizations, such as governmental agencies and public safety departments, require alternative solutions because of the geographical, legal and cultural boundaries that collaborative analyst sessions regularly cross. Therefore, there is a research need for systems that will facilitate multi-user collaborative distributed analysis safely, securely and legally.

Users must be able to trust visual analytics results. In line with the above comments regarding misleading data, 'trust' in this sense refers to the user's faith that the analytics system is transforming data in ways that maintain the original data's characteristics while foregoing adding artificial biases. Establishing and maintaining this trust is especially important for analysts who may be called to explain their analytical process to another decision maker (for example, a chief scientist, a lawmaker, a judge).

Users are dynamic and constantly change through analysis: their mental context, their model of the analyzed phenomenon and their focus or trust in various regions of data will often change through the course of analysis. This is especially true when analyzing data as a new piece of evidence, a new website discovered or a new laboratory result can quickly bring a new perspective on the current analytical context.

There are also physical constraints imposed by limited screen space with only two or three display dimensions. Limitations in human cognition capacity to communicate high-volume and high-dimensional data also present important challenges. Even with today's growing display size and resolution and the use of multiple monitors, display walls and CAVEs, the number of available pixels remains a fundamental limiting factor. The small screens on mobile devices used by first responders exacerbate the problem.

Methods for judiciously approximating or down-weighting large regions as appropriate to the analysis of interest will provide solutions to some of these demands. Clustering can provide a simple starting point toward organizing data into related groups for improved understanding and learning. Numerous clustering methods have been developed since the k-means algorithm was first published in 1955.[22] The best clustering approach is often very closely tied to the end goals of the intended users. For example, the task of binary clustering of a collection of animals may produce two completely different groups, such as mammals versus birds or predators versus non-predators, depending on the features used to represent the data.

In visual analytics, experts can often provide additional information. This can be realized by designing clustering methods that use human-specified constraints. Semi-supervised clustering formulates the problem in a way to satisfy cannot-link and must-link constraints. Methods that can incorporate additional expert input as constraints in the clustering problem formulation will provide more accurate representations of data. New approaches such as those based on multi-resolution data approximation for scalable data transformation operations using hierarchical data structures and multipole-like expansions provide promising directions.

The user-in-the-loop dimension of visual analytics is being extensively studied in the later phases of the analytical process. 'Sensemaking' systems and methods assist users in managing and making sense of their data.[23,24] Enhanced visualization techniques[25–27] are being developed to display and navigate through the complex, dynamic and temporal data prevalent today. However, all of these techniques and systems involve the user interactively only when the data have been collected and transformed into their (final) analytical representation. The possibility of including the user in the intermediate transformation and representation steps is an interesting one. The effect of this compared to fully automated approaches and the effect of this interaction on the analytical process are all open areas of research.

Recognizing and leveraging user dynamism provides significant benefit when done correctly. User modeling research[28,29] is still exploring strong guidelines for developing and maintaining an accurate model. With such a model, systems can adapt to the user's context and the machine's processing capability.[30,31] Systems could also use such modeling techniques to capture the user's mental state in the analytical process[32] and provide support for following best analytical practices. Integrating user modeling with visual analytics systems is still in its infancy and holds great potential.

Challenge synopsis

Challenges in data transformations and representations include:

- Maintaining transformation performance to sustain interactive rates, even when handling huge volumes of raw data.
- Because the same item may be interpreted differently across heterogeneous data stores, reconciling semantic inconsistencies across data stores.

- Uncertainty is caused by a number of different data characteristics. Estimating this uncertainty and communicating it in a meaningful way is an important challenge. Deriving value when the quality of the data varies significantly. For example, human language differences change the meaning of words in text, video and audio; noise in sensors affects numeric data.
- Developing provenance and context of data source(s) and data evolution.
- Computing with data in situ to minimize the impact of extract, transform and load.
- Transforming information into knowledge.
- Keeping the user clearly involved in the analytic loop to not only provide the results from various types of transformations but to also allow the user to guide the transformation process itself.

Transformations and Representations for Visualization

The first two sections described the raw data characteristics and the methods to transform the data to efficient representations. The final step, described in this section, is to develop visual representations of the transformed data that gives the end user an easy-to-analyze visual form.

