SUPPORT EFFECTIVE DISCOVERY MANAGEMENT IN VISUAL ANALYTICS

by

Yang Chen

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2013

Approved by:

____

Dr. Jing Yang

____

Dr. William Ribarsky

____

Dr. Robert Kosara

____

Dr. Xiaoyu Wang

____

Dr. Min Jiang

ABSTRACT

YANG CHEN. Support effective discovery management in visual analytics.
(Under the direction of DR. JING YANG)

Visual analytics promises to supply analysts with the means necessary to analyze complex datasets and make effective decisions in a timely manner. Although significant progress has been made towards effective data exploration in existing visual analytics systems, few of them provide systematic solutions for managing the vast amounts of discoveries generated in data exploration processes. Analysts have to use off line tools to manually annotate, browse, retrieve, organize, and connect their discoveries. In addition, they have no convenient access to the important discoveries captured by collaborators. As a consequence, the lack of effective discovery management approaches severely hinders the analysts from utilizing the discoveries to make effective decisions.

In response to this challenge, this dissertation aims to support effective discovery management in visual analytics. It contributes a general discovery management framework which achieves its effectiveness surrounding the concept of patterns, namely the results of users' low-level analytic tasks. Patterns permit construction of discoveries together with users' mental models and evaluation. Different from the mental models, the categories of patterns that can be discovered from data are predictable and application-independent. In addition, the same set of information is often used to annotate patterns in the same category. Therefore, visual analytics systems can semi-automatically annotate patterns in a formalized format by predicting what should be recorded for patterns in popular categories. Using the formalized annotations, the framework also enhances the automation and efficiency of a variety of discovery management activities such as discovery browsing, retrieval, organization, association, and sharing. The framework seamlessly integrates them with the visual interactive explorations to support effective decision making.

Guided by the discovery management framework, our second contribution lies in proposing a variety of novel discovery management techniques for facilitating the discovery management activities. The proposed techniques and framework are implemented in a prototype system, ManyInsights, to facilitate discovery management in multidimensional data exploration. To evaluate the prototype system, two long-term case studies are presented. They investigated how the discovery management techniques worked together to benefit exploratory data analysis and collaborative analysis. The studies allowed us to understand the advantages, the limitations, and design implications of ManyInsights and its underlying framework.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor Assoc. Prof. Jing Yang who provided continuous support and encouragement throughout my doctoral studies. Over the years she has regularly offered a complete, fair, and accurate assessment of my work. Our discussions, her ideas and suggestions make my Ph.D. experience productive and stimulating. Besides research, she teaches me how to be maturer and do better at getting along with people. Her advices will follow me through my personal and professional life.

Special thanks to my committee, Dr. William Ribarsky, Dr. Robert Kosara, Dr. Xiaoyu Wang, and Dr. Min Jiang for their support, guidance, and helpful suggestions.

I sincerely acknowledge Dr. Shixia Liu for offering me great opportunities to do internships at Microsoft Research Asia, China. It was a good experience to work in a world class industry lab. She also advises me on many aspects of my professional life. I also thank Yangqiu Song, Furu Wei, Hao Wei, and all my colleagues in MSRA.

I would like to thank Assoc. Prof. Ye Zhao for his valuable advices and collaboration. The collaboration work was successfully turned into three papers. I also thank Jamal Alsakran for his support in algorithm development.

I would like to thank my colleagues and friends in Charlotte Visualization Center. It's been a wonderful and unforgettable experience to share my life with them. I appreciate the ongoing support from Chong Zhang, Jinbo Feng, and Siyuan Chen.

Lastly, I wish to thank my family for all their love and encouragement. My parents, Pu Chen and Xiaoming Yang, deserve special mention for their inseparable support. They support me in all my pursuits. Finally, I would like to thank my wife Sihui Zhang for her endless patience during the fulfillment of this dissertation. Without her love and faithful support, I would not have complete this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Nowadays gigabits of digital data are generated per person per year. People need to get information from the massive data to make decisions or solve problems. With the rapid advancement in data storage, data integration, and data mining techniques, people can effectively access and manage the data that was previously unavailable or too difficult to process [1]. This presents tremendous opportunities to discover new insights for making effective decisions and solving unexpected problems.

In response to these new opportunities, an emerging research area, known as visual analytics, has been proposed to address the grand challenge of analyzing the massive amounts of data [2]. Its basic approach is to create interactive visualizations so that human perception abilities and domain knowledge can be exploited together with computational powers to improve the reasoning process [3]. Consequently, people can derive profound insights from massive, dynamic data and make effective decisions.

A number of visual analytics approaches have been developed in a wide range of data analysis applications, such as health care, homeland security, terrorism detection, and financial market analysis. Many applications provide sophisticated forms of visualizations to facilitate the exploration of massive structured data (e.g., multidimensional data) and unstructured data (e.g., text collections). Examples include the visual analysis of massive text documents with the ThemeView 3D visual landscape in In-SPIRE [4] and ThemeRiver [5], large graph and tree analysis with Treemaps [6] and TreeJuxtaposer [7], high dimensional tabular data analysis with Xmdv Tool [8] and Polaris [9], and spatial and temporal data analysis with Oculus [4].

In addition, visual analytics can benefit from collaboration [10]. By partitioning the tasks between multiple collaborative works across different time and locations, collaborative visual analytics offers greater analysis scalability and ensure richer analytic outcomes. In practice, researchers in the visual analytics area have explored co-located synchronous systems (e.g., large displays and shared workspaces [11]), remote synchronous systems (e.g., real-time networked displays [12]), and asynchronous collaborative visualizations (e.g., online visualization communities such as ManyEyes [13] and sense.us [14]) to support different forms of collaboration.

With the advanced visualization and collaboration techniques, the scalability and productivity of visual analytics have been significantly increased. People need to carry out complex analytic tasks over days or even months and manage vast amounts of discoveries generated from various datasets and collaborators. A report from ManyEyes, a popular online visualization community, showed that the site received over 460 comments about discoveries, regarding to 2,100 datasets from 1,463 registered users in its first two months of life [13]. Information with such exploding volume and velocity poses significant new challenges to effective discovery management. For example, how can we easily record new discoveries and share them with collaborators? How can we effectively search and browse useful information from a large collection of discoveries? How can we flexibly organize the massive discoveries and explore their connections? How can we integrate these discovery management activities into the data exploration process to support decision making? Few of existing visual analytics systems provide a general solution to addressing these problems.

Our belief is that significant progress can be made toward the emerging gap by taking advantage of automated data analysis techniques, the wide bandwidth of human perception abilities, and human computer communication abilities enabled by visualization and interaction techniques. In this dissertation, we investigate a general, interactive visual exploration paradigm to address the discovery management challenges. In the reminder of this introductory section, we will formalize the problem and highlight the contributions and outline of the dissertation.

1.1 Research Problem and Approach

In this dissertation, we define a discovery as a piece of new knowledge that is useful for solving problems and making decisions. We define the application domain of discovery management techniques to be the applications where many individual discoveries can be explored from data and multiple discoveries need to be managed, i.e., to record, retrieve, organize, associate, and share among collaborative analysts for decision making activities such as creating and evaluating hypotheses. In particular, discovery management consists of two principles:

- Looking backward and looking forward: Decision making involves iterative information foraging and sense making loops. Users need to continuously gather information during the exploration, dynamically adjust exploration foci according to their new findings and new insights, and associate interrelated findings to form hypotheses. We refer to this process as a dynamic knowledge construction process. *Looking backward* supports the process by allowing users to retrieve and recall discoveries from past analysis steps. As important discoveries are retrieved, the users can associate them to build comprehensive, integrated insights. The integrated insights are what to drive new hypotheses and future analysis directions, namely *looking forward*. For instance, a financial analyst who has been monitoring stock market data over months would frequently connect previous patterns with the current market state to make predictions. Security and law-enforcement organizations would build integrated views of emerging threats and events from all available data sources to take timely actions.
- Constructing common ground: In collaborative visual analytics, many analysts collaboratively investigate data with different visualization tools and various types of expertise. In order to ground their analytic actions for making better decisions, the analysts need to review, manipulate, organize each other's findings to reach a shared understanding of them [15]. This process is known

as common ground construction [16]. Effective common ground construction reduces the cost of collaboration by avoiding redundant discoveries and minimizing the need to verbally confirm actions among the analysts [16]. This is especially critical for asynchronous collaboration since verbal communications between the collaborators are usually difficult or even impossible [10]. Therefore, discovery management should support multiple users to construct common ground in collaborative analysis.

The above two principles have been widely studied in the areas of social psychology, knowledge management, and sensemaking (see Chapter 2). By synthesizing the results from these studies, we argue that for effectively managing discoveries, analysts must be able to:

- *annotate* the key information and rich context of discoveries for reusing them;
- *retrieve* and *browse* discoveries to get useful information from them;
- *organize* and *associate* discoveries to connect them for drawing hypotheses;
- *exchange* and *share* the discoveries and hypotheses with collaborators.

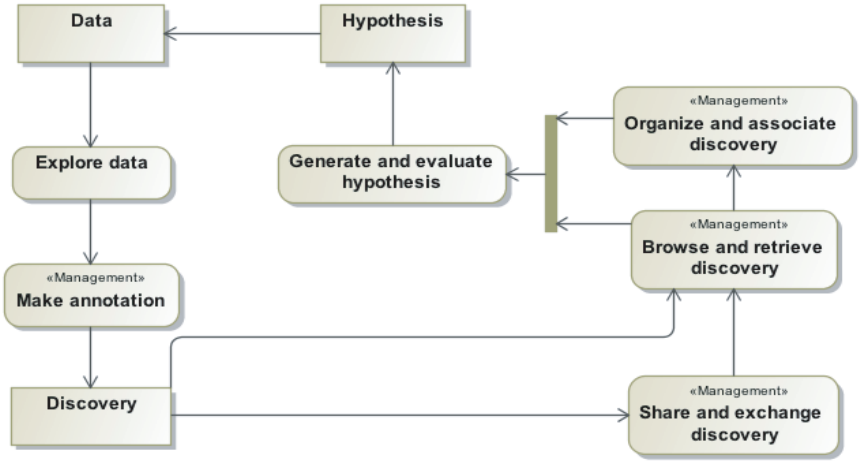

Data
Hypothesis
«Management»
Organize and associate discovery
Explore data
Generate and evaluate hypothesis
«Management»
Browse and retrieve discovery
«Management»
Make annotation
Discovery
«Management»
Share and exchange discovery

Figure 1.1: The workflow of discovery management activities in a decision making process.

Figure 1.1 shows a full picture of how the discovery management activities work together to facilitate the decision making process. When analysts explore data using interactive visualizations, they annotate their discoveries about the data. Later on, the analysts retrieve and browse useful information from the annotations of individual discoveries and use them for generating and evaluating hypotheses. The retrieved discoveries are further organized and associated based on the current analysis needs for new findings. The new findings are either used to evaluate the current hypotheses, or guide the exploration towards a new direction that may lead to more interesting discoveries and hypotheses. The analysts can also organize their discoveries and engage in collaboration by sharing or presenting the discoveries to their collaborators. By performing these management activities, the analysts can successfully evaluate hypotheses and make effective decisions.

Therefore, in this dissertation, we mainly focus on the problem that how to enable analysts to effectively *annotate*, *browse*, *retrieve*, *organize*, *associate*, and *share* discoveries in visual analytics processes.

1.2 Contributions

This dissertation contributes a general framework, novel techniques, and a system to support effective discovery management in visual analytics:

- We propose a general framework that enhances the effectiveness of a variety of discovery management activities and tightly integrates them in the visual data exploration. The framework addresses the challenges of effective discovery management surrounding the concept of pattern, namely the result of users' low-level analytic tasks [17]. Patterns are essential components of discoveries and convey the rich semantics of users' analytic tasks. From our observations in user studies, for the same type of data (e.g., multidimensional data), users can effectively classify most patterns into a small number of categories, independent from the domains/applications and visualization tools. In addition, the same set of information is often used to annotate patterns in the same category. Therefore, visual analytics systems can semi-automatically annotate patterns in a formalized format by predicting what to be recorded for patterns in popular categories. Based on the formalized annotations, discoveries can be browsed, retrieved, associated, organized, and shared effectively. These discovery management activities are seamlessly integrated with the interactive visual explorations to support the visual data exploration.

- We propose a set of novel discovery management techniques by integrating the pattern taxonomy, automated data analysis techniques, state-of-the-art visualization techniques, and novel interaction techniques. The techniques provide support, both visually and computationally, for facilitating discovery annotation, browsing, retrieve, organization, association, and sharing.
- ManyInsights is a multidimensional data exploration prototype we developed using the proposed framework and discovery management techniques. The individual discovery management techniques of ManyInsights, such as annotation and association, were evaluated through a set of formal user studies. In addition, experts from various application domains used ManyInsights to perform long-term exploratory data analysis using real datasets and real analytic tasks. The observations from these studies provided an in-depth understanding of how the proposed discovery management techniques work together to facilitate real-world exploratory data analysis.

1.3 Outline

The reminder of the dissertation begins by the following chapters:

- Chapter 2 discusses the background work related to the discovery management, including its theoretical basis, experiment designs, and state-of-the-art discovery management techniques. The limitations of the techniques are pointed out.
- Chapter 3 presents a general framework to support effective discovery manage-

ment in visual analytics. The framework leverages the efficiency of discovery management around the concept of pattern. It employs a pattern taxonomy to enhance the automation and efficiency of different discovery management activities. Based on the taxonomy, a visual exploration paradigm is provided to integrate the discovery management with interactive visual exploration. We also present ManyInsights, a multidimensional visual analytics prototype that support the discovery management using the proposed framework.

Guided by the discovery management framework, Chapter 4 through 7 present a set of novel techniques that are implemented in ManyInsights for managing discoveries in multidimensional datasets:

- Chapter 4 presents a pattern taxonomy for multidimensional data as our first step toward effective discovery management. The taxonomy characterizes the vast number of patterns that could be discovered in multidimensional datasets and defines the characteristics of each category of patterns.
- Chapter 5 introduces Click2Annotate, a semi-automatic discovery annotation approach. The core component of Click2Annotate is a set of annotation templates generated based on the pattern taxonomy. For annotating a certain type of discoveries, the template guides the system to retrieve the rich context information of discoveries from data and encode it in highly formalized annotations. We present a formal user study to prove the effectiveness of Click2Annotate.
- Chapter 6 introduces two novel techniques that utilize the rich context in annotations to retrieve and browse discoveries. The faceted discovery search allows users to search discoveries using custom navigation based on the context of discoveries. The scented discovery browsing technique allows users to flexibly access discoveries on data visualizations.
- Chapter 7 introduces a suite of toolkits to explore correlations among discoveries. We present an automatic technique to calculate discovery correlations

based on formalized annotations. Next, we present two interactive views that enables the exploration of correlations at different levels of detail. The dynamic discovery clustering display provides an overview of discovery clusters, their semantics, and their temporal evolution. The region graph enables the detail exploration of correlations for visual decision making. Finally, we present a case study and a user study to demonstrate the usefulness of the toolkits.

