An Annotation Process for Data Visualization Techniques

Geraldo Franciscani Jr., Rodrygo L. T. Santos, Raphael Ottoni, João Paulo Pesce,
Wagner Meira Jr. and Raquel Melo-Minardi
Department of Computer Science - DCC
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{gfrancis, rodrygo, rapha, jpesce, meira, raquelcm}@dcc.ufmg.br

***Abstract*—As the area of information visualization grows, a massive amount of visualization techniques has been developed. Consequently, the choice of an appropriate visualization has become more complex, usually resulting in unsatisfactory data analysis. Although there exist models and classifications that could guide the choice of a visualization technique, they are mostly generalist and do not present a clear methodology for evaluation and evolution. In contrast, we propose an annotation process for data visualization techniques based on an initial capability-driven collection of terms and concepts that encompasses visual components of both well established as well as modern visualization techniques. To demonstrate the initial collections expressiveness, we present a qualitative analysis of an experiment with specialist users at annotating visualization techniques from the D^3 (Data-Driven Documents) library. Furthermore, to show the completeness of the collection, we automatically assess its coverage of all published papers from six major international information visualization conferences since 1995. Our results attest the expressiveness of the initial collection and its coverage of over 99% of the analysed literature. Finally, we discuss the limitations and alternatives for semi-automatically evolving the annotation process as new visualization techniques are developed and how the spread of this type of methodology could benefit the information visualization community.**

Keywords–Annotation Process; Data Visualization; Ontologies; Taxonomies.

I. INTRODUCTION

There has been an increasing need to extract relevant information from data and make sense of it in different contexts. At the same time, it is becoming increasingly difficult to identify frequent patterns and exploit large databases. Human abilities of visual perception and cognition come into play as the need to extrapolate textual forms and explore the graphic field become a necessity. As the information visualization area grows, a vast number of visualization techniques are developed. Nonetheless, ordinary users are not prepared to decide which visualization is the most appropriate for the required analysis and tend to express data unsatisfactorily. As a result, the development of strategies and tools to help users choose visualization techniques that can effectively help in data analysis and sense making has become crucial.

It is vital to organize the knowledge of visualization methods and capabilities being produced, with a focus on making visualization development easy, more tangible and effective. We are reaching a juncture of information overload where it has become challenging, even for experts, to cope with the many approaches on visualizing data produced by the academic and design communities. With that in mind, the knowledge being produced by information scientists in the creation of concepts of classification models, taxonomies and ontologies is a straightforward approach.

According to the *Oxford English Dictionary*, a *taxonomy* is a classification, especially in relation to its general laws or principles; that department of science, or of a particular science or subject, which consists in or relates to classification, especially the systemic classification of living organisms. An *ontology* is the science of study of being; this is a department of metaphysics that relates to the being or essence of things, or to being in the abstract. Researchers have been using such a collection of concepts and terms in the biology field since at least 13 years ago, when Gene Ontology was proposed and broadly adopted [1].

From our perspective, the information visualization area requires a unified annotation process that allows its community to annotate or associate terms to both traditional visualization techniques as well as novel techniques being developed. We believe the collection of terms needed by the information visualization field should primarily be able to describe visualization methods in terms of two main elements: visual components and capabilities. Examples of visual components are dimensionality, the objects used in the visual composition, the types of displays and pre-attentive attributes. By capabilities, we mean broader features that encompass the quantitative relationships being described and visual patterns being revealed, as well as the analytical, navigation and interaction techniques that could be used.

According to Gilchrist [2], the definitions of the terms taxonomy and ontology have been subverted and overlap significantly. Previous works focused on the specification of taxonomies, models and ontologies to describe and study the relationships between terms and visualization techniques [3]–[10]. Most importantly, there were attempts to use such classifications and models to generate recommendation systems and to evaluate techniques [11] [12]. Although those works have some important implications in helping users to express data in a more satisfactory way, they do not represent a consensus between specialists and do not address a clear methodology for progressive evaluation and evolvement, regarding the emergence of new techniques and concepts.

In the present work, we describe the methodology used to propose an annotation process for data visualization techniques based on a collection of terms and concepts that covers visual components of visualization techniques and their capabilities. Next, we select a diverse set of visualizations to be annotated with the proposed collection of terms. Note that, here, we import the term annotation from the biology field where it

 ISBN: 978-1-61208-358-2

means the association of terms of the controlled vocabulary with biological objects. We also propose an FP-tree-based [13] algorithm to organize the set of visualizations in a tree where internal nodes are the collection terms and leaves are the visualizations themselves. The tree is a type of visual index that helped us to evaluate both the collection of terms and the selected set of visualization characteristics. We characterize this tree and show how it provides a macro view of the visualization capabilities. Furthermore, we also automatically assess the coverage of our proposed collection of terms in all published papers from six major international information visualization conferences since 1995. Our results attest the expressiveness of the proposed collection and its coverage of over 99% of the analysed literature. In addition, we discuss alternatives for semi-automatically evolving the annotation process as new visualization techniques are developed and finally, how the spread of this type of methodology could benefit the information visualization community.

The remainder of this paper is organized as follows. Section 2 reviews some works related to the development of models and classifications in the information visualization and visual analytics fields. Section 3 describes all the methods, including the proposed annotation process and, in particular, the use of this process, used to build a tree of visualization techniques and their related terms. Section 3 also describes the algorithms we built and used with that purpose as well as the strategy to automatically assess the presence of the terms in the literature. Section 4 presents an evaluative study of the proposed process and discussions about the adopted methods. Section 5 presents the evaluation results. Finally, Section 6 presents our concluding remarks and future directions.

II. Related Work

Many studies focused on the definition of a consistent ontology / taxonomy to categorize visualizations. Our goal is to identify areas already covered by the ontologies / taxonomies existing in the literature and find related examples that serve as basis for our annotation process. Voigt and Polowinski [14] systematically reviewed existing models and classifications, comparing the strengths and weaknesses of each, as well as establishing relationships among them. As a result, the authors specified an initial unified visualization ontology for classification and synthesis of graphical representations. Although it is complete and comprehensive, the authors do not present a methodology for evaluation and evolvement of the concepts presented.