From data to visual display

The overall goal of creating visual representations is to use cognitive and perceptual principles that increase the communication impact of the results of the data transformation process to enable visual analysis and knowledge synthesis. These techniques need to use visual representations that ease the user's cognitive burden through the creation of effective external cognitive artifacts, work across problems and data at multiple scales, and semi-automatically adapt to the task at hand. Therefore, a clear understanding of the principles of effective visual information depiction is needed.

Incorporating these principles into visual analytics systems allows the creation of appropriate visual representations. The level of abstraction and choice of visual representation are keys to success. The goal is to not only present the deluge of data that the analyst receives but also extract the relevant information from these data in a format that enables reasoning and analysis. Therefore, improved visual representation can be generated that incorporates both advanced techniques for showing complex 3D structures. In addition, techniques are needed for abstracting the representation, focusing the user's attention and providing contextual information for reference. All of these techniques must adapt to the large variety of types and kinds of information to be simultaneously analyzed and scaled across both data size and display size (PDA to wall display). In creating effective analytic environments, visual metaphors are needed for different data representations, including

- raw data;
- data signatures and transformed data;
- metadata information including related data, transformations and algorithms applied to generate the data signatures, as well as data lineage.

To be effective, these visual representations must accommodate the users' perceptual preferences and characteristics (for example, color acuity, form dominance) and their cognitive analysis style, the characteristics of the display device (for example, cell phone versus display wall), and the characteristics of the task they are performing (for example, time frame for decision making, discovery task, analysis task, verification task, situational awareness task). The key issues are centered on developing principles and techniques to enable cognition amplification.[33] Creating useful and appropriate cognitive artifacts enhances both reflective and experiential cognition.[33] The design task must use cognitive principles such as the appropriateness principle the naturalness principle and the matching principle.[34]

Human adapted display of data to enhance analysis – The balance between automated data processing and human reasoning

Each data type (raw data, appropriately transformed data – using techniques from the previous section – and metadata) offers the challenge of determining an effective visual representation. Decision making is the ultimate goal. The decision-making environment must allow visual cognition and analysis in a way that lets the user guide additional data analysis and transformation to complete the task at hand. Over the past 10 years, this has become an active area of research, but many challenges still remain.

There have been some good systems that use data characteristics to determine appropriate visual mappings.[33] These are often based on low-level perceptual characteristic mappings for the classes of data (for example, ordinal, nominal, interval, ratio). Over the past several years, these techniques have begun appearing in commercial products to aid users in understanding their data (for example, ShowMe in Tableau[35]). Several systems match task and data characteristics to appropriate visualizations[36,37] and there is new work in evaluations of their effectiveness.[38] Numerous systems provide abstract, illustrative renderings of data by attempting to harness the power and conciseness of the representations developed by medical and technical illustrators.[39–41] A number of efforts have been made to use design principles for visualization over the past 10 years.[42] All of these approaches have been used on a limited basis and represent only initial steps at solving the problems of creating the most effective visual representation for multi-source, multivariate, multi-modal, incomplete and temporal data.

Purpose-driven visual representation

As mentioned above, a key component in determining effective visual representations is the purpose of the visualization – what is the task the user is performing? Cognitive task analysis is a highly active research area, and many classifications of tasks have been developed. Two key factors in determining the appropriate visual representation are the type of task and time-scale of the task. Discovery, verification, inquiry and situational awareness tasks all have different characteristics that lead to different visual representations. For instance, in situational awareness displays, the representation needs to succinctly convey all of the relevant situational variables at a summary level, while highlighting unusual values/events. In contrast, in a verification or inquiry task, detailed information presented clearly and enabling comparative or correlative analysis is necessary.

The time-scale of the task is equally important. For displays that users interact with for many hours per day for in-depth analysis, complex, rich visual representations can be used. However, in real-time decision-making environments, pre-attentive or slightly longer visual information transfer may be necessary to convey the information quickly enough for effective decision making. In this case, low-level perceptual cueing through simple visual representation such as size, course shape, color and transparency may be the only viable choices. The frequency of system use also factors into the visual representation that is appropriate if complex visual mappings are used.