The system are evaluated and concluded in Chapter 8 and 9:

- Chapter 8 reports two long-term case studies of ManyInsights conducted by domain experts with real datasets and real research tasks. In the first case study, a domain expert used ManyInsights to conduct a 8-week data exploration for his own datasets and analytic tasks. In the second case study, a group of collaborative workers used ManyInsights to explore datasets and share their discoveries for collaborative reasoning. The studies provide an in-depth understanding of how the discovery management techniques work together to facilitate real-world exploratory data analysis.
- Chapter 9 concludes the dissertation and presents the remaining challenges for effective discovery management.

Parts of this dissertation have been published before, including:

- Y. Chen, J. Yang, and W. Ribarsky. "Toward Effective Insight Management in Visual Analytic Systems." *In IEEE Pacific Visualization Symposium*, 2009, pages 49-56.
- Y. Chen, J. Yang, S. Barlowe, and D.H. Jeong. "Touch2Annotate: Generating Better Annotations with Less Human Efforts on Multi-touch Interfaces." *In ACM Conference on Human Factors in Computing Systems (CHI) Extended Abstracts*, 2010, pages 3703-3708.
- Y. Chen, S. Barlowe, and J. Yang. "Click2Annotate: Automated Insight Externalization with Rich Semantics." *In IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2010, pages 155-162.

- Y. Chen, J. Alsakran, S. Barlowe, J. Yang, and Y. Zhao. "Supporting Effective Common Ground Construction in Asynchronous Collaborative Visual Analytics." *In IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011, pages 23-28.

CHAPTER 2: RELATED WORK

The research of discovery management for effective decision making has long resided in the realm of intelligent systems, organizational research, and social science. Recently, it has been receiving more attention from the visual analytics community. This chapter presents an in-depth survey of the related work in these areas. The survey begins with theories and empirical studies. They serve as the theory foundation and design guidelines of this work. Then, we present the state-of-the-art in visual analytics and discuss the limitations.

2.1 Multidimensional Data Visual Exploration

Visual analytics is an emerging research area that targets the grand challenge of analyzing massive amounts of data [2]. It combines techniques from multi-disciplinary fields, such as information visualization, statistics, machine learning, and cognitive psychology, for facilitating analytical reasoning. Among the motivations of the generation of this field, the need for analyzing large-scale multidimensional datasets is among the most significant ones since these datasets are standard in many application domains such defense, health, governance, business, and cyberspace. In this dissertation, we focus on the discovery management for multidimensional data and explore a visual exploration paradigm to facilitate multidimensional data exploration.

A number of techniques can be used in the proposed paradigm for exploring multidimensional data and generating discoveries. For example, automatic knowledge discovery techniques, such as subspace clustering algorithms [18], k-nearest neighbor search algorithms [19], and k-nearest match algorithms [20], can be used to partition a high dimensional data space into multiple smaller divisions. As meaningful divisions are constructed, they can be visually explored via less scalable visualization techniques (e.g., parallel coordinates [21], scatterplot matrices [22]), categorical data visualization techniques (e.g., parallel sets [23]), and geospatial and time series data visualization techniques (e.g., GeoTime [24]). In addition, multiple view techniques [25] can be used to handle divisions with mixed data characteristics, such as divisions with mixed numeric, categorical, and geospatial attributes.

2.2 Discovery Management Theory

Discovery management and decision making are widely studied in areas such as intelligent systems and organizational research. The proposed work has been inspired by various efforts from those areas. For example, Hori [26] found that knowledge evolves dynamically depending on the context. Such dynamic nature requires information workers to effectively manage their knowledge, such as capturing the knowledge, categorizing and linking information corresponding to the knowledge, and presenting them in a meaningful way [27]. Gavetti and Levinthal [28] used computer simulations to examine the role and interrelationship between search processes that were forward-looking, based on the actors' cognitive map of action-outcome linkages, and those that were backward-looking, or experience-based. In sensemaking research, Weick and Sutcliffe [29] also pointed out that sensemaking is a mixture of retrospect and prospect. These efforts provide a solid theory foundation to our "look forward and look backward" paradigm for dynamic knowledge construction.

In social and organizational research, researchers have investigated how management activities benefit collaboration in a variety of collaborative tasks, such as emergency task management [30], tactical operations planning [31], and collaborative information synthesis [15]. Often, collaborative workers come into collaboration only having completed their own individual work. They are unaware of what has been done and found by others. Therefore, collaborative workers need to manage and share their individual work to reach a common ground of the collaboration [16]. Effective common ground construction minimizes the need to verbally confirm actions and reduces the cost of collaborative effort [16]. Thus, our approach to support discovery management in visual analytics also benefits the collaboration environment.

2.3 Empirical Study of Discovery Management

During a visual analytics process, discoveries are captured from interactive visual exploration and used for supporting hypothesis generation and evaluation toward problem solving and decision making. A significant challenge faced by analysts is that large amounts of discoveries are often involved in the analysis process and need to be handled in a timely manner. To explore this challenge, researchers have conducted a set of empirical studies to examine how analysts manage their discoveries in different analytic tasks and analysis environments.

For example, Saraiya et al. [32] conducted a two-month study to examine how analysts use visualizations to gain insights into bioinformatics data. The study showed that analysts started the analysis by capturing as many interesting patterns as possible from the data. As new insights were discovered, they were connected with past analysis for additional questions and, hence, further directions [32]. In the later stage of the analysis, the analysts focused on reviewing and exploring the insights that have been captured. More specifically, they needed to create readable graphs to present the correlations between the insights and used different data formatting methods to detect their conflicts. The analysts considered the latter process equally important to the formal, but were inadequately supported by visualization tools.

Kang et al. [33] conducted controlled experiments to compare the use of visual analytics system Jigsaw with other three traditional text analysis tools in sensemaking of small document collections. They observed that analysts frequently extracted interesting entities from documents and added annotations to these entities. They also needed to draw connections between the entities to reveal their correlations. Systems missing these functionalities would hinder analysts from tracking and reusing entities. An important analysis stage, named "schematize", was also identified during the study. In this stage, the analysts used their preferred organizational scheme, such as a timeline and map, to organize captured entities. They claimed that visual analytics tools should enable flexibility and room for customizing organizational metaphors to support this analysis stage [33].

Robinson [15] investigated how analysts collaboratively synthesized individual visual analytic results in a collaborative environment. In the study, ten geography and disease biology analysts worked in pairs to synthesize analytic artifacts that were created individually. Based on analysis of video coding results, he identified a set of management activities commonly taken by the analysts, such as describing the information development process, reviewing individual artifacts, grouping similar artifacts, and identifying the overlaps of the artifacts. The study also provided several design implications for supporting collaborative synthesis, such as the use of multiple visual metaphors for organizing analytic results and the support of role assignment in collaboration.

Mahyar et al. [34] studied analysts' note taking and note organization behaviors in collocated collaboration. The results indicated that users often use multiple approaches (e.g., ordering by chronological history) to organize notes, which help them better communicate and discuss with each other. Experimental evidence of these studies, regardless of the specific tasks, resulted in parallel lists of tasks and design implications critical for effective discovery management in visual analytics. These tasks and implications, guide us in the design and implementation of the proposed discovery management techniques.

2.4 State of the Art

To support effective decision making, initial efforts have been directed towards managing discoveries in visual analytics approaches. In this section, we review a number of discovery management approaches to annotate, retrieve, organize, associate, and share discoveries in visual analytics.

2.4.1 Taxonomy

There exists considerable work on information visualization taxonomies. For example, Keim and Kriegel [35], Chuanh and Roth [36], Dix and Ellis [37], Ward and Yang [38], and Yi et al. [39] propose taxonomies on visualization interaction techniques. Card et al. [40], and Chi [41] present taxonomy on visualization models. Keim et al. [42] classify factors that differentiate various visualization techniques.

Among the existing taxonomic work, the taxonomies of users' analytic activities and tasks are closest to our work, since users often generate discoveries by performing analytic tasks. Gotz and Zhou [43] propose a visual analytic activity taxonomy based on Activity Theory [44] and observational experiments. In this taxonomy, users' visual analytic activity is classified into four levels: tasks, sub-tasks, actions, and events. They range in semantic richness and abstraction levels from high to low. Tasks correspond to an analyst's highest-level analytic goals, such as investigating the financial market. They are often domain or application specific. Sub-tasks correspond to more concrete analytic goals, such as detecting clusters, outliers, or correlations for multidimensional data. They are also called low level analytic tasks in other literatures [17]. Actions represent individual executable analytic steps such as zooming and panning a visualization view. Events correspond to the lowest-level of user interaction events, such as mouse clicks and button presses.

Based on Gotz and Zhou's taxonomy, we focus on managing discoveries at the sub-task level. Among existing sub-task taxonomies, there are Shneiderman's task by data type taxonomy [45], Wehrend and Lewis' cognitive task taxonomy [46], Zhou and Feiner's low level visualization system tasks [47], Lee et al.'s graph exploration tasks [48], Amar and Stasko's low level analytic task taxonomy for multidimensional data with analytic goals [17].

2.4.2 Discovery Annotation and Retrieval

Numerous visual analytics systems have been equipped with history mechanisms to capture users' low-level interaction events or parameter settings during a data exploration process. Users can revisit linear history using an undo-redo mechanism or visually explore it in tree visualizations [49] and graph visualizations [50]. Exploring the history records helps the users to infer the high-level logical constructs of the analysis and track their findings [43]. However, it is difficult to scale up these approaches to handle the vast amounts of low-level interaction events generated in a complex visual analytics process.

There is also a growing interests in recording analysts' analytic activity at the action level. Actions contain semantically meaningful behaviors. The management of actions is more efficient than low-level interaction events. Gotz and Zhou [43] identify three categories of actions common to different visual analytics tasks. Visual analytics systems, such as HARVEST [43] and Aruvi [51], utilize the action categories to automatically capture analysts' actions in a data exploration process. In Aruvi, a sequence of actions is visually conveyed by a horizontal-vertical tree, where nodes of the tree represent visualization states, and edges between adjacent nodes indicate the navigation resulted by actions. Users can revisit the visualization states sequentially in the tree using the undo-redo mechanism.

Visual analytics researchers have suggested the use of augmenting visual representations with annotations to record analysis details and discoveries. Compared to the analytic activity history, annotations summarize higher levels of knowledge and contain richer context and semantic information about the discoveries [52]. Moreover, annotations can be more easily shared and reviewed among collaborative workers. Most existing visual analytics systems rely on human beings to manually generate annotations. For example, Many Eyes [13] allows visualization users to share their discoveries or free thoughts by posting comments in a discussion forum. A URL bookmarking mechanism is used to point back from the comments to the associated visualizations so that users can revisit and review their discoveries. Aruvi [51] enables users to create notes to record analytic artifacts such as findings, assumptions, hypotheses, and causal relations. These notes are linked to a visualization state to facilitate revisit and recall. They can also be organized into groups to form a highly structured and systematic argumentation. Systems such as Sandbox [53] and sense.us [54] allow users to jot down their observations and opinions into visualization views. Ellis and Groth [55] propose using annotations to share discoveries among collaborators in collaborative data analysis. Analysts need to manually create annotations in a separate layer on top of data. Elias et al. [52] propose a "context aware" annotation approach for complex visualization dashboards. In this approach, annotations are transparent to data dimensions and data items so that users can browse and retrieve the same annotation from multiple correlated dashboard charts.

Recently, a few preliminary efforts have been made to take advantage of automatic analysis and visual exploration techniques to annotate discoveries. For example, the Nugget Management System [56] allows users to extract, refine, and record nuggets (subsets of multivariate data) with the help of automated analysis techniques. Useful statistical information about the nugget is automatically attached in addition to manual annotations given by users. Currently this system supports the annotation of clusters in multivariate data.

The annotations generated from the above approaches contain rich context information surrounding the discoveries. The information can be used to efficiently retrieve discoveries of specific visualizations or data. For example, Many Eyes [13] and sense.us [54] utilize keyword search to retrieve discoveries with comments containing keywords of interest. The comments are indexed and attached to both datasets and view parameters of visualization states so that all the comments associated with a visualization view or dataset can be promptly retrieved. Aruvi [51] organizes analysts' notes in a node-link diagram where the users can access the notes using keyword searches and text similarity metrics. However, the effectiveness of discovery retrieval in these approaches is highly depending on the quality of annotations.

2.4.3 Discovery Organization and Association

To make informed decisions, analysts often have to organize discoveries into coherent groups and reveal the interrelations between or within the groups [15]. A common approach to organizing discoveries is to use annotations. For example, web-based collaborative visualization systems such as sense.us [54] and Many Eyes [13] allow users to link free comments and graphic annotations to specific visualization views. The comments and annotations are usually manually generated by the users and contain high level semantic information about the discoveries associated with the views. The comments are frequently organized in a discussion forum where they can be retrieved by other users through browsing or keyword searches. CommentSpace [57] and Sandbox [53] go further by allowing users to tag discoveries and link them for supporting or conflicting hypotheses. Using the tags, users can also easily group and review discoveries for specific hypotheses. Nevertheless, these approaches rely on human being's effort to manually organize discoveries. Users often have to manually examine lengthy annotations for grouping and relating discoveries.

Shrinivasan and Wijk [58] propose an automated discovery association approach using exploration histories. Whenever users record a note about a discovery, the relevant analytic action trail is automatically recorded to form a context description of the note. Therefore, multiple notes can be grouped and associated by conducting automatic analysis techniques on their associated context descriptions. Such a context description also helps analysts to make inferences about their collaborators' high-level analysis strategies [58].

Many existing visualization systems provide highly formalized schemes, such as graph and matrix, to visually organize discoveries and represent their correlations. For example, systems such as Aruvi [51], Analyst Notebook [59], and Nugget Management System [60] allow users to manually organize and relate discoveries in a network structure. Tree Trellis and Table Trellis [61] support aggregation and comparison of linked free-text claims. Sandbox [53] allows analysts to jot down hypotheses and evidence and organize them in an automatically generated concept map. Evidence matrices [62] aggregate and make inferences according to analytic evidence. Rows contain multiple hypotheses and columns contain collected evidence. The cells of the matrix are populated with scores representing the degree to which the discovery supports or disputes the hypothesis. Text visual analytics systems such as Jigsaw [63] and CzSaw [64] provide a variety of organizational metaphors, such as network and table, to explore the relationships between entities extracted from documents. A shoebox is used to capture entities and documents, to record hypotheses, and to organize them into groups.

2.4.4 Discovery Sharing and Exchange

In collaborative visual analytics, a number of approaches have been developed to support sharing and exchanging discovery in both synchronous and asynchronous settings. In synchronous collaboration, real time shared views and instant communication tools are often used for sharing discoveries. For example, VizCept [12] allows users to keep track of each other's findings and relations in a shared concept map. Users can refer to such a shared view to ground their actions. Reality Instant Messaging [65] integrates an online social tool into visualization systems, helping users to coordinate their activities and interests in the decision making process.