Duke, Brodlie and Duce [15] built an initial skeleton for a vocabulary that would identify the communication between user and system. Concepts and relationships were considered in more restricted areas such as data, tasks and visual representations. In their study, the authors describe how the relationships between published studies may contribute to the construction of this unified ontology and presented, as a major challenge, the consensus among researchers in this area. Although it was an important attempt to organize and categorize existing knowledge, it presents an early version of the vocabulary that would require more specificity to classify a large set of techniques.

Shu, Avis and Rana [9] presented the design of an ontology focused on providing semantics to aid the discovery of visualization services based on the initial concept proposed by Duke, Brodlie and Duce [15]. Their study defined classes mostly for modeling data and visualizations techniques. However, the presented class names were unreadable for users and some concepts were not addressed, such as tasks and interactions.

Shneiderman [16] proposed the Task by Data Type Taxonomy (*TTT*) for information visualizations, dividing the visualization techniques into seven data types (one-, two-, and three-dimensional data, temporal and multi-dimensional data, tree and network data) and seven tasks (overview, zoom, filter, details-on-demand, relate, history, and extracts). The data types characterize the task-domain information objects and are organized by the problems the users are trying to solve. The seven tasks are at a high level of abstraction and represent user interaction with the visualization or data. In 2012, Shneiderman and Jeffrey [4] proposed an update for *TTT* by presenting a taxonomy of interactive dynamics to help users in evaluating and creating visual analysis tools. The taxonomy consists of 12 task types grouped into three high-level categories (1) data and view specification (visualize, filter, sort, and derive); (2) view manipulation (select, navigate, coordinate, and organize); and (3) analysis process and provenance (record, annotate, share, and guide). Although *TTT* was an interesting step towards categorizing and organizing existing visualizations, from the perspective of visualization annotation, it is still too generalist and could benefit from the addition of more detailed and discriminative terms.

Chi [5] presented another taxonomy based on what they called the Data State Reference model. This model divides each technique into four data stages (value, analytical abstraction, visualization abstraction and view) and three types of data transformation operators (data transformation, visualization transformation and visual mapping transformation). Within each data stage, there are four types of operators that do not change the underlying data structures, the within stage operators (within value, within analytical abstraction, within visualization abstraction and within view). Data transformation operators are used to transform data from one stage to another, and within stage operators are used to transform data without changing the underlying data structure. The contribution of this model is in the sense that the authors classified each visualization technique by not only its data type but also its processing operating steps, which helps in understanding the operating steps for each classified visualization technique and in defining sequential ordering of operations and their dependencies. However, this model is limited in comparison to our proposal regarding visualization annotation process in the sense that it does not take into account important factors about the expressive power of visualization techniques in terms of what quantitative relationships they are able to represent, what type of data they can present and what type of visual patterns they can evidence. Additionally, this model does not consider visual objects and pre-attentive attributes involved in the representations.

A different taxonomy-based approach is to focus on the visualization algorithm instead of the data to be visualized. Tory and Möller [6] proposed a model divided into four categories: object of study, data, design model and user model. This model does not attempt to consider the data-oriented approach, instead emphasizing a more flexible system that

 ISBN: 978-1-61208-358-2

highlights the users' conceptual model of the visualization.

Fujishiro, Furuhata, Ichikawa and Takeshima [11] presented a semi-automatic approach for the development of data visualization applications. The authors proposed the GADGET/IV system, based on a goal-oriented taxonomy. This taxonomy has been constructed by combining the Wehrend Matrix [17] with the concepts introduced in TTT [16]. Moreover, this system was an extension to the GADGET (Goal-oriented Application Design Guidance for modular visualization EnvironmenTs) system [18], which used only the above matrix as a reference to aid the development of data visualization applications. This research presented an interesting perspective, although the use of the system was not evaluated.

Pfitzner, Hobbs and Powers [8] built a taxonomy-based framework that encompasses several aspects in information visualization: data, tasks, interactions, context and human capacities of cognition. Although the study seems promising and complete, the usefulness of the taxonomy created was not evaluated and it lacks a clear methodology for evolving the taxonomy with the area.

Gilson, Silva, Grant and Chen [19] proposed an ontology as part of a tool that automatically generates visualizations from web pages in specific areas without prior knowledge of the content of these pages. Although the proposed ontology presents properties of graphical representations and visual objects, some important topics such interactions, tasks to be performed on data and user goals were not considered.

Amar, Eagan and Stasko [20] presented a set of ten low-level analysis tasks (retrieve value, filter, compute derived value, find extremum, sort, determine range, characterize distribution, find anomalies, cluster and correlate). According to the authors, these tasks capture people's activities while employing information visualization techniques to understand data. These tasks were obtained using an affinity diagramming approach from 200 sample questions from students about how they would analyze five different datasets from different domains with information visualization tools. Despite being very interesting, this taxonomy focuses only on analytical tasks and not on visualization techniques, which is what this work focuses on.

Zhou and Feiner [10] developed a visual task taxonomy that extends the one proposed by Wehrend and Lewis [17]; additional tasks were defined, parameterized, and grouped in three dimensions (organization, signaling and transformation). These dimensions were composed by types and subtypes where elemental tasks were defined (for instance, associate, cluster, locate, categorize, cluster, distinguish, among others). Morse, Lewis and Olsen [12] showed that this type of taxonomy can be used in the evaluation of visualization techniques. In this research, a methodology is developed to create a set of taxonomy-based tasks for evaluating visualization techniques for information retrieval. According to this research, the taxonomies are very useful for addressing the complexity of the visual tasks.

From our point of view, we will consider all related works to compose our annotation process, as they are an important inheritance in the area. However, in this work, we will not consider data preparation or transformation tasks. We are mainly interested in visual components, which are not considered in most previous works, and the capabilities of the visualization techniques, which have been considered with different perspectives. We tried to conserve important terms regarding data type, but the majority of the terms we kept concerns important analytical interaction techniques that can be applied to the visualizations and consequently can give them important capabilities.