Data characteristics for visual representations

The previously described data transformations that are adapted to visual display are critical to a visual analytics environment's success. Even with advanced data transformations, many data characteristics still make the visual representation challenging to enable effective visual analysis. For instance, in multi-source data integration and fusion, it is vital that the data transformations enable the fused data to be visually fused and compared – they need to have similar scales, magnitudes of error and standard deviations, and they need to permit linear visual interpretation when mapped to 2D, 3D, and perceptual color spaces. Enabling visual comparison and integration of the resulting data signatures is one key difference between automated data transformations and visual-analytic data transformations. Linearizable transformations for uncertainty, confidence, erroneous and missing[43] data are also needed to enable correct visual interpretation. For instance, in syndromic surveillance, there is uncertainty in syndrome classification from free text, coarseness, errors and missing data in geographical address information, as well as confidence in the values in the data from self-reported illnesses. All of this must be numerically or categorically represented in the transformations and then visually conveyed effectively to the user.

Just as challenging is creating visual representations that enable the user to analyze data across multiple scales, described in a previous article in this volume. Cross-scale reasoning is necessary in many systems that require visual analytic solutions to manage the complexity of the analysis task. Appropriate abstraction and aggregation of data to enable this cross-scale visual reasoning is crucial.

Visual representation solutions

A large toolbox of visual representation techniques can be brought to bear on visual analytic problems with large and challenging data characteristics. Shape and color have been well studied for representing data values. Some less tested, more interesting techniques include the following:

- Transparency – potential for showing temporal data (past/future), data certainty. Poor at showing defined attribute values.
- Texture patterns – potential for showing aggregation, clustering, categorical information, uncertainty with marks.
- Line style variation – heavily used in architecture, technical and medical illustration for showing certainty/uncertainty, known and missing information, and temporal characteristics of data.
- Ghosting – great potential value for showing temporal and certainty information.

The above are standard graphical techniques. The key to visual representations is the integration of graphics design when building visual analytics systems to increase the overall communication impact.

Challenge synopsis

Transforming data into effective visual representations includes the following challenges:

- Classifying when the best visual representation can be automatically chosen.
- Choosing effective visual representations for cross-scale analysis.
- Defining visual representations classes that scale from real-time to in-depth slow analysis.
- Characterizing visual representation for confidence, uncertainty and erroneous data.
- Developing effective visual representations for reasoning about temporal data.

Conclusion

The task of transforming and representing massive amounts of data into comprehensible forms remains a challenge. The magnitude of the transformation and representation problem is increasing because the rate at which data of all types discussed in this article is growing faster than the research effort.

Visual analytics relies on effective, reliable transformation and representation methods to distill raw data into forms from which humans can gain insight. As discussed in this article, there is no single transformation or representation method to uniformly address all data issues. As well as improving and extending transformation and representation methods required for computation, additional investigation is needed to understand the most appropriate data transformation and representation method(s) for specific visual analytics tasks.

Acknowledgements

Our thanks to Professor Richard Fujimoto, Computational Science and Engineering Division, Georgia Institute of Technology for his valuable input on this manuscript. The anonymous reviewers provided excellent comments that improved the quality of this paper.

This work has been supported by the Boeing Company, the Purdue Regional Visualization and Analytics Center™ (PURVAC™) located at the Purdue University in West Lafayette, IN and the Foundations on Data Analysis and Visual Analytics (FODAVA) research initiative located at the Georgia Institute of Technology in Atlanta GA. PURVAC is sponsored by the U.S. Department of Homeland Security (DHS) Science and Technology (S&T) Division. FODAVA is sponsored jointly by the National Science Foundation and the U.S. Department of Homeland Security.

References

1 Thomas, J.J. and Cook, K.A. (eds.) (2005) Data representations and transformations. In: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, Chapter 4. Los Alamitos, CA: IEEE Computer Society Press, pp. 105–136.

2 Robertson, G., Ebert, D., Eick, S., Keim, D. and Joy, K. (2009) Scale and complexity in visual analytics. *Information Visualization* 8(4): 247–253.

3 Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24(7): 498–520.

4 Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R. (1990) Indexing by latent semantic analysis. *Journal of American Society for Information Science* 41: 391–407.

5 Golub, G. and Loan, C. V. (1996) *Matrix Computations*. 3rd edn. London: Johns Hopkins University Press.

6 Howland, P. and Park, H. (2007) Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(8): 995–1006.

7 Kim, H. and Park, H. (2008) Nonnegative matrix factorization based on alternating non-negativity-constrained least squares and the active set method. *SIAM Journal on Matrix Analysis and Applications* 30(2): 713–730.