In asynchronous collaboration, sharing and exchanging discoveries are challenging tasks since there lacks instant communication among asynchronous users. It is difficult for them to collaboratively identify significant discoveries and capture relationships among discoveries through face to face discussion or real-time communication as in synchronous collaboration. As a consequence, users have to manually retrieve, review, and organize each other's notes or annotations for sharing discoveries [13].

2.5 Summarization

In this chapter, we presented background work related to discovery management and reviewed a number of discovery management approaches supported by existing visual analytics systems. In summary, existing discovery management approaches suffer from the following problems:

- Manual annotation is often required for capturing the rich semantics of discoveries. Manual annotation is time-consuming and reduces users' interests in annotating discoveries. Moreover, manually generated annotations can be incomplete, imprecise, and hard to understand, which leads to difficulties in subsequent discovery management activities such as discovery retrieval and association. Although a few efforts have been directed toward automatic discovery annotation, the automation of these approaches is conducted at the action or event level, based on based on Gotz and Zhou's taxonomy [43]. Since information captured from the action or event level has limited semantic meanings to users, the generated annotations can be more difficult to retrieve and understand than the annotations generated at the sub-task level. To the best of our knowledge, there exists no general annotation approach that conducts the automation at the sub-task level.
- Most existing approaches require users to manually detect and organize correlations among discoveries. It is difficult to use manual approaches to handle complex sensemaking tasks where a large amount of discoveries and multiple users are involved. Although a few efforts have been made to automatically associate discoveries using exploratory histories, it can be difficult to organize and summarize discoveries according to their high level semantic meaning, if the exploration histories consist of exploration steps with little semantic meaning. In addition, the large volume of exploratory steps toward each discovery may

hinder a system of organizing and associating a large number of discoveries. Moreover, existing visual analytics systems merely provide static views to organize and associate discoveries. They are difficult to scale to the fast growing discoveries for users with diverse information needs.

- It is time consuming to search and browse recorded discoveries with existing approaches, especially in an asynchronous collaboration environment. In such environments, constructing queries to fetch stored discoveries is often challenging, since different users may use various terms to express similar meanings when manually annotating discoveries. Users may also have difficulties in understanding discoveries recorded by others, since the annotation process is not well regulated.
- Few, if any, existing approaches provide a general discovery management framework that seamlessly integrates the discovery management activities to support the dynamic knowledge construction process.

The above challenges need to be addressed to achieve effective and efficient discovery management. Toward this goal, we propose a general discovery management framework and a set of discovery management approaches based on this framework. They are summarized in Chapter 3.

CHAPTER 3: A GENERAL DISCOVERY MANAGEMENT FRAMEWORK

Discovery management is an essential step in the process of transferring information from massive data to the human mind for making effective decisions. However, it is poorly supported in existing visual analytics systems. In this chapter, we propose a general framework that employs taxonomy and a visual exploration paradigm to achieve effective discovery management. Based on the framework, we propose a set of techniques to facilitate various discovery management activities, such as discovery annotation, retrieval, browsing, organization, association, and sharing. The framework and the techniques are integrates in a prototype system, ManyInsights, to support the sensemaking of multidimensional data. A concrete scenario of visual sense making on real datasets illustrates how the system works.

3.1 Introduction

Recently, numerous visual analytics approaches have been developed to facilitate sensemaking of complex, massive data. A vast amount of discoveries is often captured from the data using these approaches. To effectively support analytic activities such as hypotheses evaluation and collaborative reasoning, discovery management, the process of annotating, retrieving, associating, and organizing discoveries, becomes essential in visual analytics approaches. We argue that effective discovery management should allow users:

- To keep found things found [66], i.e., to allow users to capture, annotate, retrieve, and inspect discoveries;
- To reveal the correlations among discoveries and allow users to interactively explore the correlations; and
- To aid collaborative workers in sharing and exchanging discoveries.

A few efforts have been directed towards effective discovery management in visual analytics systems. However, as we state in Chapter 2, existing approaches suffer from several problems: (1) Manual discovery annotation is often required. It reduces users' interests in annotating discoveries and leads to difficulties in subsequent discovery activities, such as discovery browsing and retrieval; (2) Most existing approaches require users to manually detect and organize relationships among discoveries. It is difficult to use manual approaches to handle complex analytic tasks where a large amount of discoveries and multiple users are involved; and (3) It is time consuming to search and reuse recorded discoveries with existing approaches, especially in an asynchronous collaboration environment.

In this chapter, we present a general discovery management framework to achieve effective discovery management in visual analytics. The framework addresses the above challenges surrounding the concept of pattern, namely the result of users' low-level analytic tasks. Patterns are essential components of discoveries and permit the construction of discovery. The type of patterns that can be discovered from data is predictable and application-independent. Thus, it is possible to develop general approaches to allow users to effectively annotate, browse, retrieve, associate, and share patterns. Toward this goal, we first propose pattern taxonomy to categorize various patterns and capture their common features. Such taxonomy serves as the foundation of the framework and enhances the automation and efficiency of a variety of discovery management activities. Based on the taxonomy, we explore a visual exploration paradigm that integrates the discovery management activities with interactive visual exploration to support the dynamic knowledge construction process.

Guided by the framework, a variety of automated discovery management techniques is developed, such as a semi-automatic discovery annotation technique, flexible discovery browsing and retrieval techniques, and automatic discovery correlation exploration techniques. They are implemented and integrated in a multidimensional data exploration prototype, ManyInsights, to manage discoveries in multidimensional data. In the following sections, we first provide a refined definition of discovery and introduce the concept of pattern. Next, we present the discovery management framework and its two essential components: the taxonomy and the visual exploration paradigm. Finally, we introduce the ManyInsights and present a use case scenario to demonstrate its usefulness.

3.2 Discovery - A Close Look

In visual analytics, discovery can be defined as a piece of new knowledge or insight that is useful for solving problems and making decisions. Discovery can be "complex" and "deep" [67] and have different levels of abstractions regarding to different data and application domains [68]. In order to develop general and effective discovery management approaches, a close look must be taken at what is a discovery and how to present a discovery.

3.2.1 Definition

Researchers have made initial efforts towards defining and classifying users' discoveries in visual analytics processes. For example, Pousman et al. [69] identified four types of insights that are commonly generated in visual analytics process: analytic insight, awareness insight, social insight, and reflective insight. Among these types, analytic insight is the most traditional sense of users' discoveries supported in visual analytics systems [69]. It comes from the exploratory analysis and consists of a body of data that has been given meaning through users' analytic tasks or activities [69]. The management of analytic insights is tightly coupled with data analysis techniques and visualization techniques, and can benefit a variety of types of application domains. Therefore, this dissertation focuses on addressing the challenge of discovery management for analytic insights.

Since analytic insights are direct results from users' analytic tasks and activities, they can be characterized based on analytic activity taxonomies. Gotz and Zhou categorized user's analytic activities into four abstraction levels: task, sub-task, action, and event [43]. They range in semantic richness from high to low. Based on this categorization, we argue that managing discoveries at sub-task level is a promising research direction. First, information captured from sub-task level, such as clusters and outliers, have higher semantic richness than action and event levels, such as zooming and mouse clicks. Therefore, the former will be easier to understand and reuse than the latter. Moreover, sub-tasks are less application-dependent than tasks. For example, Wehrend and Lewis [46] identified 11 low-level analytic tasks (sub-tasks) that can result in an analytic insight, such as classification and ranking. They are general for a wide range of application domains. Therefore, we propose a general discovery management approach for managing discoveries at the sub-task level.

3.2.2 Model

In real-world data analysis, a discovery contains information not only about data, but also regarding to users' mental model [70]. It is difficult to handle using a general approach. Therefore, it is necessary to identify the component of discoveries that can be effectively handled across different domains and users. For this, we propose a three-components discovery model, as shown in Figure 3.1. The model consists of a *data pattern*, pattern in short, extracted from data under analysis, such as the outliers and clusters, a *domain/application knowledge base* against which the pattern is evaluated, and *objective and subjective evaluations* of the pattern against the knowledge base. In a typical case, an analyst discovers a pattern as a result of low-level analytic task during an interactive visual exploration process. The analyst then evaluates the pattern against the knowledge base to see if it is a significant and reliable piece of evidence that can be used in the sensemaking process. The pattern, the knowledge base applied, and the evaluations construct a discovery for the sensemaking process.

Among the three components, the knowledge base is difficult to handle using a general approach, since it varies significantly between datasets, applications, and an-

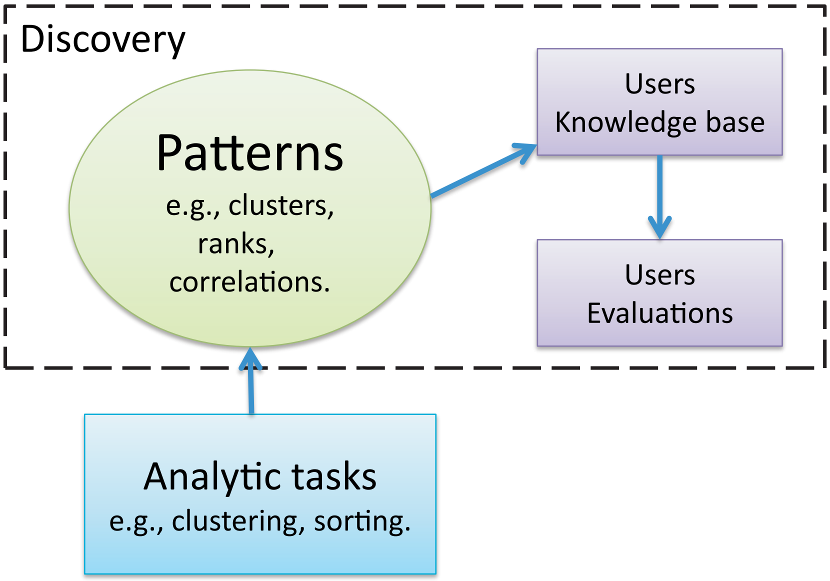


Figure 3.1: The three-component discovery model.

alysts. Subjective evaluations are dependent on users' knowledge base. On the other hand, the pattern composes the essence of discovery. The types of pattern that can be discovered from data is predictable [17] and are independent from datasets, applications, and analysts. In addition, patterns are direct products of users' analytic tasks in the visual exploration process and thus their management can be tightly integrated into the visualization system. Therefore, we believe that general approaches can be developed to allow visualization users to effectively and efficiently detect, annotate, associate, retrieve, and share patterns using automatic or semi-automatic approaches. A general discovery management framework is proposed based on this idea. Since patterns are the fundamental components of discoveries and bridge the visual exploration process and discovery management, it will be feasible to extend the general approach in various visual analytics applications by adding real-world knowledge and evaluation from mental model and thus lead to effective and efficient discovery management.

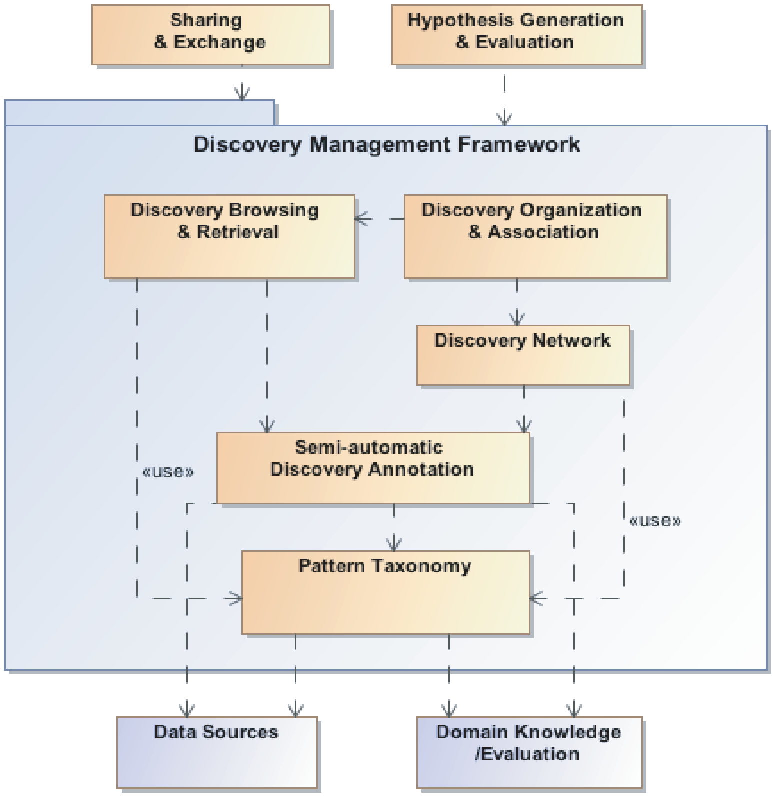


Figure 3.2: The general discovery management framework.

3.3 A General Discovery Management Framework

The general discovery management framework is shown in Figure 3.2. The foundation of the framework is a pattern taxonomy that summarizes information about pattern categories that can be discovered from data, pattern attributes, and the relations in which a discovery can be associated with another (Section 3.3.1). The taxonomy serves three important functions. First, it enhances the automation of discovery annotation. After users discover a pattern from data and decide its category, the computer can automatically collect and extract information about the pattern following the taxonomy to capture its semantics. Users only need to provide the knowledge base and the subjective evaluations to complete an annotation. Second, since the information obtained for each category of patterns is predictable, annotations can be highly formalized by the computer. This greatly enhances the automation of other discovery management activities. For example, the information can be automatically indexed and stored for flexibly browsing and retrieval discoveries. Clusters of similar discoveries can be automatically constructed since the computer can capture correlations among the formalized annotations following the taxonomy. Finally, formalized annotations enable effective communication among multiple users and multiple systems to share discoveries. Upon the taxonomy, the framework provides a visual exploration paradigm that integrates discovery management with interactive visual exploration to support visual exploration process. In the following sections, we introduce the taxonomy and the exploration paradigm in detail.

3.3.1 Taxonomy

In visual analytics, researchers have identified a common set of low-level analytic tasks that are repeatedly performed for data, regardless of the visualization tools being used or specific application domains being involved [17]. Since patterns are direct results of users' low-level analytic tasks, it is feasible to construct a general pattern taxonomy to categorize various patterns. Such a categorization is essential for developing effective discovery management approaches. Without the categorization of patterns, it is hard to answer what discoveries are to be captured and managed by a general visual analytics system.

In our user experiments (Chapter 4), we observed that users often used a common set of information to annotate discoveries falling in the same category. The information helped them to understand and recall their discoveries and enabled them to search, organize, and associate discoveries. Based on this observation, capturing the common characteristics for discoveries is essential for effectively annotating and managing them. In particular, the following characteristics should be included:

- Data content information that describes the data related to discoveries to enable access to the data;
- Context information to describe the context of discoveries to enable access to

the context; and

- Interaction and visualization methods that lead to the discoveries.