III. Annotation Process

The proposed annotation process consists in a definition of a collection of terms and concepts related to a set of data visualizations techniques to be annotated. To this end, we conduct an experiment with experts who defined an initial selection of terms and concepts in existing literature. We also propose an initial set of data visualization techniques that will serve as a source for the study and may also evolve with the area. Finally, we present the annotation process itself as an association of the techniques with the terms and concepts.

A. Initial Collection of Terms and Concepts

We had two main objectives in proposing the initial collection of terms and concepts: terms should describe visualization techniques concerning their visual components; and terms should encompass the quantitative relationships being described and visual patterns being revealed, as well as the analytical, navigation and interaction techniques that could be used with the visualization.

First, we list all terms and concepts found in the existing models and classifications in the literature presented in Section II. Then, we enrich this set with other terms manually selected from references qualified in our research field [21]–[33]. The first reference used terms we considered useful for the two aforementioned objectives we defined for the annotation process and the second is classical in terms of visual objects. As a result, we obtained a set composed by 101 terms.

In order to adjust this initial set with the proposed objectives, we conduct an experiment with three experts (one professor and two MSc students in Information Visualization) and three data visualization research assistants. Each one evaluated the relevance (yes or no) of each term according to the two previously mentioned objectives. After that, we considered the terms that had 100 % positive reviews (63). The terms with one or more negative evaluations were discussed among the group and evaluated again. Terms with an agreement higher than 80% were considered (11), and the remaining disregarded (27). At the end of the experiment, we obtained a more appropriate initial collection composed by 74 terms.

We present the initial collection below. The following terms present visual objects and attributes that are intuitive and self-explanatory. Thus, we only cite them: *Bars, Boxes, Cells, Circle Section, Lines, Points, Ring Sector, Shape, Trails, Motion, Direction, 2D Spatial Position Representing Quantities, Spatial Grouping Position Representing Categories, Blur, Color Variation, Curvature, Enclosure, Orientation Variation, Shape Variation, Size Variation, Texture Variation, Value Variation, 1D (Dimensional), 2D (Dimensional), 3D (Dimensional), Multidimensional.*

Next, we list and explain the remaining terms:

Correlation: How variables relate to and affect one another.

 ISBN: 978-1-61208-358-2

Deviation: How one or more sets of values deviate from a reference set of values, which can be a target, a forecast, same point in the past, immediately prior period, standard or norm.

Distribution: Examining sets of quantitative values to see how the values are distributed from the lowest to highest or to compare and contrast how multiple sets of values are distributed.

Multivariate: The purpose of multivariate analysis is to identify similarities and differences among items, each characterized by a common set of variables.

Part-to-whole: Used when trying to make sense of a total amount (whole), aggregating them by the parts to see how much each part adds to the whole.

Ranking: Items ranked by value.

Time series: One or a set of time-dependent attributes.

Alternating differences: Differences from one value to the next begin small then shift to large and finally shift back again to small.

Center: Estimation of the middle of the set of values.

Co-variation: When two sets of values relate to one another so that changes in one are reflected by changes in the other, either immediately or later, this is called co-variation.

Cycles: Patterns that repeat at regular intervals, such as daily, weekly, monthly, quarterly, yearly, or seasonally.

Exceptions: Values that fall outside the norm.

Gaps: Empty regions where we would expect to find values.

Increasingly different: Differences from one value to the next decrease.

Non-uniformly different: Differences from one value to the next vary significantly.

Rate of change: The percentage difference between one value and the next.

Shape: Shows where the values are located. If it is a curve, for instance, is it curved or flat? If curved, upward or downward? If curved upward, single or multiple peaked? If single peaked, symmetrical or skewed? Concentrations? Gaps?

Spread: A measure of dispersion, that is, how spread out the values are.

Trend: The overall tendency of a series of values to increase, decrease or remain relatively stable during a particular period of time.

Uniform: All values are roughly the same.

Uniformly different: Differences from one value to the next decrease by roughly the same amount.

Variability: The average degree of change from one point in time to the next throughout a particular span of time.

Directed (Analytical Navigation): Begins with a specific question (perhaps a particular pattern), and then produces the answer.

Exploratory (Analytical Navigation): Begins by simply looking at the data without predetermining what might be found. Then, when something that seems interesting is noticed and questioned, we proceed in a directed fashion to find an answer to that question.

Hierarchical (Analytical Navigation): To navigate through information from a high level view into progressively lower levels along a defined hierarchical structure and back up again.

Accessing details on demand: When details are called up instantly when needed but kept out of the way before they are needed and after they have been read. Select a group or item and obtain details when needed.

Adding variables: Adding one or more attributes.

Aggregating: When we aggregate or disaggregate information, we are not changing the amount of information but rather the level of detail at which it is viewed. We aggregate data to view it at a high level of summarization or generalization; we disaggregate to view it at a lower level of detail.

Annotating: To document objects of the display, adding notes to them.

Bookmarking: To allow users to save automatically particular views, including its filters, sorts, and other features, so they can easily return to them later.

Brushing and linking: To highlight the same subset of data in multiple graphs at the same time.

Comparing: Encompasses comparing (looking for similarities) and contrasting (looking for differences).

Drilling: Involves moving down levels of summarization (and also back up) along a defined hierarchical path.

Filtering: The act of reducing the data we are viewing to a subset of what is currently there.

Focus and context together: When we are focusing on details, the whole does not need to be visible in high resolution, but we need to see where the details are focusing or reside within the bigger picture and how they relate to it.

Highlighting: To cause particular data to stand out without causing all other data to go away.

Re-expressing: When we change the way we delineate quantitative values that we are examining (e.g.: changes of units of measure).

Re-scaling: Changes the scale: linear, quadratic, or logarithmic.

Re-visualizing: Changing the visual representation in some fundamental way, such as switching from one type of graph to another.