8 Lee, D.D. and Seung, H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755): 788–791.

9 Choo, J., Bohn, S., Park, H. (2009) Two-stage framework for visualization of clustered high dimensional data. Proceedings of IEEE Symposium on Visual Analytics Science and Technology: 12–13 October, Atlantic City NJ.

10 Scholkopf, B. and Smola, A. (2001) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: Massachusetts Institute of Technology Press.

11 Littlestone, N. (1988) Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning* 2: 285–318.

12 Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining association rules between sets of items in large databases. In: P. Buneman, and S. Jajodia (eds.), Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data; 25–28 May, Washington, DC, NY: Association for Computing Machinery Press, pp. 207–216.

13 Roth, D. (1999) Learning in natural language. In: T.L. Dean (ed.) Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '99); 31 July–6 August, Stockholm, Sweden. San Francisco, CA: Kaufmann Publishers, pp. 898–904.

14 Ganiz, M.C., Lytkin, N.I. and Pottenger, W.M. (2009) Leveraging higher order dependencies between features for text classification. In: ECMLPKDD '09: The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases; 7–11 September, Bled, Slovenia.

15 Ganiz, M.C., Kanitkar, S., Chuah, M.C. and Pottenger, W.M. (2006) Detection of interdomain routing anomalies based on higher-order path analysis. In: C.W. Clifton (ed.) Proceedings of the Sixth International Conference on Data Mining (ICDM '06); 18–22 December, Hong Kong. Los Alamitos, CA: IEEE Computer Society Press, pp. 874–879.

16 Buneman, P., Khanna, S. and Tan, W.C. (2001) Why and where: A characterization of data provenance. In: J. VandenBussche, and V. Vianu (eds.) Proceedings of the 8th International Conference on Database Theory (ICDT 2001); 4–6 January, London, UK. Berlin: Springer, pp. 316–330.

17 Hastie, R., Tibshirani, X. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. New York: Springer.

18 Agresti, A. (2007) *An Introduction to Categorical Data Analysis*. 2nd edn. Hoboken, NJ: Wiley-Interscience.

19 Cowell, R.G., Dawid, A.P., Lauritzen, S.L. and Spiegelhalter, D.J. (1999) *Probabilistic Networks and Expert Systems*. New York: Springer.

20 Cox, T. and Cox, M. (2001) Multidimensional scaling. *Monographs on Statistics and Applied Probability 88*, 2nd edn. Boca Raton, FL: Chapman and Hall/CRC.

21 Munzner, T., Guimbretiere, F., Tasiran, S., Zhang, L. and Zhou, Y. (2003) TreeJuxtaposer: Scalable tree comparison using focus+context with guaranteed visibility. Proceedings of the International Conference on Computer Graphics and Interactive Techniques ACM SIGGRAPH 2003 Papers; 27–31 July, San Diego, CA. New York: Association for Computing Machinery Press, pp. 453–462.

22 Jain, A. (forthcoming). Data Clustering: 50 Years Beyond K-means. Pattern Recognition Letters, Elsevier.

23 Wright, W., Schroh, D., Proulx, P., Skaburskis, A. and Cort, B. (2006) The sandbox for analysis: Concepts and methods. In: R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson (eds.) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 22–27 April, Montreal, Quebec, Canada. New York: Association for Computing Machinery Press, pp. 801–810.

24 Shrinivasan, Y.B. and vanWijk, J.J. (2008) Supporting the analytical reasoning process in information visualization. Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems; 5–10 April, Florence, Italy. New York: Association for Computing Machinery Press, pp. 1237–1246.

25 Yang, X., Asur, S., Parthasarathy, S. and Mehta, S. (2008) A visual-analytic toolkit for dynamic interaction graphs. Proceedings of the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 24–27 August, Las Vegas, NV. New York: Association for Computing Machinery Press, pp. 1016–1024.

26 Schreck, T., Tekusova, T., Kohlhammer, J. and Fellner, D. (2007) Trajectory-based visual analysis of large financial time series data. *ACM SIGKDD Explorations Newsletter* 92: 30–37.