In the proposed framework, the taxonomy will serve the following purposes:

- Providing a standard language among the users, the systems, and the automatic analysis techniques for effectively communicating with discoveries;
- Enabling semi-automatic discovery annotation;
- Enabling flexible discovery searching and browsing; and
- Enabling automatic organization and association of discoveries generated by different users and systems.

3.3.2 Visual Exploration Paradigm

Based on the taxonomy, the discovery management framework enables a visual exploration paradigm that tightly integrates the following discovery management activities with interactive visual exploration:

Discovery Annotation: effective discovery annotation summarizes the high-level knowledge of discovery, such as the categories of the patterns, contents, and contexts. The generated annotations allow users to organize, browse, retrieve, associate, and exchange discoveries using the information contained in them. Based on the taxonomy, we propose a semi-automatic discovery annotation approach that is tightly coupled with existing visualization techniques and enables annotation efficiency. In particular, after a pattern is distinguished by visualization through interactions (such as brushing) and its pattern category is decided (manually or automatically), the system will know what needs to be extracted from the data according to the attributes of the specific pattern category listed in the taxonomy. The automatically extracted information will be used to annotate the pattern and visually present to users in a formalized format. The users will be allowed to interactively improve the automatically generated annotations for more flexibility. For example, they can attach personal tags to record their domain knowledge or evaluations in an annotation.

Discovery Browsing and Retrieval: When annotations are automatically generated, the same vocabulary will be used for all patterns and thus the discoveries can be easily indexed, browsed, and retrieved using keywords in their annotations, as if the way that tags are used in YouTube [71]. For example, we can allow users to search discoveries by using rich context information contained in their annotations and browse them using document visualization techniques by treating the annotations as documents. Moreover, users can flexibly browse and retrieve discoveries on the visualizations being explored. Visual indicators can be attached to the visualizations to represent discoveries and allow users to effectively browse them without cluttering the visualizations and flexibly drill down to detailed explorations.

Discovery Organization and Association: Users need to organize and associate discoveries to reveal their correlations. In the framework, the same vocabulary is used in annotations so that the discoveries, either generated by different users or different systems, can be automatically associated through these information. For example, discoveries can be associated according to the dimensions or the data elements they contain. Based on this, discovery clusters and discovery network can be automatically constructed according to the correlations among the discoveries. Dynamic visualization techniques can be used to visually convey the discovery clusters and help users to track their evolution over time. Graph visualization techniques can be applied to help users interactively navigate in the network and browse the discoveries using graph interactions.

Discovery Sharing and Exchange: Standard discovery exchange requests can be generated to allow efficient discovery exchange in collaborative visual analytics. For example, when a user wants to get information from other users, she first requests an automatically generated form listing attributes of a pattern in a desired type. The user fills part of the form to express her information need and leave the attributes she wants to learn from her collaborators empty. She then sends the form to her collaborators so that they can complete the form, either manually or automatically, and send it back to her.

3.4 Multidimensional Data Exploration Prototype

The ManyInsights is a fully working prototype of the general discovery management framework for multidimensional datasets. Similar to existing online visualization applications such as Many Eyes [13], ManyInsights is a web based visual analytics system that supports both individual and collaborative visual analysis. Individual users can upload multidimensional datasets, create visualizations (e.g., scatter plot and parallel coordinates), and share the datasets and visualizations with colleagues. Beyond these commonly supported tasks, ManyInsights provides rich discovery management functions.

3.4.1 System Implementation

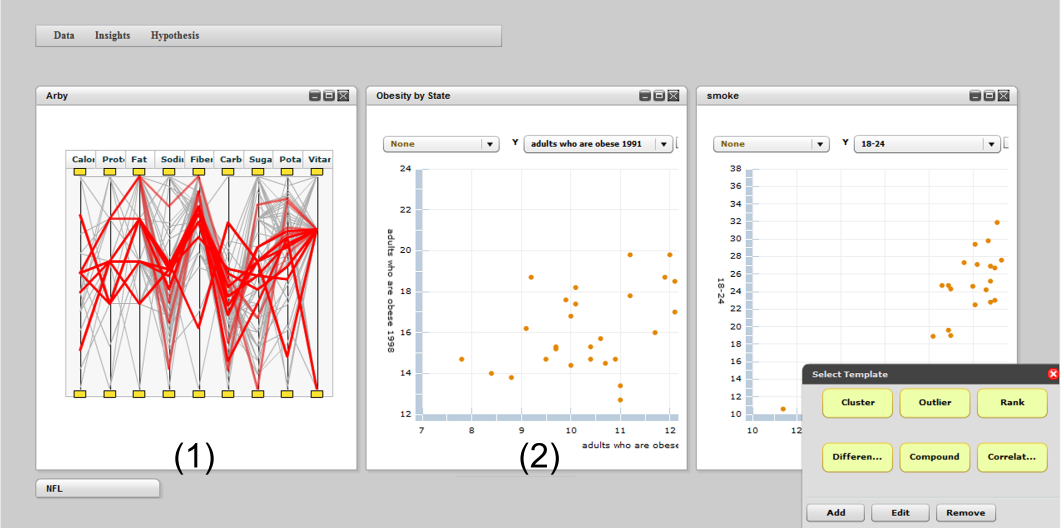


Figure 3.3: The data visualization interface. ManyInsights allows users to create multiple coordinated visualizations for exploring datasets. (a) A parallel coordinates view. (2) A scatter plot view.

ManyInsights is implemented in Flex [72], a web-based application and UI framework. The implementation architecture is based on a client-server architecture and consists of three important modules: data and visualization module,discovery management module, and visual interfaces (see Figure 3.4). They are described in the following sections.

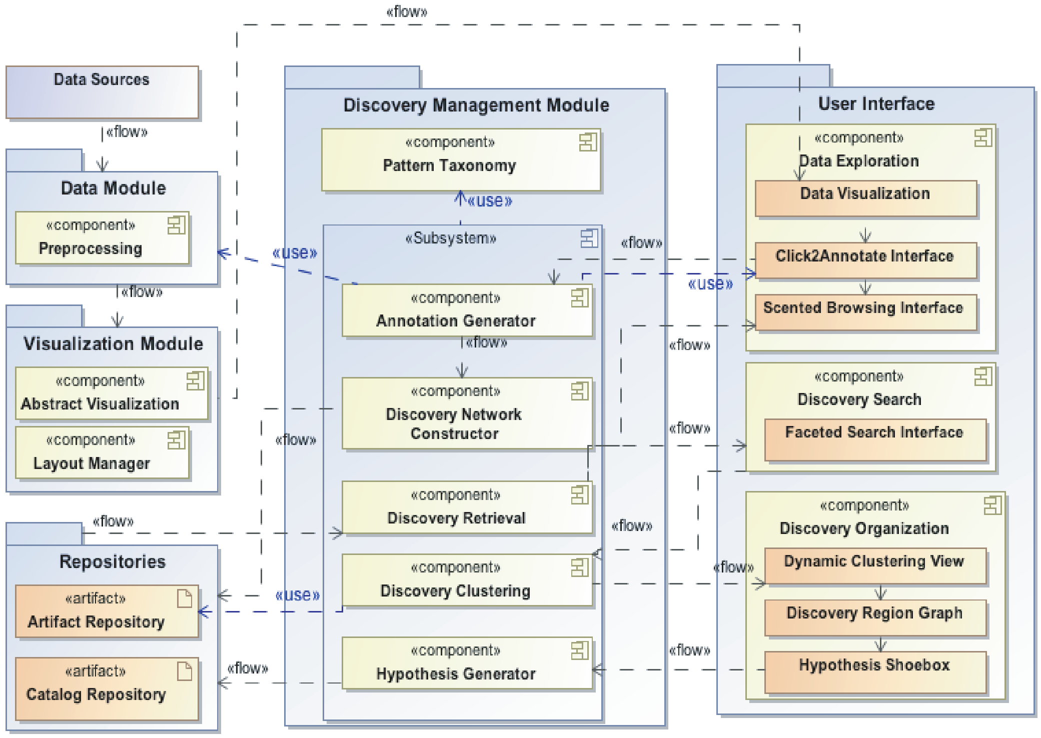


Figure 3.4: The implementation architecture of ManyInsights. The architecture consists of three modules: data and visualization module, discovery management module, and visual interface.

3.4.1.1 Data and Visualization Module

The data and visualization module processes datasets and generates visualizations for them. Multidimensional datasets serve as the input of the module. The data module provides necessary preprocessing functions, such as data normalization, data transformation, dimensionality reduction, to make the datasets suitable to visualize. Metadata, such as the size of dimensions and statistical information, is also extracted and stored. Afterwards, the processed datasets and metadata are fed into the visualization module.

The visualization module generates the suitable visualizations in response to users' analysis queries. An abstract visualization class is implemented to house the common attributes for all visualizations, such as the rendering, color-coding, and interactions. Each type of visualizations is implemented separately with minimal dependency on each other, allowing users to add more types of visualizations to the system. When multiple visualizations are requested, the layout manager controls their layouts and coordinates them using linking and brushing techniques [73].

3.4.1.2 Discovery Management Module

The discovery management module is a central element of the system's overall design. It monitors users' data exploration actions in the client visual interface, manages the discovery generated from the interface, and communicates with back-end databases to store and retrieve discoveries. The taxonomic information, such as the discovery categories and attributes, is organized and stored in a database accessible by all the components of the module for automated discovery management as required.

When users capture a discovery on the client visualizations, the annotation generator handles the annotation request by semi-automatically generating annotations using the taxonomic information. The generated annotations are fed into the discovery network constructor, where the correlations between the new discoveries and the stored discoveries are calculated. In addition, when users search for discoveries in the client visualizations, the discovery retrieval component retrieves the related information from the database and restores the annotations using the information. The clustering component computes the clusters for a collection of discoveries based on their correlations. The correlations are also mapped to physical force between the discoveries which will be used to visually present them in the client visual interface. The clustering component is optimized to achieve real time computation, which provides the instant feedback for the dynamic clustering request. Finally, the hypothesis generator generates annotations for hypotheses, links them with associated discoveries, and stores them in the database.

3.4.1.3 Visual Interface

Guided by the visual exploration paradigm, ManyInsights provides a variety of visual interfaces to support users in effectively performing discovery management tasks. The following scenario describes how they work together to facilitate a sensemaking process

- Users visually explore one or more datasets in the visualization for discoveries. After they find a discovery, they highlight the data of interest, select the type of the pattern, and enter the knowledge base and subject evaluations. ManyInsights will automatically collect content and contextual information of the pattern and use them together with other user input to generate a formalized discovery annotation. The annotation is stored in the database, which can be shared by many users in a collaborative analysis environment. Pair-wise discovery correlations are calculated between two formalized discovery annotations.
- Later on, the users retrieve and browse discoveries annotated by themselves or other users from the database via a faceted search interface. They can also browse the discoveries in related visualizations using scented insight browsing.
- After the users retrieve discoveries of interest, they can interactively explore them in an automatically generated dynamic discovery clustering display. This view reveals discovery clusters consisting of closely related discoveries, the discovery history of these clusters, and their semantics. According to the drifting interest of the users, correlations among the discoveries can be calculated differently to reveal different clusters.
- The users create hypotheses and associate the discoveries with the hypotheses. Discoveries associated with one or more hypotheses can be examined in detail in the region graph for visual sensemaking.
- The users annotate their key findings and hypotheses for future exploration.

3.4.2 Use Case Scenario

We provide a scenario of how ManyInsights and its underling discovery management framework work in visual analytics process. In this scenario, Mary and Tom are two analysts that work on the task of detecting the relationship between carbon dioxide emission and global warming in an asynchronous collaboration. The datasets used in this scenario are real data sets uploaded to Many Eyes for an ongoing discussion of a similar topic.

First, Mary uploads the dataset "USA emissions per capita by state" to ManyInsights and creates a scatterplot view to visualize it. From this view, Mary discovers the pattern that the Wyoming has the highest emissions per person among all the states. According to this pattern, Mary suspects that Wyoming might contribute more to global warming than the other states and she decides to record this discovery. She selects the *rank* category for the discovery and the system automatically creates an annotation form with most information filled. Mary manually records her hypothesis and stores the annotation into the database.

A few days later, Tom wants to know which states make significant contributions to weather warming. He logs into ManyInsights and submits a search in the faceted discovery search for discoveries in the *ranking* category and with the keyword "emissions" in dimension names. The system returns him some discoveries, including the discoveries Mary generated.

Tom reviews Mary's discovery. Since Tom knows that Wyoming has an extremely low population, he suspects that Mary might have ignored the overall emission amount of the states when she made her judgment. Thus, Tom loads the dataset "USA overall emissions by state" and creates a bar chart on it. From the bar chart, he discovers the pattern that Texas, Florida, Ohio and New York have much higher overall emissions than Wyoming. He thus records this pattern as a *difference* and attaches it to Mary's discovery.

Later, Mary gets a notification about the new finding captured by Tom and reviews it in the system. She remembers that she has discovered some interesting findings about New York and Texas before. To review how these findings relate to the discovery, she submits a search in the faceted discovery search to retrieve all her previous discoveries and groups them by data items in the dynamic discovery clustering display. By browsing the semantics for each group, she observes a cluster of discoveries about Texas. Mary sends the discovery cluster to the region graph, highlights the Tom's discovery in the graph, and examines the links between her discoveries and the highlighted discovery. Finally, she records her findings and hypotheses and shares them to Tom.

3.5 Conclusion

In this chapter, we presented a general discovery management framework to support effective discovery management in visual analytics. The framework leverages the efficiency of discovery management around the concept of pattern, and provides systematic taxonomy and exploration paradigm to facilitate different discovery management activities. Using this framework, users can generate formalized annotations to capture the rich context of discoveries. They can flexibly search discoveries and interactively browse them during the data exploration. The framework also allows users to flexibly organize and associate discoveries to explore their correlations. Finally, it allows collaborative workers to effectively share and browse each other's discoveries.

Guided by the framework, a set of novel discovery management techniques is developed and integrated in a multidimensional data exploration prototype, ManyInsights. In the following chapters, we will present these techniques in detail. In Chapter 4, we present a pattern taxonomy that categorizes and characterizes patterns in multidimensional data. In Chapter 5, we present Click2Annotate, a semi-automatic discovery annotation approach based on the taxonomy. In Chapter 6, we introduce two techniques to effectively retrieve and browse discoveries. In Chapter 7, we introduce a suite of visual analytic toolkits to explore the correlations of discoveries.

CHAPTER 4: PATTERN TAXONOMY FOR MULTIDIMENSIONAL DATA

In Chapter 3, we propose a general discovery management framework to achieve the effectiveness of discovery management surrounding the concept of pattern. Our assumption was that the type of pattern that can be discovered from data is predictable and application-independent. Therefore, categorizing patterns and summarizing their essential attributes will greatly enhance the automation and efficiency of discovery annotation, which ultimately benefit other management activities, such as discovery retrieval and organization. In this chapter, we present a pattern taxonomy for multidimensional data as a proof of the approach. The taxonomy is integrated in ManyInsights, serving as a foundation for all the discovery management techniques.