Sorting: Sorting from low to high or high to low.

Zooming and panning: When we enlarge the portion of the display that we wish to see more closely.

Clustering items by similarity: Clustering is the process of segmenting data into groups whose items share similar features.

 ISBN: 978-1-61208-358-2

Comparison of individual and cumulative values: Useful when we assess how well things are going by comparing actual values to targets.

Multiple concurrent views and brushing: The visualization of a single dataset from different perspectives concurrently using multiple graphs.

Overlapped time scales: We can strengthen our ability to detect and compare cyclical patterns stretching across multiple cycles in a line graph by displaying each cycle as a separate line and overlapping time scales.

Ranking items by similarity: To order items according to their relative similarity to enhance visual analysis.

Reference lines and regions: Objects used to give context to the analysis making comparisons easy. Reference lines usually represent expected values as well as averages or means.

Trellises and cross-tabs / Small multiples: When we divide the data set we wish to examine into multiple graphs, either because we can't display everything in a single graph without resorting to a 3-D display, which would be difficult to decipher, or because placing all the information in a single graph would make it too cluttered to read. By splitting the data into multiple graphs that appear on the screen at the same time in close proximity to one another, we can examine the data in any one graph more easily, and we can compare values and patterns among graphs with relative ease.

B. Visualization Techniques

The visualization techniques used in the study were collected in December of 2012 from D^3's (*Data-Driven Documents*) [34] web site [35]. This dataset was used due to the extensive and varied set of visualizations techniques made available by D^3's collaborators. We removed examples that were not true visualization techniques and represented only examples of how to use the library. A total of 53 visualization techniques remained.

C. Association Process

As noted previously, the annotation term was borrowed from biology and means to associate terms of an ontology with objects of interest. In our case, the ontology is represented by the initial collection of terms and concepts and the objects of interest by the visualization techniques. The annotation process consisted of using a web form to associate a set of terms with each visualization. It was performed by the same team of experts and research assistants and annotations with more than 80% of agreement were considered. We decided to associate with each visualization not only terms that are readily implemented in the visualizations but also every term that could be easily incorporated into the implementation because our purpose is to annotate visualizations according to their capabilities rather than their implementation. Our goal is to open the system to the scientific community to integrate other researchers' opinions about the current annotations in a way that the process will be more robust and reliable, analogous to what happened in biology.

IV. Evaluation Strategy

In this section, we describe our strategies to evaluate the annotation process and its components. Firstly, to evaluate the expressiveness of the initial collection of terms and the performed annotation procedure, we produce a visual index represented in a tree structure. The nodes are the terms, and the leaves are the techniques. Then, to evaluate the completeness of the proposed collection of terms and concepts, we present a methodology for automatically assessing the terms coverage of all published papers from six major international information visualization conferences since 1995.

A. Expressiveness evaluation

To evaluate the expressiveness of the proposed collection of terms and the performed annotation procedure, we produce a visual index in a tree structure. The tree we produced was based on the classical FP-tree which is commonly used to find frequent patterns [36] and to cluster objects [37] in a parameter-independent way. The FP-tree is an appropriate data structure for representing our data because we would like to build a visual index of visualizations and terms capable of grouping similar visualizations in terms of similar visual components and capabilities (the main objectives of our collection of terms and concepts). Additionally, we would like to distinguish popular (and non-discriminative) terms from specific (and discriminative) ones.

We use a modification of the original FP-tree data structure implemented by Pires et al. [37]. Due to space limitations, we will not explain the FP-tree construction algorithm, which can be found with examples in [36]. Each transaction in the database is represented as a path in the tree, where each node is an attribute and the attributes are organized in non-increasing frequency order from root to leaves. The path length (i.e., the number of attributes per transaction) may vary. The attributes are then sorted by their frequency in the database and inserted so that transactions with attributes in common share a path in the tree. Consequently, globally common attributes are at the highest levels and less frequent attributes are at lower levels. The generated tree structure is shown in Figure 2.

B. Completeness evaluation: automatic assessment of the literature coverage

An important drawback of any proposed collection of concepts is the difficulty to assess its completeness. We automatically assess the terms coverage of all published papers from six major international information visualization conferences: IEEE Symposium on Information Visualization (INFOVIS); IEEE Conference on Visual Analytics Science and Technology; EuroVis / Joint Eurographics - IEEE TCVG Symposium on Visualization; International Conference on Information Visualization; Asia Pacific Symposium on information visualization; and Computer Graphics, Imaging and Vision. We download all available papers since 1995, totaling 5,061 publications. To normalize the comparison between the terms and the full-text extracted from each paper, we preprocess all text content by applying standard text processing techniques, such as punctuation removal, stop-words removal, lemmatization and stemming [38]. Finally, in Figure 1, we present terms' coverage of papers in which bars represent