27 Adrienko, G., Adrienko, N. and Bartling, U. (2008) Interactive visual interfaces for evacuation planning. In: P. Bottoni, and S. Levialdi (eds.) Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '90); 28–30 May, Napoli, Italy. New York: Association for Computing Machinery, pp. 472–473.

28 Stumpf, S. *et al.* (2008) Integrating rich user feedback into intelligent user interfaces. In: J. Bradshaw, H. Lieberman, and S. Staab (eds.) Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI 2008); 13–16 January, Gran Canaria, Spain. New York: Association for Computing Machinery Press, pp. 50–59.

29 He, D., Brusilovisky, P., Grady, J., Li, Q. and Ahn, J.-W. (2007) How up-to-date should it be? The value of instant profiling and adaptation in information filtering. In: T.Y. Lin (ed.) Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence; 2–5 November, Silicon Valley, CA. Los Alamitos, CA: IEEE Computer Society Press, pp. 699–705.

30 Wen, Z., Zhou, M. and Aggarwal, V. (2007) Context-aware, adaptive information retrieval for investigative tasks. In: A. Puerta, and T. Lau (eds.) Proceedings of the 12th International Conference on Intelligent User Interfaces; 28–31 January, Honolulu, HI. New York: Association for Computing Machinery Press, pp. 121–131.

31 Wen, Z. and Zhou, M. X. (2008) An optimization-based approach to dynamic data transformation for smart visualization. In: J. Bradshaw, H. Lieberman, and S. Staab (eds.) Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI 2008); 13–16 January, Gran Canaria, Spain. New York: Association for Computing Machinery Press, pp. 70–79.

32 Perry, J. M., Janneck, C. D. and Pottenger, W. M. (2009). Supporting Cognitive Models of Sensemaking in Analytics Software Systems. Rutgers University, Center for Discrete Mathematics & Theoretical Computer Science. DIMACS Research Report 2009-12.

33 Card, S., Mackinlay, J. and Shneiderman, B. (1999) *Readings in Information Visualization: Using Visualization to Think*. San Francisco, CA: Morgan Kaufmann.

34 Norman, D. (1993) *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Reading, MA: Perseus Books.

35 Mackinlay, J., Hanrahan, P. and Stolte, C. (2007) Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics* 13(6): 1137–1144.

36 Roth, S. F. and Mattis, J. (1991) Automating the presentation of information. *Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications; 24th--28 February, Miami Beach, FL*. Los Alamitos, CA: IEEE Computer Society Press, pp. 90–97.

37 Zhou, M. (1999) Visual planning: A practical approach to automated visual presentation. in: T.L. Dean (ed.) Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '99); 31 July–6 August, Stockholm, Sweden. San Francisco, CA: Kaufmann Publishers, pp. 634–641.

38 Bier, E. A., Card, S. K. and Bodnar, J. W. (2008) Entity-based collaboration tools for intelligence analysis. In: D. Ebert, and T. Ertl (eds.) IEEE Symposium on Visual Analytics Science and Technology: VAST '08; 21–23 October, Columbus, OH. Los Alamitos, CA: IEEE Computer Society Press, pp. 99–106.

39 Svakhine, N., Jang, Y., Ebert, D. S. and Gaither, K. (2005) Illustration and photography-inspired visualization of flows and volumes. *IEEE Visualization (VIS '05); 23--28 October, Minneapolis, MN*. Piscataway, NJ: IEEE, pp. 687–694.

40 Svakhine, N., Ebert, D. S. and Andrews, W. M. (2008) Illustration-inspired depth enhanced volumetric medical visualization. *IEEE Transactions on Visualization and Computer Graphics* 15(1): 77–86.

41 Chen, W., Zhang, S., Correia, S. and Ebert, D. S. (2008) Abstractive representation and exploration of hierarchically clustered diffusion tensor fiber tracts. *Computer Graphics Forum* 27(3): 1071–1078.

42 Agrawala, M. *et al.* (2003) Designing effective step-by-step assembly instructions. *ACM Transactions on Graphics TOG* 22(3): 828–837.

43 Kim, H., Golub, G. and Park, H. (2005) Missing value estimation for DNA microarray expression data: Local least squares imputation. *Bioinformatics* 21(2): 187–198.

Disclaimer: The work of authors from Georgia Institute of Technology was supported in part by the NSF/DHS FODAVA-Lead grant CCF-0808863. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.