4.1 Introduction

With the explosion of new information visualization techniques and the increasing complexity of visual analysis process, taxonomy is playing an important role in comprehensively understanding the new techniques and the flow of human reasoning. In practice, researchers have intensively produced taxonomies for users' visual analytic tasks [17], analytic activities [43], and interaction techniques [39]. However, there are few, if any, general taxonomies for users' discoveries in visual analytics. Without taxonomy of discoveries, it is impossible to build effective discovery management approaches. For example, without a categorization of discoveries, it is hard to answer what is to be managed by a general visual analytics system.

However, it is often non-trivial to construct general taxonomy for discoveries. A significant challenge is that discovery is a complex concept that is associated with not only data under analysis, but also objective and subjective evaluations and real-world knowledge, which are stored in users' mental model. They vary significantly from different application domains, data, and users [70]. For examples, an economists and an environmentalist might have varied, or even opposing views toward a country that is considered as an outlier of greenhouse gas emission.

To address this challenge, we propose a three-component discovery model (Chapter 3). In this model, we identify pattern as the domain-independent component that captures the essential semantics of a discovery. Characterizing patterns for a certain type of data provides methods to access the data and context of discoveries across different applications, different visualizations, and different users. Therefore, pattern taxonomy forms the foundation of the general discovery management framework, enabling the automation of a variety of discovery management activities.

This chapter presents our work towards constructing such general pattern taxonomy for multidimensional data. Our goal is to categorize the vast number of patterns that are frequently discovered from multidimensional data and define the characteristics of each category of patterns. Different from existing taxonomy construction approaches, we adopt a multi-stages approach that includes extensive literature surveys, user experiments, and domain expert interviews. The resulted taxonomy provides a solid basis for the discovery management framework and enables the development of a variety of discovery management techniques.

4.2 Taxonomy Construction

A pattern taxonomy for multidimensional data categorizes various patterns that can be discovered from multidimensional data and describes their essential attributes. We argue that a general pattern taxonomy needs to meet the following criteria:

- Completeness: the taxonomy should cover the majority of patterns that can be discovered using various visualization tools and from multidimensional data sets of various sizes and dimensionalities in different application domains;
- Unambiguous: the taxonomy should accurately and clearly distinguish different types of patterns;

- Independence: the taxonomy should be independent from the application domains that generate the multidimensional data sets, the visualization and interaction techniques that are used to discover the patterns, and the users who discover the patterns; and
- Utility: the taxonomy should be feasible to use in pattern and discovery management.

Toward the above goals, we used a multi-stages process to construct pattern taxonomy for multidimensional data, which is described as follows:

- A literature survey on existing visualization taxonomy work and visualization techniques was conducted to generate an initial pattern categorization;
- The initial categorization was evaluated and refined through an experiment and a user study using real discoveries from real users;
- Interviews of domain experts were conducted to further evaluate the categorization and to learn the attributes of patterns that are essential in their discovery management tasks; and
- A literature survey on existing statistical and data mining work was conducted to summarize essential attributes for each category of patterns.

4.2.1 Literature Survey for Categorizing Patterns

We constructed an initial pattern categorization by conducting a literature survey of existing visualization taxonomy work and existing visualization techniques. We noticed that the taxonomy of visual analytic tasks is the most related to our pattern taxonomy among all taxonomy work since there is a strong tie between patterns and visual analytic tasks: users often discover patterns from visualizations by performing visual analytic tasks, i.e., visual analytic tasks are the analytical processes and patterns are the consequences.

Besides examining existing task taxonomies, we also reviewed 98 papers on multidimensional visualization from 00-07 IEEE InfoVis and VAST conferences and symposiums, which are the main avenues of information visualization techniques. These papers either present new or evaluate existing multidimensional visualization and interaction techniques. We examined these papers for patterns that can be discovered using the techniques under discussion in them.

After this turn of literature review, we constructed an initial pattern categorization that captures the results of most tasks considered in the task taxonomies and covers most patterns discovered from the technique and evaluation papers. In the initial categorization, there are ten big categories, namely *value/derived value, distribution, difference, extreme, rank, categories, cluster, outliers, association, and trend.* After our user experiment and user study (see Section 4.2.2 for more details), two other categories, namely *compound pattern* and *meta pattern*, were added. We define rows in a multidimensional dataset as items and columns in it as dimensions. Most categories of patterns exist in both the item space and the dimension space. For each category we gave a formal definition, along with examples extracted from real user discoveries posed in Many Eyes [13].

4.2.2 User Study for Evaluating Pattern Categorization

Although we conducted an extensive literature survey, the completeness and unambiguousness of the initial categorization are still in doubt. First, few existing task taxonomies have been evaluated in diverse real applications involving real users, real data, and real tasks. Second, few existing visualization and interaction techniques were designed for discovering all kinds of patterns. As a consequence, the initial pattern categorization needs to be evaluated and refined with patterns from a diversity of real users, real data sets, and real tasks. Toward this goal, we sampled patterns discovered by users of Many Eyes [13] and conducted an experiment and a user study.

Many Eyes [13] is a pubic collaborative information visualization web site where users visually explore data sets contributed by themselves or others and share their findings by posting comments in a discussion forum. Since Many Eyes is quite popular, a large number of discoveries are reported daily as comments by a large number of users ranging from scientists, managers, to sports fans [13]. These discoveries come from a wide range of data sets, most of which are real data sets from real application domains. In addition, the quality of the discoveries can be examined since visualization is attached to each comment. We thus considered Many Eyes comments as a good source of patterns from real users, real data sets, and real tasks.

For our experiment and user study, we collected all comments posted to Many Eyes between January 2007 and January 2008 and manually picked out patterns embedded in them. For duplicative patterns that have same data elements and same categories, we just picked out one of them. Patterns about data types other than multidimensional data were also removed. As the result, we got a sample containing 215 patterns which were collected from 56 multidimensional data sets. Some data sets contained temporal and geographical dimensions.

4.2.2.1 Experiment for Completeness Testing

An experiment was conducted to examine if the initial categorization covered the majority of the patterns contained in the Many Eyes sample. In particular, we reviewed all 215 patterns and tried to fit them into the pattern categorization. For example, the pattern "big drop in males becoming eye doctors in the past ten years" was classified into the *trend* category and the pattern that "relatively fewer number of females are going into business school than male" was classified into the *difference* category. We also counted the number of patterns falling into each category.

Among the 215 patterns, there were 63 patterns that did not fit into any categories in the initial categorization. They fell into one of the following situations:

- Compound patterns: there were 46 patterns that were patterns about patterns. For example, the pattern "it's interesting how different the second letter distribution is from the first letter distribution" contains a *difference* pattern about two *distribution* patterns.
- Patterns about meta data: there were 17 patterns about data itself such as missing values or errors in the data sets, appearance or disappearance of di-

Table 4.1: The result of comments classification.

Knowledge type	**Number of comments**	**Percentage**
Trend	55	25.6%
Compound pattern	46	21.4%
Outliers	41	19.1%
Difference	31	14.4%
Association	27	12.6%
Extreme	25	11.6%
Meta pattern	17	7.9%
Value/Derived value	16	7.4%
Categories	9	4.2%
Cluster	7	3.3%
Distribution	5	2.3%
Rank	3	1.3%

mensions, and meanings of labels. For example, the pattern that “a change happened between 1999 and 2000 when a bunch of new categories showed up” was about the appearance of new dimensions. The pattern that “the Soviet Union has no action movies? Can that be right?” was about data quality.

As a consequence, we added two additional categories into the initial categorization, namely *compound pattern* and *meta pattern* to fit those patterns in. In addition, we decomposed each compound pattern into multiple elementary patterns and counted them not only in the *compound pattern* category, but also in the elementary pattern categories. Table 2 shows the final result. In this table, categories are sorted according to the total number of related patterns in the Many Eyes sample.

4.2.2.2 User Study for Unambiguous Testing

A formal user study was conducted to evaluate the improved categorization for its ambiguity. In this user study, subjects were asked to classify Many Eyes patterns into the pattern categories and their classification results were compared with the classification we did in the above experiment, with the assumption that mismatching indicated ambiguity of the categorization.

Five graduate students of computer science major (3 males and 2 females) participated in the user study. Three students studied in the field of visualization and two students studied in the field of data mining. The subjects took the user study one by one on the same computer in the same office following the same process. First, a pre-test training was given. The definition of each pattern category was explained and pattern examples were given. After the training, each subject was asked to select a category from the 12 categories in our categorization for each of 60 patterns that were randomly sampled from the Many Eyes patterns one after another. The classification results and the time spent for each pattern were automatically recorded.

The classification results were compared against the classification we did in the experiment. The comparison showed that there were only 5 conflicts. Two of them were between the categories *extreme* and *rank*. Three of them were between the categories *difference* and *outliers*. Although it seemed that the category *rank* could cover *extreme* according to their definitions, we decided not to merge them since the latter is a significant category according to our previous experiment (see Table 8.1). For *difference* and *outliers*, we reduced the ambiguity by modifying the definition of *outliers* to emphasize that the difference between the sizes of the sets in comparison should be big.

The average and maximum time the subjects used to classify a pattern was 223 seconds and 360 seconds respectively. It indicated that the subjects were able to make the classification without much effort.

4.2.3 Domain Expert Interview

We conducted interviews with domain experts from a variety of research fields for the following goals: (1) to evaluate the generalized categorization using patterns sampled from specific application domains, and (2) to determine which information about patterns is essential for visual sense making in real applications.

Sixteen participants (10 male and 6 female) were interviewed, including 7 PhD students, 5 research scientists, and 4 analysts working in companies. They were working on a wide variety of research fields including neurology, biology, bioinformatics, cytology, GIS, remote sensing, financial analysis, telecom planning and designing, civil designing, economics, biology, and networking. All participants had self-identified as having experience of sense making with the help of visualization in their research. All of them analyzed multidimensional data sets in their research. Six participants claimed that their data had temporal dimensions and four claimed that their data contained geographical dimensions.

The interviews were conducted person by person in July 2008, including 9 phone interviews and 7 face-to-face interviews. Each interview took about 20 to 30 minutes, following a structured interview guide. An interview began by collecting the participant's background information such as analytic goals, data, and visualization tools. Then the participant was asked to provide specific examples of patterns collected in their analytic tasks. The participant was also asked to provide a list of attributes about the patterns that were important for their analytic tasks. Towards the end of the interview, our existing pattern categorization was explained and the participant was asked to classify his/her reported patterns into existing categories. When the participant encountered any patterns that did not fit, the patterns were placed in a list for future analysis. Extensive field notes were taken during the interview. Some participants provided screenshots to example patterns after the interview.

After the interviews, the patterns and attribute lists were analyzed. Eighty-one domain specific patterns were collected from the interviews and sixty-eight of them fitted into our categories. The thirteen patterns that did not fit into any categories fall into one of the following categories: (1) Patterns about other data structures derived from the multidimensional data, such as a pattern about the hierarchical structure derived from the multidimensional data; (2) Patterns about high level knowledge that were not directly related to the multidimensional data, such as the pattern that ”K means clustering is much better than SOM in sorting out the dynamics of data”. Since these patterns were either beyond the range of multidimensional data or about high level knowledge, we exclude them from categorization and claim that our categorization covered the majority of domain specific patterns we collected.

For the attributes in the list, we divided them into two categories:

- Content: this category includes information characterizing the content of patterns, such as sizes and averages of clusters, values of anomalies, and names of correlated dimensions.
- Context: this category includes information capturing the context of patterns. For example, the distribution of the whole data sets provides a context to an *outliers* pattern. The significance of most patterns can only be evaluated among their contexts. Quality is a special context attribute of patterns. Many participants suggested that quality information is important since it helps them index, retrieve, and filter patterns.

The above study showed that the content and context attributes are essential in discovery management. We thus decided to summarize them for each pattern category and include them into our pattern taxonomy. We conducted the following literature survey for this purpose.

4.2.4 Literature Survey for Summarizing Pattern Attributes

A literature survey has been conducted on statistics and data mining textbooks [74, 75, 76, 77] to learn what information should be captured as content and context attributes for different categories of patterns. The attribute lists collected from the domain expert interviews were also referenced. The essential content and context attributes for each category are listed.

4.3 Resulting Pattern Taxonomy

The constructed pattern taxonomy, which includes the categorization, formal definition, examples, content attributes, and context attributes, is presented in Table 4.2. In this table, X indicates a set of all elements. For a pattern in the item space, it refers to the set of all items. For a pattern in the dimension space, it refers to the set of all dimensions. D indicates a set of all attributes. For a pattern in the item space, it refers to the set of all dimensions. For a pattern in the dimension space, it is the set of all items. V indicates values of elements on their attributes. X_i indicates a subset of X. D_j indicates a subset of D. x_i indicates a element in X. d_j indicates a element in D. f indicates distance calculation function. ∂ indicates user defined constant. Table 4.2 also shows tasks related to each pattern category for users' reference.

4.4 Conclusion

In this chapter, we presented the construction of pattern taxonomy for multidimensional data. The taxonomy is the basis of the discovery management framework and enables significant automaticity in different management activities such as pattern annotation, retrieval, organization, association. In the future, we will construct pattern taxonomies for other data types such as trees and graphs and extend the discovery management framework to those data types. We will also explore more pattern categories that are essential for specific application domains, such as network analysis and bioinformatics. We will compare the domain specific discovery categories with the general taxonomy presented in this chapter, which helps us to understand users information needs and extend the discovery management framework to fulfill the needs.

Table 4.2: A pattern taxonomy for multidimensional data.