 ISBN: 978-1-61208-358-2

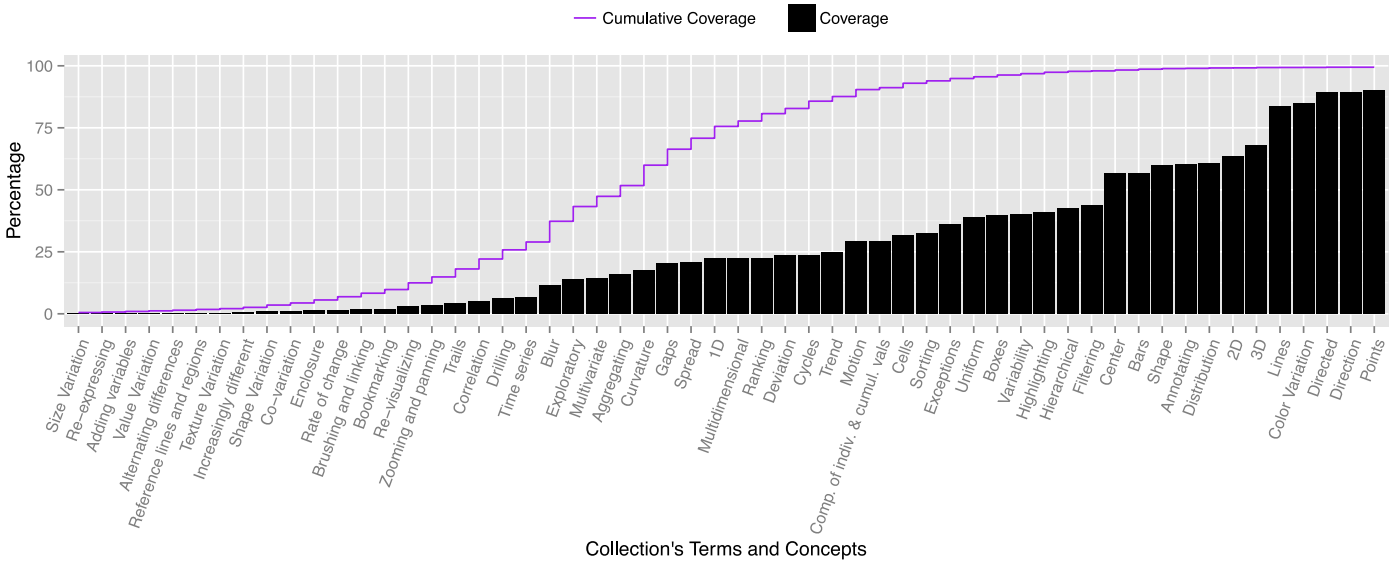

Figure 1. Terms' coverage of papers: bars represent individual terms coverage and line represents the cumulative coverage of papers (percentage) from the current term and all previous ones.

individual terms coverage (the percentage of papers in which the term appears) and line represents the cumulative coverage of papers (percentage) from the current term and all previous ones. Terms that appear in five or less papers (18) were not exposed for presentational reasons. We demonstrate that the suggested collection covered about 99% of the papers, in other words, that 99% of the papers mention at least one of the collection terms. As the most frequent terms can be very general words, we also considered the 75% of the least frequent terms. In this case, the collection still covers 94% of the papers.

V. Evaluation Results

In this section, we present some qualitative results obtained with the proposed collection of terms and concepts in the annotation of a set of visualizations as well as some quantitative results from the automatic assessment of literature coverage.

A. Use of the Annotation Process and Expressive Power of Visualization Techniques

From the 74 terms of the complete collection, 68 were used at least once to describe a visualization. The average frequency of use of a term was 27.22, the minimum was 0 and the maximum was 53, which is the number of visualizations. Hence 5 terms (*Accessing details on demand*, *Annotating*, *Bookmarking*, *Comparing* and *Filtering*) were used to describe all the techniques, which is a result of our strategy of associating each technique with every visualization capable of implementing it, even when the technique was not actually implemented. For instance, the D^3 *Line chart* has no implementation of *Details on demand*, but this analytical interaction technique could be easily implemented in that technique. There were 6 terms with no association as for instance *Texture Variation*. This lack of associations for such a small number of terms does not lower the strength of the proposed collection as the terms were all pre-attentive attributes or visual objects possibly meaning that the visualization set is not too diverse.

Figure 2 depicts the obtained annotation tree, which contains circles that represent the terms of the initial collection and squares representing each annotated visualization technique. The size and color of the squares encode, redundantly, the distance from root from dark blue (high) to light blue (low) on a continuous scale, specified by the number next to their names. Leaves that are farther from the root have more terms assigned to them and the number of terms assigned to a visualization is proportional to its *expressive power*.

On average, 27 out of 68 ($\sim 39\%$) terms are used to annotate each technique. Approximately $\sim 25\%$ of the visualizations have 19 or fewer associated terms, $\sim 50\%$ have 25 terms or fewer, 75% have 31 terms or fewer, and $\sim 90\%$ have 40 terms or fewer. Only 5 techniques are associated with more than 40 terms. We regard these 5 visualizations as special techniques concerning their high expressive power and ubiquity. These 5 techniques are all bar charts or a combination of other representations with a bar chart. The *Grouped bar chart* [39] for instance is an example of high expressive power, represented by 43 terms. At the other extreme and very close to the root of the tree, we have a *Voronoi Diagram* [40] plotted in the US map, dividing the space into a number of regions of points closer to their seed than to other seed (seeds are the US airports in 2008). Although it is a beautiful and informative visualization, it is very specific in terms of applicability.

In conclusion, in our annotation tree, a longer path between a technique and the root indicates a higher expressive power and greater potential ubiquity of that technique. The ubiquity of bar charts is well known, which in some sense demonstrates the correctness and usefulness of our methodology in analyzing this phenomenon.