Category	**Formal definition**	**Examples**	**Content**	**Context**	**Related Task**
Derived value	*Derived value* is defined on a 3-tuple (X_i, d_n, R) where R is a derived value of X_i on d_n. When X_i contains only 1 element, R is the value of the element on d_n.	The average salary of graduated students in laws school is 60k per year.	X_i; d_n; R (calculated using distributive, algebraic, holistic, mathematic function, or other functions).	N/A	Retrieve value [17], Compute derived value [17].
Distribution	*Distribution* is defined on a 3-tuple (X_i, D_j, R). R describes the distribution of $V_{D_j}(X_i)$.	The distribution of consumption month by month in Italy is fairly even.	X_i; D_j; R (density description such as skewed, clumpy, sparse, and striated [78]; shape description such as convex, skinny, and stringy [78]).	N/A	Characterize distribution [17, 46]

Table 4.2 (continued)

Category	Formal definition	Examples	Content	Context	Related Task
Difference	*Difference* is defined on a 4-tuple $(X_i,\ D_j,\ f,\ \partial)$. $\forall\ x_m,\ x_n \in X_i,\ f_{Dj}(x_m,\ x_n) \geq \partial$ where f calculates the distance between x_m and x_n on D_j.	In USC, there is still a greater absolute enrollment in the social sciences than the biological sciences.	X_i; D_j; f; ∂.	Statistical distribution of X on D_j; Distances between elements $\in X_i$ and other elements.	Distinguish [47, 46]
Extreme	*Extreme* is defined on a 3-tuple $(x_m,\ X_i,\ d_n)$. $\forall\ x_l \in X_i$ and $x_l \neq x_m$, $V_{dn}(x_m) \geq V_{dn}(x_l)$ or $\forall\ x_l \in X_i$ and $x_l \neq x_m$, $V_{dn}(x_m) \leq V_{dn}(x_l)$.	The lowest average salary of a department is 92k for the Romance Languages and Literature Department in university.	x_m; d_n; $V_{dn}(x_m)$.	Statistical distribution of X on d_n.	Find extreme [17]
Rank	*Rank* is defined on a 4-tuple $(x_m,\ X_i,\ d_n,\ R)$. R is the order of $V_{dn}(x_m)$ in sorted $V_{dn}(X_i)$.	Between 1970 and 1971, Human resources budget surpassed National Defense to be the No.2 budget category.	x_m; X_i; d_n; $V_{dn}(x_m)$; R.	N/A	Ranking [46, 47], Sort [17]

Table 4.2 (continued)

Category	**Formal definition**	**Examples**	**Content**	**Context**	**Related Task**
Categories	*Categories* is defined on a 3-tuple $(X_i,\ D_j,\ C_k)$. C_k is a set of categories. Elements in X_i are classified into the categories in C_k based on their values on D_j.	All in all, jobs in this data can be classified into 4 categories: rich, middle, lower middle and lower.	X_i; D_j; C_k.	N/A	Categorization [47]
Cluster	*Cluster* is defined on a 4-tuple $(X_i,\ D_j,\ f,\ \partial)$. $\forall\ x_m,\ x_n \in X_i,\ f_{D_j}(x_m,\ x_n) \leq \partial$, where f calculates the dissimilarity between x_m and x_n on D_j.	Countries in Western Europe tend to group together according to their consumption amounts in 1999.	X_i; D_j; ∂; statistics of $V_{D_j}(X_i)$ such as average values, minimum values, and maximum values.	Statistical distribution of X on D_j; dissimilarity between this cluster and other clusters; quality measures such as *recall* that measures the proportion of the relevant elements in the cluster and *precision* that measures the fraction of elements in the cluster that are actually relevant.	Clustering [47, 46, 17]

Table 4.2 (continued)

Category	**Formal definition**	**Examples**	**Content**	**Context**	**Related Task**
Outliers	*Outliers* are defined on a 3-tuple (X_i, D_j, R) where R is a considerable dissimilarity, exception or inconsistency of $V_{Dj}(Xi)$ with respect to the remaining elements.	Uganda's consumption is high given the relatively low consumption of its neighbors.	X_i; D_j; $V_{D_j}(X_i)$; R.	Statistical distribution, distances, density differences, or deviation differences between outliers and other elements according to the outlier analysis approach used.	Find anomalies [17]
Association	*Association* is defined on a 3-tuple (X_i, D_j, R). R is the relationship among elements in X_i on D_j.	In US, there is a negative correlation between income and obesity when income is less than 50k.	X_i; D_j; R.	The support and confidence of association [75]; quality measures such as correlation coefficient [77] for the continuous scale data, or statistics chi square test [79] for categorical data.	Associate [46, 47], Correlate [17]

Table 4.2 (continued)

Category	**Formal definition**	**Examples**	**Content**	**Context**	**Related Task**
Trend	*Trend* is defined on a 6-tuple (X_i, D_j, T, $t1$, $t2$, R). R describes the movement feature of $V_{D_j}(X_i)$ on T in the segment defined by $t1$ and $t2$. T is usually a temporal attribute.	Veterans' benefits are going down over the past ten years.	X_i; D_j; T; $t1$; $t2$; R (rise/fall/stable, cyclic, seasonal, or irregular movements, slope, or shapes described by formal language [75]).	Globe trend; trend of other attributes in the same segment.	N/A
Meta	*Meta fact* is a fact about data itself, such as missing dimensions or values, data qualities, meanings of labels, and etc..	A change happened between 1999 and 2000 when a bunch of new categories showed up.	Determined by users.	Determined by users.	N/A
Compound patterns	*Compound fact* is a fact that contains two or more facts.	It's interesting how different the second letter distribution is from the first letter distribution.	Split into other types of facts and then analyze.	Split into other types of facts and then analyze.	Compound tasks [17]

CHAPTER 5: SEMI-AUTOMATIC DISCOVERY ANNOTATION

During a complex visual analytics process, users need to annotate their discoveries for reusing them, organizing them, or presenting them to others [51]. Often, automated annotation techniques are desired to reduce human beings' efforts for annotating discoveries. Most existing discovery annotation approaches achieve the annotation automation through an automatic record of users' interaction events (e.g., clicks and key presses) or analysis actions (e.g., panning and zooming). In this chapter, we present a novel automated discovery annotation approach, named Click2Annotate. It allows semi-automatic discovery annotation that captures low-level analytics task results (e.g., clusters and outliers), which have higher semantic richness and abstraction levels than actions and interaction events. Therefore, Click2Annotate reduces human effort required in annotation and generates annotations easy to understand. We also present a formal user study to prove this benefit.

5.1 Introduction

During a complex visual analytics process, capturing discoveries from data and using them as evidence for the hypothesis generation and evaluation are important steps for decision making and problem solving. Since a visual analytics process may involve a large number of discoveries, discovery annotation, namely the process of capturing and recording the semantics of discoveries [51], is important for discovery revisiting, association, comparison, and exchange.

Discovery annotation in most existing visual analytics systems, such as Many Eyes [13] and Name Voyagers [54], requires users to type notes, draw marks, or connect associated discoveries manually. When the number of discoveries grows larger, these manual approaches become tedious, inefficient, inaccurate, and time consuming

[43]. To address these problems, initial efforts have been made towards automatically annotating discoveries.

Existing automated approaches can be classified according to the four-tier visual analytic activity model proposed by Gotz and Zhou [43]. In this model, visual analytic activities are abstracted into four levels namely tasks, sub-tasks, actions, and events. They range in semantic richness and abstraction levels from high to low. *Tasks* correspond to a user's highest-level analytic goals. *Sub-tasks* correspond to more objective, concrete analytic goals, such as finding clusters, outliers, or correlations. They are also called low level analytic tasks in other literatures [17]. *Actions* refer to atomic analytic steps such as zooming and panning. *Events* correspond to the lowest-level of interaction events, such as mouse clicks and button presses. Automation in most existing discovery annotation approaches is conducted at the action level or event level. To the best of our knowledge, there exists no general approach that conducts the automation at the sub-task level.

We argue that conducting automated discovery annotation at the sub-task level is a promising research direction. The reasons are:

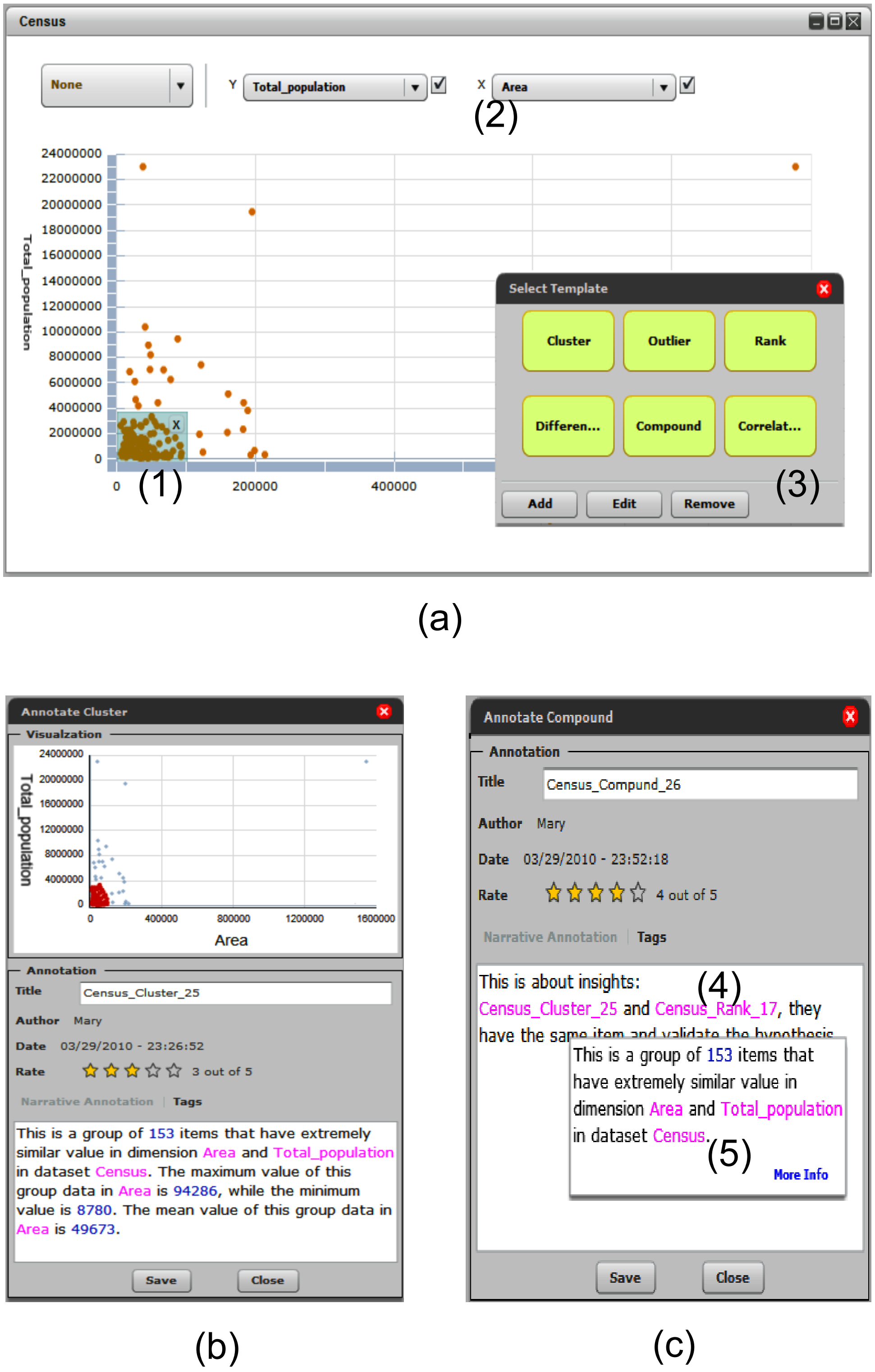
- Sub-tasks are less application-dependent than tasks. For example, according to Amar and Stasko [17], there exists a set of low-level analytic tasks (sub-tasks) that are common to most multidimensional datasets. Therefore, it is possible to develop automated annotation approach independent from particular domains and applications at the sub-task level.
- Information captured from the sub-task level, such as clusters and outliers for multidimensional datasets, can have higher semantic richness and abstraction levels than that from the action and event levels, such as zooming and mouse clicks. The former will be easier to understand, recall, retrieve, and reuse in the visual analytics process than the latter.
- Annotations with information from the sub-task level can be decoupled from the

low level user exploration behaviors. For example, we can annotate a discovery as a cluster without recording how this cluster is found. As a consequence, the annotations are independent from the visualization platforms on which the discoveries are captured. Thus the share and exchange of discoveries among different visualization systems can be enabled. In addition, the implementation of the discovery management approach can be made simpler by not capturing the exploration process. For example, the storage of the generated annotations can also be more efficient without the exploration process captured.

In this chapter, we propose a novel semi-automatic discovery annotation approach, Click2Annotate, which conducts the automation at the sub-task level. Guided by the proposed pattern taxonomy (Chapter 4), Click2Annotate can semi-automatically annotate patterns in a formalized format by predicting what should be recorded for facts in popular categories, and allow users to annotate knowledge bases and evaluations with light human effort. Click2Annotate has been integrated in ManyInsights to annotate discoveries from different multidimensional visualizations, such as scatter plot and parallel coordinates. The user evaluation of ManyInsights proved that Click2Annotate could enhance annotation efficiency and the annotations generated could be easy to understand. Besides, the semantic-rich information automatically captured at the sub-task level by Click2Annotate facilitates different discovery management techniques in ManyInsights.

5.2 Approach

Based on the proposed discovery model, Click2Annotate enhances the automation in **pattern** annotation, and allows users to annotate knowledge bases and evaluations with light human efforts. The automation of Click2Annotate in pattern annotation is based on the following observations reported from our user experiments and expert interviews (see Chapter 4). First, most patterns extracted from multidimensional data fall into multiple categories independent from the domains/applications and



(a)

Title: Census_Cluster_25

Author: Mary

Date: 03/29/2010 - 23:26:52

Rate: ★★★☆☆ 3 out of 5

This is a group of 153 items that have extremely similar value in dimension Area and Total_population in dataset Census. The maximum value of this group data in Area is 94286, while the minimum value is 8780. The mean value of this group data in Area is 49673.

(b)

Title: Census_Compund_26

Author: Mary

Date: 03/29/2010 - 23:52:18

Rate: ★★★★☆ 4 out of 5

This is about insights: Census_Cluster_25 and Census_Rank_17, they have the same item and validate the hypothesis

This is a group of 153 items that have extremely similar value in dimension Area and Total_population in dataset Census.

(c)

Figure 5.1: Semi-automatic annotation generation using Click2Annotate. (a) A scatterplot with a cluster in it and the annotation process. (b) The automatically generated annotation for the cluster. (c) The annotation generated for a compound pattern.

visualization tools. Second, users can effectively and efficiently classify patterns into these categories. Third, the same set of context and content information is often used to annotate patterns falling into the same category.

According to the above observations, the core components of Click2Annotate are annotation templates, each of which guides the semi-automatic annotation of a certain type of patterns. They are either pre-defined for popular pattern types or interactively created by users. During the annotation process, the users only need to highlight data composing a pattern, to decide the type of the pattern, and to select the corresponding annotation template. The system will automatically follow the template to fetch information, to encode it, and to generate a narrative annotation for the discovery.

We briefly describe how users use Click2Annotate to annotate an discovery in a scatter plot view. When a user discovers a pattern of interest during the visual exploration, such as the cluster shown in Figure 5.1(a), she brushes the relevant data (see Figure 5.1(a-1)), specifies the dimensions (see Figure 5.1(a-2)), judges the type of the pattern, and selects the template for this type by a mouse click (see Figure 5.1(a-3)). The system will then automatically create an annotation based on the template and present the annotation to the user (see Figure 5.1(b)). The user reviews the annotation and interactively improves it, such as typing domain-related information and her evaluations (this step can be customized for individual applications to increase the level of automation). Since mouse clicks rather than intensive typing effort are required from the user to accomplish the majority of the annotation process, our approach is named Click2Annotate.

5.3 Annotation Templates

Annotation templates are the key components of Click2Annotate. Each annotation template is associated with a pattern type. It tells the system what information needs to be retrieved from the data and how to generate a semantic-rich annotation for this type of pattern. A template can be either pre-defined or user-defined.

5.3.1 Pre-Defined Templates

Click2Annotate provides pre-defined templates for popular pattern types detected from the taxonomy presented in Chapter 4. The templates are generated with the following steps. Determining Popular Pattern Types: Six pattern types, namely *cluster*, *outlier*, *rank*, *difference*, *correlation*, and *compound pattern*, are determined to be popular pattern types. Patterns of these types were frequently posted on Many Eyes [13] as revealed by our experiments in constructing the taxonomy. Their definitions are self-explained by the type names. These pattern types are further classified into three categories, namely dimension-oriented patterns, data item-oriented patterns, and compound patterns. Pattern types within the same category share common features. Dimension-oriented patterns, such as *correlation*, describe relationships of dimensions. Data item-oriented pattern, such as *cluster*, *outlier*, *rank*, and *difference*, describe clusters, anomalies, patterns, and relationships of data items. Compound pattern describe relationships among multiple pattern, such as that the pattern A is related to the pattern B. The type hierarchy guides the generation of the templates by extracting the common features among the types in the same category. It also guides the use of annotations in other discovery management activities.

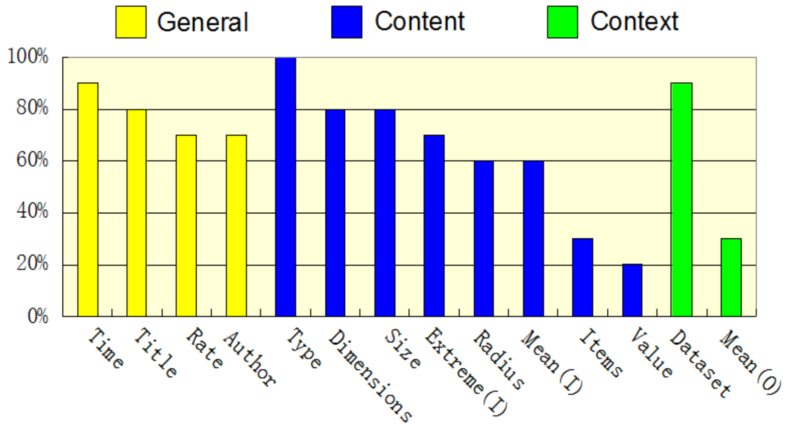


Figure 5.2: The frequencies of attributes used in cluster annotations.

Table 5.1: Content attributes, context attributes, and narrative sentences for popular pattern types. The attributes with "I" are about the data inside the pattern. The attributes with "O" are about the data outside the pattern.

Type	**Content Attributes**	**Context Attributes**	**Narrative Sentences**
Cluster	Type, Dimensions, Size, Extreme(I), Radius, Mean(I)	Dataset	This is a group of xx items that have (extremely/very/slightly) similar values in dimensions xx in dataset xx.
Outlier	Type, Dimensions, Size, Items, Mean(I)	Dataset, Mean(O), Distance	This is a group of xx items that are (extremely/very/slightly) different from the others in dimensions xx.
Rank	Type, Dimensions, Items, Value, Rank	Dataset	Item xx ranks xx in dimension xx in dataset xx from highest to lowest.
Difference	Type, Dimensions Items, Difference, Distance	Dataset	There is an/a (extremely large/large/slightly) difference between item xx and item xx in dimension xx. The value of item xx is higher by xx.
Correlation	Type, Dimensions Coefficient	Dataset	For xx percent of data items in dataset xx, the higher their value in dimension xx, the (higher/lower) their values in dimension xx.
Compound	Pointers to the related patterns	N/A	This is about the pattern xx.

Predicting Information in Annotations: The templates tell the system what information needs to be retrieved from the data for generating an annotation. We predict such information for each pre-defined template based on the results of our domain expert interviews from the taxonomy construction process. In these interviews, the experts reported what information they used to annotate the patterns. We summarized the results and got an attribute list for each popular pattern type. The percentage of the experts that used an attribute to annotate each type of patterns was calculated. For example, Figure 5.2 shows how often the attributes listed were used to annotate a cluster by the experts.

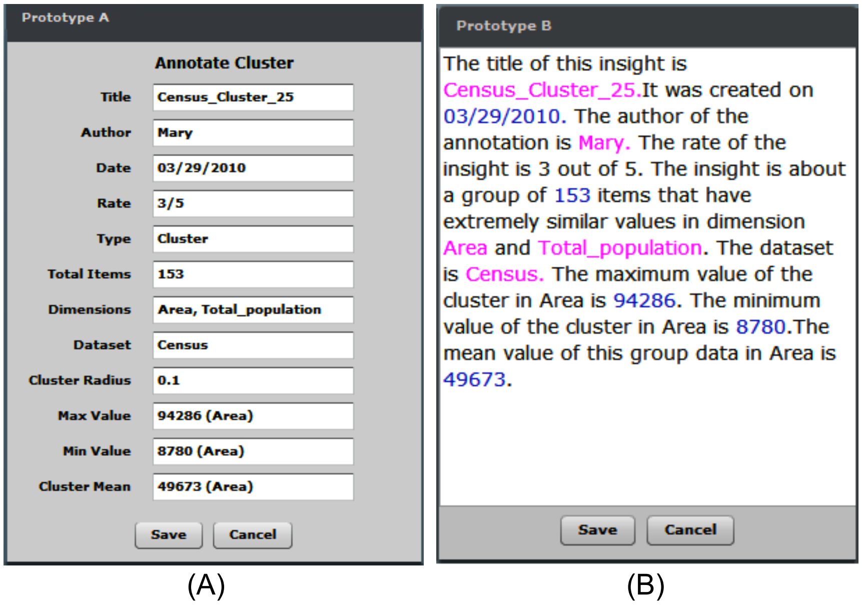
As shown in Figure 5.2, an annotation of a cluster often consists of the following attributes, in descending order of their frequency: Type (the type of the pattern, such as "cluster" in this example); Time (when the cluster was discovered); Dataset (the dataset where the cluster was discovered); Title (the title of the annotation); Dimensions (the dimension names of the subspace where the cluster existed); Size (the number of items in the cluster); Rate (users' subjective evaluation); Author (who discovered the cluster); Extreme(I) (the extreme of data inside the cluster); Radius (the radius of the cluster); Mean(I) (the mean of data inside the cluster); Items (the data item names of the cluster); Value (the data values of the cluster); and Mean(O) (the mean of data outside the cluster).

According to the statistics, we identify three categories of attributes for each popular pattern type (shown in yellow, blue, and green in Figure 5.2). They include: general attributes, such as Author, Time, Title, and Rate, which are important information for all types of patterns and they are not directly related to the data; context attributes, such as Size, Items, and Mean(I), which are frequently used to describe the content of a certain type of discoveries; and context attributes, such as Mean(O), which are frequently used to capture the context of a certain type of patterns.

Table 5.1 summarizes the frequently used content and context attributes (with percentage $\geq 50\%$) of the popular pattern types. They are semantic-rich information widely used by the experts to describe patterns. We include them into the templates of the types together with all the general attributes. Users can customize a template (refer to Section 5.3.2) if the information they desire is not included in the template.

Shaping the Templates: From the previous step, a fairly large amount of attributes are determined to be included into the templates. How should they be presented to users so that the users can enjoy reading the annotations and grasp their content effectively and efficiently? To address this problem, we conducted a user study.

First, we designed three template interface prototypes with the following goals: (1) Completeness: all attributes should be represented; (2) Clearness: the information should be easy to read and understand; and (3) Briefness: key information of the attributes should be easily accessed. In Prototype A, each attribute was represented as a form entry, such as “cluster radius: 0.1” for the radius of a cluster. Prototype B employed a narrative annotation that represents information textually [80]. All the attributes were presented in sentences that described the attributes using natural



Annotate Cluster

Title	Census_Cluster_25
Author	Mary
Date	03/29/2010
Rate	3/5
Type	Cluster
Total Items	153
Dimensions	Area, Total_population
Dataset	Census
Cluster Radius	0.1
Max Value	94286 (Area)
Min Value	8780 (Area)
Cluster Mean	49673 (Area)

The title of this insight is Census_Cluster_25.It was created on 03/29/2010. The author of the annotation is Mary. The rate of the insight is 3 out of 5. The insight is about a group of 153 items that have extremely similar values in dimension Area and Total_population. The dataset is Census. The maximum value of the cluster in Area is 94286. The minimum value of the cluster in Area is 8780.The mean value of this group data in Area is 49673.

Figure 5.3: Examples of prototypes A and B.

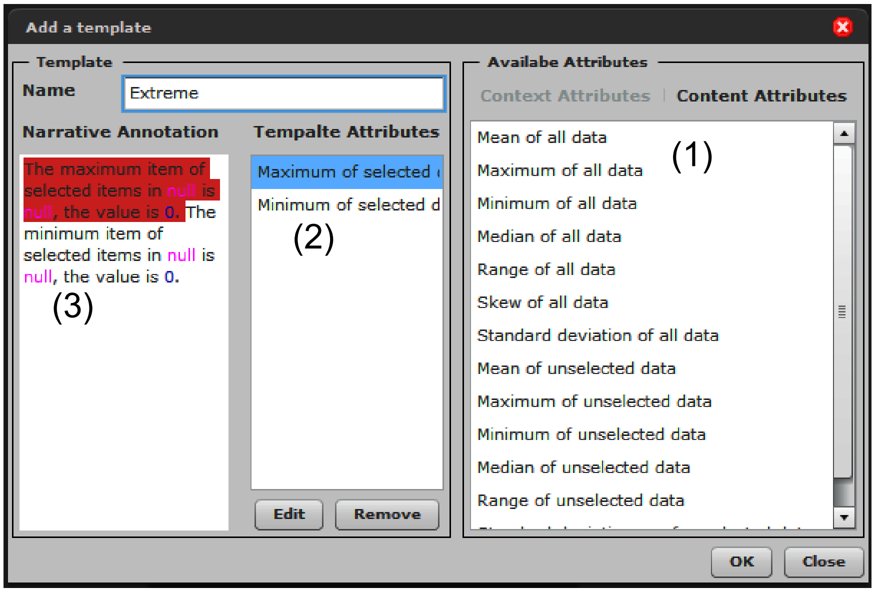
language. For example, the entry "cluster radius: 0.1" in the previous example was expressed as "The items in this group have very similar values". Prototype C used a mixed design. The general attributes were represented as form entries while the content and context attributes were represented textually. Two annotations were generated for each of the pattern types, including cluster, outlier, and correlation following each prototype. A total of 18 annotations were generated. Figure 5.3 shows examples of the generated cluster annotations for prototype A and B. The annotation of the same content for prototype C is shown in Figure 1(b).

Twenty users who had good experiences with reading annotations in visualizations participated in the study one by one. The subject was asked to grade prototypes A, B, and C according to the following three criteria: (1) the annotations are pleasant to read; (2) the values of the attributes in the annotations can be quickly perceived; and (3) it is easy to compare patterns of the same type and pattern of different types. A 7-point scale was used for the rating (0=strongly disagree, 6=strongly agree). User feedbacks were also collected.

The results showed a stronger preference to Prototype C. In particular, the average scores of Prototype A, B, and C in the first criterion were 2.8, 3.4, and 4.8, respectively; the average scores in the second criterion were 3.8, 3.2, and 4.4, respectively; and the average scores in the third criterion were 4.4, 3.0, and 4.2, respectively.

According to user feedback, Prototype C had the following advantages: First, it represented the general attributes as form entries and thus reduced the number of sentences in the narrative annotation. Users had no difficulty in understanding general attributes, such as author and dataset name, in the form entries. Second, it represented context and content attributes using natural language, which makes them easy to understand by users who were not familiar with terms such as cluster radius. Therefore, Prototype C is used in Click2Annotate for shaping the templates.

Encoding Attributes Using Natural Language: The context and content attributes are encoded into human-readable sentences in the templates to compose narrative annotations. Our encoding process is similar to the one described in [80] but improved from three aspects: First, multiple context and content attributes can be encoded in one sentence. This produces a less wordy annotation. Second, the numerical attribute values (e.g., the radius value) are explained in an easy to understand manner (e.g., "very similar"). Third, the key information in the sentences, such as dimension names, is automatically highlighted and hyperlinked so that users easily browser and retrieve related discoveries sharing the common content in the search interface (see Chapter 6). For example, the narrative of a cluster may start by a sentence that describes the size, the quality (indicated by radius), the dimension labels, and the dataset of the cluster: **this is a group of 8 data items that have extremely similar values in dimensions A and B in dataset NFL.** The information in pink is automatically extracted by the system after a user selected the data and the template. We summarize examples of narrative sentences in Table 5.1.



Name: Extreme

The maximum item of selected items in null is null, the value is 0. The minimum item of selected items in null is null, the value is 0.

Maximum of selected
Minimum of selected d

Mean of all data
Maximum of all data
Minimum of all data
Median of all data
Range of all data
Skew of all data
Standard deviation of all data
Mean of unselected data
Maximum of unselected data
Minimum of unselected data
Median of unselected data
Range of unselected data

Figure 5.4: Interactive generation of a user-defined template.

5.3.2 User-Defined Templates

Although a set of templates is pre-defined for most popular pattern types, it is impossible to predict all useful pattern types as well as all possible attributes for each pattern type. Therefore, Click2Annotate allows users to interactively modify pre-defined templates or create new templates from scratch.

Figure 5.4 shows an example of how to create a new template for a user-defined pattern type named *extreme*. In this window, there a lists of available attributes (see Figure 5.4(1)), including all possible context and content attributes reported by the domain experts in the interviews from the taxonomy construction. They can be added to the template attributes list (see Figure 5.4(2)). In this example, the general attributes are automatically included and the maximum and minimum of the relevant dimension are manually added into the template. The narrative sentences of these attributes are represented in the annotation area, providing a preview for the annotations generated by this template (see Figure 5.4(3)). Users can interactively modify these sentences or change their order. The modification of an existing template can be accomplished in the same interface.

5.4 Semi-Automatic Annotation Generation

Click2Annotate semi-automatically generates annotations based on pre-defined or user-defined templates. To generate an annotation, users brush the relevant data items and dimensions and select a template according to the type of the pattern. To allow quick access to the templates, a list of buttons are provided in a separated panel (see Figure 5.1(a-3)), which is shared by all created views. Each button corresponds to a template. Users can add or remove buttons from the panel so that it only contains the buttons for templates they need. The users click on a button to select a template. After the template is selected, the system will automatically fetch information from the data and encode it to fill the incomplete information in the template. Thus, an annotation is automatically generated.