 ISBN: 978-1-61208-358-2

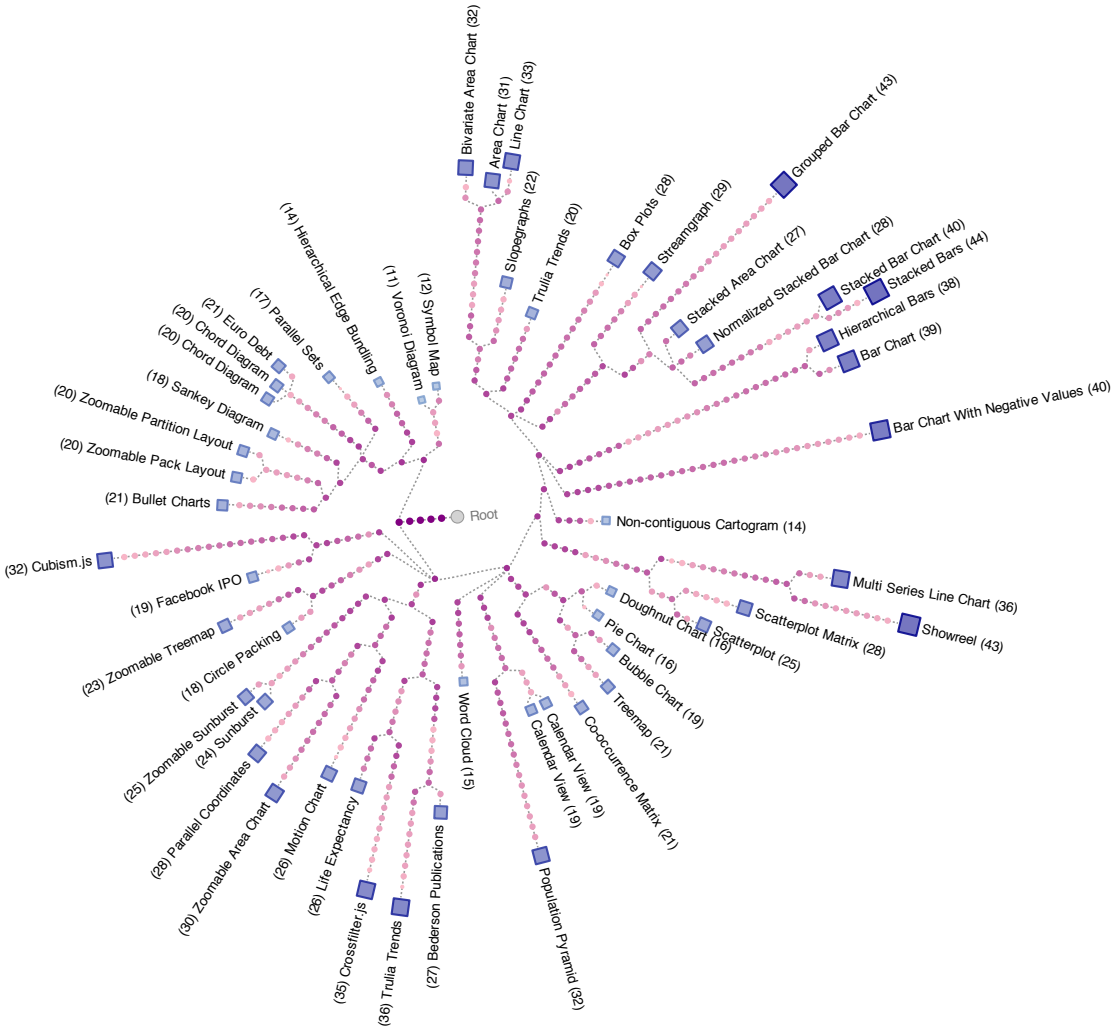

Figure 2. Modified FP-tree for annotating visualization techniques. The circles (internal nodes) represent the terms and the squares (leaves), visualization techniques. The circle colors encode the ratio *DNF* / *TNF* from purple (high) to pink (low) on a discrete scale.

B. Tree Characterization and Discriminative Power of Terms

We used the following metrics to characterize the tree and evaluate the terms of the proposed annotation process: *Dataset Node Frequency (DNF)* is the frequency of the term in the annotation of techniques in the whole dataset; *Tree Node Frequency (TNF)* is the frequency of the node representing the term in the tree, which is lower than or equal to the Dataset Node Frequency due to the compactness of FP-trees and *Mean Distance From the Root (MDR)* is the mean distance of the nodes representing the terms in the tree from the root. All the metrics have the ability to distinguish terms that are very popular in the dataset from more discriminative ones. For instance, the five top nodes of the tree (*Accessing details on demand*, *Annotation*, *Bookmarking*, *Comparing* and *Filtering*) were previously mentioned to describe every single visualization in the dataset. They are not discriminative in that they can be used everywhere and represent interesting and ubiquitous analytical interaction techniques. On the contrary, less frequent terms commonly appear far from the root and tend to be more discriminative. For instance, the term *Rate of change*, which is the percentage difference between one value to the next, presents a *MDR* of 27 and is set only for three techniques: *Line chart*, *Multi Series Line Chart* and *Showreel*, which can show the rate of change when using a logarithmic scale. The same happens for the analytical technique *Comparison of Individual and Cumulative Values* and for the visual patterns *Uniformly Different*, *Non-uniformly Different*, *Increasingly Different* and *Alternating Differences*, which we found very particular of bar charts. The term *Color Variation* is not so frequent in the dataset (60%) but is the most frequent node in the tree, appearing 18 times in various branches because it is the most used pre-attentive attribute in visualization techniques in general.

In Figure 3, we present a distribution of the values for each metric, which are all skewed. Both *DNF* and *TNF* are skewed to the left. The *DNF* has a mean of 11 and a *TNF* of 5. 95% of the terms have frequencies below 52 in the dataset, whereas 95% of the terms are presented fewer than 15 times in the tree. The compression of the tree is apparent here. *Distance from root* is skewed to the right, as the majority of the terms are far from the root, with a mean of 20.

We analyzed the tree under the perspectives of the different

 ISBN: 978-1-61208-358-2

metrics on a continuous scale ranging from higher values to lower values (results not presented due to space limitations). At a first glance, it was difficult to extract interesting patterns from the tree visualizations due to their complexity. The tree visualizations only revealed a color pattern that goes from the root to the leaves, except for some extreme cases, such as the five top nodes that have a very high frequency in the dataset.

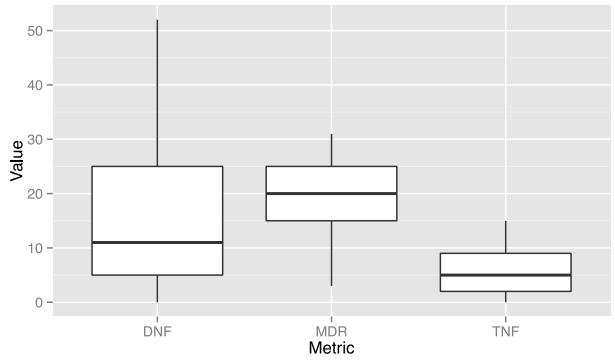

Figure 3. Distribution of the metrics: Dataset Node Frequency (*DNF*), Tree Node Frequency (*TNF*) and Mean Distance From the Root (*MDR*).

An interesting analysis came up when we colored the tree by the ratio between *DNF* and *TNF*, and the result is presented in Figure 2. When the ratio was presented on a continuous scale, its distribution was very skewed and was not easy to spot a pattern. We then used a non-uniform discretization (cuts are presented in Figure 4). The dark purple group (ratio $\geq$ 53) has already been discussed and comprises the five terms that apply to all annotated visualizations. *Exceptions*, *Directed*, *Highlighting*, *Aggregating* and *Trend* are terms presented in light purple ($4 \leq$ ratio $<$ 53), which represents highly discriminative items.

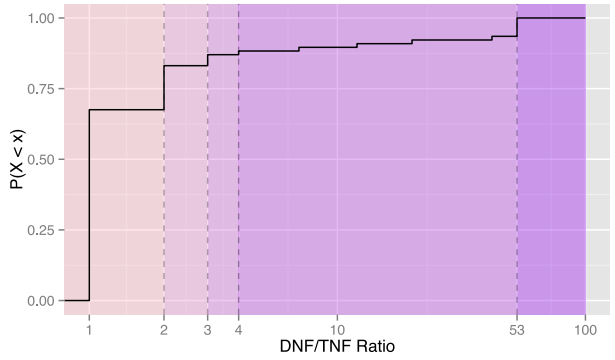

Figure 4. Discretization scheme for the *DNF* / *TNF* Ratio color scale (pink to purple). Note that x-axis is log scaled.

A broad but easily applicable characteristic of visualizations in general is how straightforward they are in communicating the underlying data and producing the desired insights; we call this characteristic *Directed* (ratio 20). The *Bubble Chart* is a good case of a visualization that does not share this characteristic, as it evolves and answers multiple questions along its dynamic life-cycle. These four attributes are comprehensive enough that they are not usually related to a single visualization, but are instead related to a large group. The darker shade of pink ($3 \leq$ ratio $<$ 4) is composed of terms that still have a large discriminative power, but already show some sort of specialization capability. *Size variation* (ratio 2.13) is a good example of this group: it is still discriminative enough to put the *Line chart* and *Bar chart* into separate groups but also specializes the whole group of bar charts (*Stacked Bar Chart*, *Hierarchical Bar Chart*), separating it from the *Streamgraph*, a "cousin" visualization that shares many terms. Terms that fall in the pink group (ratio $\leq$ 2), the largest one, do not have a strong discriminative bias to be close to the root of the tree and are sometimes very specific, being applied to a single technique. The attribute *Sorting* (ratio 1.93) is a relevant example from this group, as roughly half of the visualizations implement or could implement this functionality, but it still discriminates the *Treemap* from the *Doughnut* and *Pie Chart*. A visual object term, such as *Cells* (ratio 1.29), or a display, like *Bar graphs* (ratio 1.2), denotes high specialization.

VI. Conclusion and Future Work

We propose an embryonic version of an annotation process based on an initial collection of terms and concepts extracted from the existing literature that encompasses the visual components and capabilities of visualizations. We select a set of diverse visualizations from the D^3 gallery and annotate them with the proposed terms and concepts. We propose a visual index in form of an annotation tree that helped us to visualize the whole set of techniques and the terms associated to each of them. We characterize the proposed tree, more specifically the terms and the visualizations, concerning three metrics and were able to identify interesting patterns: the discriminative power of terms in relation to the visualizations being described and the expressive power for the visualization techniques. Qualitatively, our results demonstrate the utility of the proposed annotation process in describing visualizations as well as in understanding their capabilities and applicability. In the future, we intend to study how the proposed annotation tree can be used in automatic recommendation tasks to help users to select visualizations for specific problems and to represent data in a satisfactory way. Furthermore, to show quantitatively the completeness of the initial collection of terms and concepts, we automatically assess its coverage across all published papers from six major international information visualization conferences since 1995. Our results attest the expressiveness of the proposed collection and its coverage of over 99% of the published literature.

Finally, we acknowledge our challenge in achieving a consensus from most users of the area and our limitations concerning evaluation and evolution of the annotation process and its components. For that, we developed a platform, *CrowdVIS*, based on crowdsourcing [41]. The main goal of this platform is to use the annotation process's methodology and dynamically evolve the proposed collection of concepts and data visualization techniques, as well as their annotations [42]. Moreover, it should allow users to continuously evaluate each term and technique and to add new ones [43]. A prototype of the system is available at www.crowdvis.dcc.ufmg.br. We believe that the participation of the information visualization community, by annotating the existing visualizations in a similar way and including new visualizations in a public repository will represent a valuable contribution to future studies that could arise from ours. We intend to keep the dataset and annotations open. Certainly, this annotation process, the initial collection of terms and concepts, the annotation procedure and the dataset

 ISBN: 978-1-61208-358-2

could evolve significantly with community involvement and become intrinsic to the field in the future.

REFERENCES

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," Nature Genetics, vol. 25, no. 1, May 2000, pp. 25–29.

[2] A. Gilchrist, "Thesauri, taxonomies and ontologies _ an etymological note," Journal of Documentation, vol. 59, no. 1, 2002, pp. 7–18.

[3] B. Shneiderman, Designing the User Interface: Strategies for Effective Human-computer Interaction. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1986.

[4] J. Heer and B. Shneiderman, "Interactive dynamics for visual analysis," Queue, vol. 10, no. 2, Feb. 2012, pp. 30:30–30:55.

[5] E. H. Chi, "A taxonomy of visualization techniques using the data state reference model," in Proceedings of the IEEE Symposium on Information Vizualization 2000, ser. INFOVIS '00. Washington, DC, USA: IEEE Computer Society, 2000, pp. 69–69.

[6] M. Tory and T. Möller, "Rethinking visualization: A high-level taxonomy," in Proceedings of the IEEE Symposium on Information Visualization, ser. INFOVIS '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 151–158.

[7] G. Ellis and A. Dix, "A taxonomy of clutter reduction for information visualisation," IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 6, Nov. 2007, pp. 1216–1223.

[8] D. Pfitzner, V. Hobbs, and D. Powers, "A unified taxonomic framework for information visualization," in Proceedings of the Asia-Pacific Symposium on Information Visualisation - Volume 24, ser. APVis '03. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2003, pp. 57–66.

[9] G. Shu, N. J. Avis, and O. F. Rana, "Bringing semantics to visualization services," Adv. Eng. Softw., vol. 39, no. 6, 2008, pp. 514–520.

[10] M. X. Zhou and S. Feiner, "Visual task characterization for automated visual discourse synthesis." in CHI, M. E. Atwood, C.-M. Karat, A. M. Lund, J. Coutaz, and J. Karat, Eds. ACM, 1998, pp. 392–399.

[11] I. Fujishiro, R. Furuhata, Y. Ichikawa, and Y. Takeshima, "Gadget/iv: A taxonomic approach to semi-automatic design of information visualization applications using modular visualization environment." in INFOVIS, J. D. Mackinlay, S. F. Roth, and D. A. Keim, Eds. IEEE Computer Society, 2000, pp. 77–83.

[12] E. Morse, M. Lewis, and K. A. Olsen, "Evaluating visualizations: Using a taxonomic guide," Int. J. Hum.-Comput. Stud., vol. 53, no. 5, Nov. 2000, pp. 637–662.

[13] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," Data Mining and Knowledge Discovery, vol. 8, no. 1, Jan. 2004, pp. 53–87.

[14] M. Voigt and J. Polowinski, Towards a Unifying Visualization Ontology, ser. Technische Berichte. Techn.Univ., Fakultät Informatik, 2011.

[15] D. J. Duke, K. W. Brodlie, and D. A. Duce, "Building an ontology of visualization." in IEEE Visualization. IEEE Computer Society, 2004, p. 7.

[16] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in Proceedings of the 1996 IEEE Symposium on Visual Languages, ser. VL '96. Washington, DC, USA: IEEE Computer Society, 1996, pp. 336–336.

[17] S. Wehrend and C. Lewis, "A problem-oriented classification of visualization techniques," in Proceedings of the 1st Conference on Visualization '90, ser. VIS '90. Los Alamitos, CA, USA: IEEE Computer Society Press, 1990, pp. 139–143.

[18] I. Fujishiro, Y. Takeshima, Y. Ichikawa, and K. Nakamura, "Gadget: Goal-oriented application design guidance for modular visualization environments," in Proceedings of the 8th Conference on Visualization '97, ser. VIS '97. Los Alamitos, CA, USA: IEEE Computer Society Press, 1997, pp. 245–252.

[19] O. Gilson, N. Silva, P. W. Grant, and M. Chen, "From web data to visualization via ontology mapping," in Proceedings of the 10th Joint Eurographics / IEEE - VGTC Conference on Visualization, ser. EuroVis'08. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2008, pp. 959–966.

[20] R. A. Amar, J. Eagan, and J. T. Stasko, "Low-level components of analytic activity in information visualization." in INFOVIS, J. T. Stasko and M. O. Ward, Eds. IEEE Computer Society, 2005, p. 15.

[21] E. R. Tufte, Envisioning Information. Cheshire, CT: Graphics Press, 1990.

[22] W. Cleveland, Visualizing data. AT&T Bell Laboratories, 1993.

[23] W. S. Cleveland, The elements of graphing data. Murray Hill, N.J.: AT&T Bell Laboratories, 1994.

[24] E. R. Tufte, Visual Explanations: Images and Quantities, Evidence and Narrative. Cheshire, CT: Graphics Press, 1997.

[25] S. Few, Show me the numbers : designing tables and graphs to enlighten. Oakland, Calif.: Analytics Press, 2012.

[26] E. R. Tufte, Beautiful Evidence. Graphics Press, 2006.

[27] R. Spence, Information Visualization: Design for Interaction (2Nd Edition). Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2007.

[28] C. Ware, Visual Thinking: For Design. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

[29] S. Few, Now You See It: Simple Visualization Techniques for Quantitative Analysis, 1st ed. USA: Analytics Press, 2009.

[30] J. Bertin, Semiology of graphics: diagrams networks and maps. Esri Press, 2010.

[31] J. Steele and N. Iliinsky, Beautiful Visualization: Looking at Data Through the Eyes of Experts, 1st ed. O'Reilly Media, Inc., 2010.

[32] D. Wong, The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures. W W Norton & Company Incorporated, 2010.

[33] C. Ware, Information Visualization, Third Edition: Perception for Design, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012.

[34] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 12, 2011, pp. 2301–2309.

[35] M. Bostock. Data-driven documents: Gallery. http://bit.ly/18kMazA. Accessed April, 2014. (2012)

[36] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '00. New York, NY, USA: ACM, 2000, pp. 1–12.

[37] D. E. V. Pires, L. C. Totti, R. E. A. Moreira, E. C. Fazzion, O. L. H. M. Fonseca, W. M. Jr., R. C. de Melo Minardi, and D. O. G. Neto, "Fpcluster: An efficient out-of-core clustering strategy without a similarity metric," JIDM, vol. 3, no. 2, 2012, pp. 132–141.

[38] R. A. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[39] M. Bostock. Data-driven documents: Grouped bar chart. http://bit.ly/QsqidV. Accessed April, 2014. (2012)

[40] M. Bostock. Data-driven documents: U.s. airports, 2008 voronoi diagram. http://bit.ly/1ttgoI4. Accessed April, 2014. (2012)

[41] J. Howe, "The rise of crowdsourcing," Wired magazine, vol. 14, no. 6, 2006, pp. 1–4.

[42] D. Karampinas and P. Triantafillou, "Crowdsourcing taxonomies," in The Semantic Web: Research and Applications, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7295, pp. 545–559.

[43] J. Mortensen, "Crowdsourcing ontology verification," in The Semantic Web – ISWC 2013, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, vol. 8219, pp. 448–455.

 ISBN: 978-1-61208-358-2