The above process does not apply to compound patterns because they contain pointers to other patterns. To annotate a compound pattern, an interactive approach is employed. In particular, users first open a compound pattern annotation dialog (see Figure 5.1(c)) by clicking on a "compound" button and then use drag-and-drop interactions to add the flags of desired discoveries to the dialog. After a discovery is added, its title will be displayed in the dialog (see Figure 5.1(c-4)). It is hyperlinked to the related annotation so that users can click on it to examine the annotation in detail (see Figure 5.1(c-5)). Users can add discoveries into the dialog and type their notes to complete the annotation.

5.5 Annotation Review and Modification

After an annotation is generated by the system, it will be presented to users in an annotation window (see Figure 5.5(a)) within which the annotation can be reviewed and improved by the users.

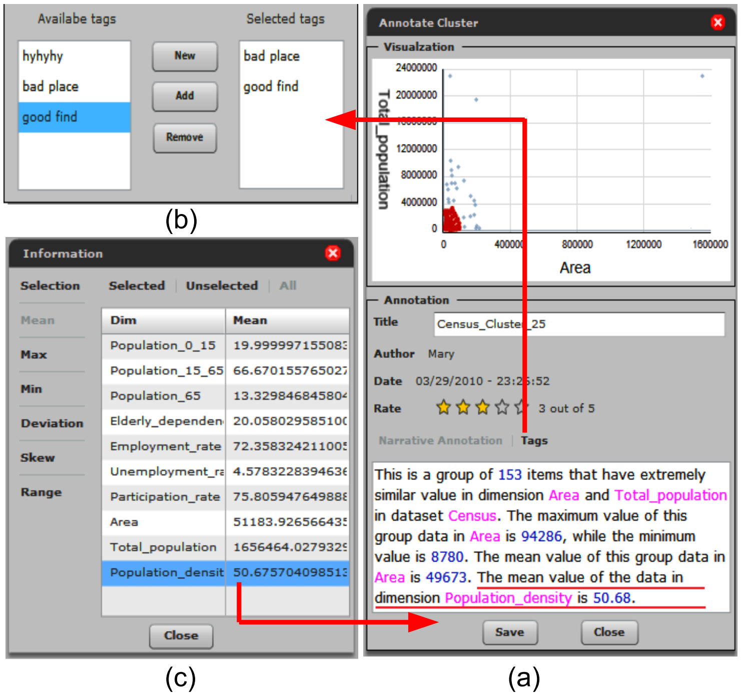
The annotation window directly mimics the design of Prototype C with a thumbnail added. The thumbnail is a screenshot that captures the visualization at the moment when the pattern was discovered to help users recall this discovery. The general attributes are represented below the thumbnail, followed by a set of sentences that textually represent the context and content attributes.

If users are not satisfied with the automatically generated annotation, they can interactively improve it. In particular, the users can open a statistics window (see Figure 5.5(c)) which presents a list of all available statistics about the pattern and the whole dataset, and a list of the information that has already been included in the current annotation. Users can use drag-and-drop interactions to add or remove statistics into or from the annotation and adjust the order of their presentations in the annotation. The statistics in the annotation is represented textually according to pre-defined templates. Users can manually customize the text representations if they are not satisfied with the pre-defined ones. For example, in Figure 5.5(c), a user drags and drops the mean value of the dimension *population density* to the annotation. A new sentence that conveys this mean value is then automatically added to the annotation, as shown in the sentence with the red underline in Figure 5.5(a).

The automatically generated annotation only captures the pattern of a discovery. To allow users to record the knowledge base and subjective evaluations of the discovery, an interactive tagging function is supported. In particular, a user can click on a button in the annotation window to trigger a tagging interface (see Figure 5.5(b)). Through the interface, the user can create tags or select existing tags to annotate the discovery. A tag is generated once and reused later on. Thus users can type frequently used information once, save it as a tag, and reuse the tag in the future with light human effort.

5.6 User Study of Click2Annotate

A formal user study has been conducted to evaluate how Click2Annotate helped users to generate annotations and if the generated annotations were understandable. The study was a 2×2 (system types×datasets) between-subjects design. We com-



Availabe tags: hyhyhy, bad place, good find

Selected tags: bad place, good find

Dim	Mean
Population_0_15	19.99999715508:
Population_15_65	66.67015576502?
Population_65	13.32984684580<
Elderly_dependen	20.05802958510(
Employment_rate	72.35832421100!
Unemployment_ra	4.578322839463(
Participation_rate	75.80594764988{
Area	51183.92656643!
Total_population	1656464.027932!
Population_densit	50.67570409851:

Title Census_Cluster_25

Author Mary

Date 03/29/2010 - 23:25:52

Rate ★★★☆☆ 3 out of 5

This is a group of 153 items that have extremely similar value in dimension Area and Total_population in dataset Census. The maximum value of this group data in Area is 94286, while the minimum value is 8780. The mean value of this group data in Area is 49673. The mean value of the data in dimension Population_density is 50.68.

Figure 5.5: The review, tagging, and modification of generated annotations. (a) A modified annotation. (b) The tagging interface. (c) The statistics window.

pared two systems: ManyInsights, which supported Click2Annotate, and a simple system, which provided users a text editor similar to those commonly found in many visualization systems to manually type notes for annotating discoveries. Our hypotheses were: (1) Click2Annotate will reduce the time spent on annotating; and (2) annotations generated by Click2Annotate will reduce the time-cost and errors for understanding the annotations.

5.6.1 Datasets and Discoveries

Two datasets were used in the user study: a small dataset (51 items, 4 dimensions) on state health measures and a large dataset (279 items, 10 dimensions) on the US census data. Before the user study, we manually extracted six discoveries, including a cluster, an outlier, a rank, a difference, a correlation, and a compound pattern, from each dataset. The compound pattern was about the difference between two clusters. All extracted discoveries were used in the user tasks described in the next section.

The numbers of data items and dimensions involved in the discoveries were controlled according to the size of dataset. For example, a cluster in the large dataset had more data items and dimensions involved. This made annotating and comprehending discoveries in the large dataset more difficult.

5.6.2 Tasks

The experiment included two sessions.

Annotation session: Each participant was asked to annotate the six discoveries for each dataset on a computer. The discovery was annotated one by one. For each discovery, the type was explicated and the relevant data was highlighted in a parallel coordinates or scatterplot view of the dataset according to the number of dimensions involved in the discovery. The participant was asked to record all possible information that could help them comprehend the discovery. The task completion time was recorded.

Comprehension session: Each participant was asked to understand the annotations generated by other participants. There were six tasks to complete for each dataset, each of which for a discovery used in the annotation session. In each task, an annotation (randomly picked from annotations generated by other participants and text only) was provided along with four images on paper. One image was the screenshot of the view with the discovery described by the annotation highlighted, namely the original view provided to the participants when the discovery was annotated. The other three images presented different views, such as the same display with other data highlighted or a different display with a similar pattern. The participant was asked to find the view with the discovery described by the highlighted annotation. The task completion time was recorded.

5.6.3 Analysis Condition and Procedure

A total of 8 subjects (5 male and 3 female) participated in the study. All of them were graduate students and had strong English writing and reading abilities. Before the study, the subjects were evenly divided into two groups. One group of subjects used ManyInsights and the other group used the simple system. The same datasets and tasks were used in both groups. The subjects took the experiment one by one on the same computer following the same process.

In the annotation session, the same views were used in both groups. When making annotations, participants were allowed to read dimension names, data names, and data values on the visualizations. In the simple system, participants used the text editor to type notes for annotating discoveries. In ManyInsights, participants were allowed to edit existing templates and interactively modify generated annotations.

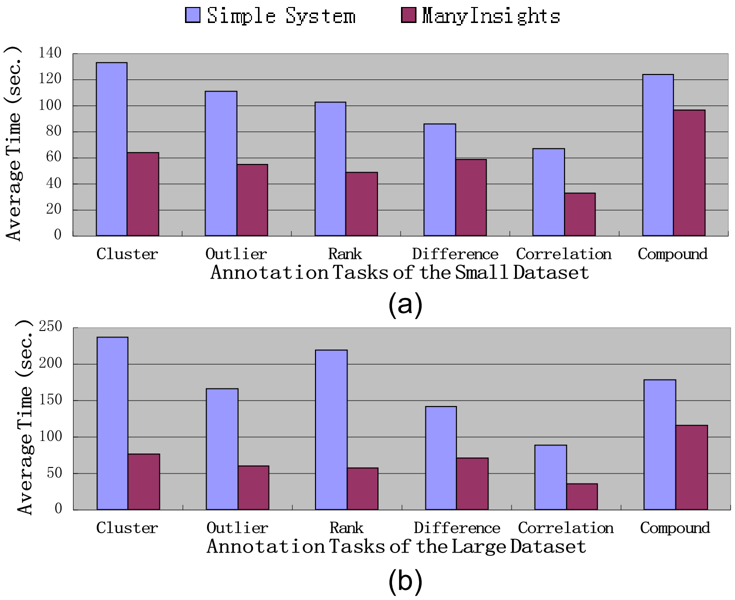
At the beginning of the study, a tutorial was provided by an instructor to explain the definition of each pattern type and show examples of how to annotate a discovery. The annotation session was conducted right after the tutorial. The comprehension session was conducted three months after the annotation session. In each session, there were first practical tasks, second experimental tasks (the small dataset followed by the large dataset), and then survey questions specific to that session.

5.6.4 Results

We present two types of results from the study, namely quantitative data (completion time and correctness) captured through the system and the subjective preferences reported from survey questions, in the following sections respectively.

5.6.4.1 Task Completion Time and Correctness

The comparisons of the average completion time for annotating discoveries are shown in Figure 5.6(a) (for the small dataset) and 5.6(b) (for the large dataset). Figure 5.6(b) reveals the difficulty the subjects encountered in making annotations in the large dataset using the simple system, especially when annotating the cluster and rank. We observed that subjects had difficulty in manually summarizing information from complex data, such as estimating the size of a big cluster and the rank of a data item in a large dataset. Besides, in the simple system, the time the participants spent on determining the information to be recorded was often more than the time they spent on typing the note. Click2Annotate showed its strength in pre-defining the most essential information for discoveries and automatically capturing this information. Therefore, our first hypothesis was validated.



(a)

(b)

Figure 5.6: The result of annotation sessions.

The comparisons of the average completion time for the comprehension tasks are shown in Figure 5.7(a) (for the small dataset) and 5.7(b) (for the large dataset). The figures show that the participants were faster in selecting the views when reading the annotations generated by Click2Annotate, especially for the large dataset. The average correct answer rate for all tasks was 89.6% for ManyInsights (with standard deviation 7%) and only 75.0% for the simple system (with standard deviation 11%). The result suggested that the subjects understood the Click2Annotate annotations faster and better than the manually generated annotations. Thus our second hypothesis was validated.

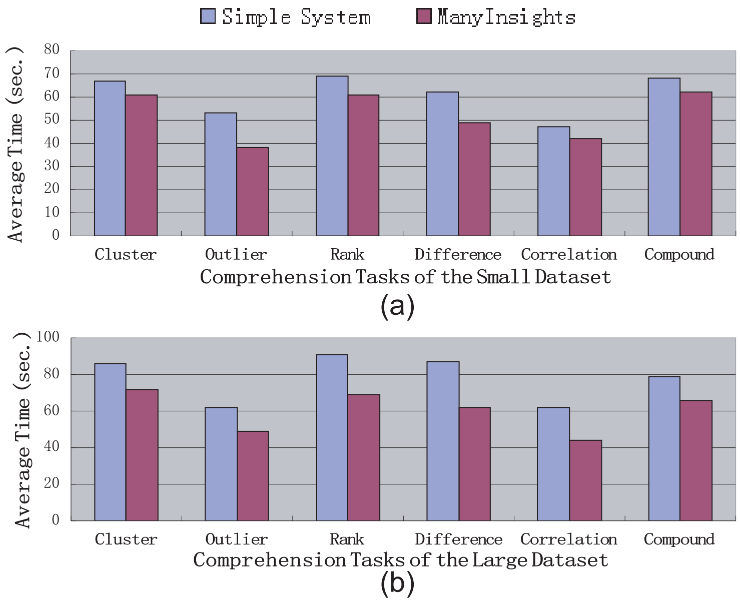


Figure 5.7: The result of comprehension sessions.

5.6.4.2 Subjective Preferences

At the end of each session, the participant was asked to answer a set of survey questions each of which was answered in a 7-point Likert scale (0=strongly disagree, 6=strongly agree). A total of 10 questions were provided. The average score for ManyInsights was 4.8 and only 2.7 for the simple system. Table 2 summarizes the pair-wise comparisons of the questions where significant differences were detected.

The significant differences indicate that Click2Annotate was judged to be more

Table 5.2: The average ratings for four survey questions that have significant differences (difference≥2.5).

Questions	**ManyInsights**	**Simple system**
Q1. This tool helped me make annotations in the large dataset.	4.8	1.3
Q2. I enjoy using this tool to make annotations.	5.0	2.3
Q3. The content of annotations is accurate.	5.3	2.3
Q4. The annotations were helpful to me in understanding discoveries.	5.0	2.5

helpful than the simple system in annotating discoveries. Annotations generated by Click2Annotate were judged to be more helpful in understanding discoveries.

5.7 Conclusion

Discovery annotation is a critical requirement for an effective decision making and problem solving process. In this chapter, we introduced a novel approach that allows users to conduct semi-automatic discovery annotation at sub-task level. We also presented a fully working prototype of the approach named Click2Annotate. The approach has two significant benefits. First, it reduces human effort and generates annotations easy to understand. Second, the rich semantic information encoded in the annotations enables various discovery management activities, such as discovery browsing and retrieval. We presented a formal user study that proved the first benefit. We will illustrate the second benefit by presenting the novel discovery management approaches based on Click2Annotate, namely scented discovery browsing and faceted discovery search, in the next chapter.

The future work includes a semi-automatic discovery tagging approaches to increase annotation efficiency. In such an approach, a user interactively creates a tag with criterion and then the system automatically assigns the tag to all discoveries meeting the criterion. Also, we will encourage users to tag discoveries by providing tag suggestions. Different approaches to generating the suggested tag lists, such as using dimension names, nominal values, statistic information, relevance to other discoveries, as well as the exploration task being executed, will be explored and compared. To support a wider range of data types, we will also extend Click2Annotate to trees, graphs, text, and geospatial data. In addition, more user studies and experiments will be conducted to investigate the advantages and limitations of Click2Annotate. For example, we will compare the performance of users who are familiar/unfamiliar with the datasets visualized. We will also investigate the effectiveness and efficiency of Click2Annotate with different design options in real analytical reasoning processes.

CHAPTER 6: DISCOVERY BROWSING AND RETRIEVAL

In Chapter 5, we presented a semi-automatic discovery annotation approach which is capable to capture the rich context information of discoveries in visual analytics. The information is potentially useful for enhancing the automation and efficiency of different discovery management activities. In this chapter, we present two concrete examples that utilize the rich context of discoveries to facilitate discovery browsing and retrieval.

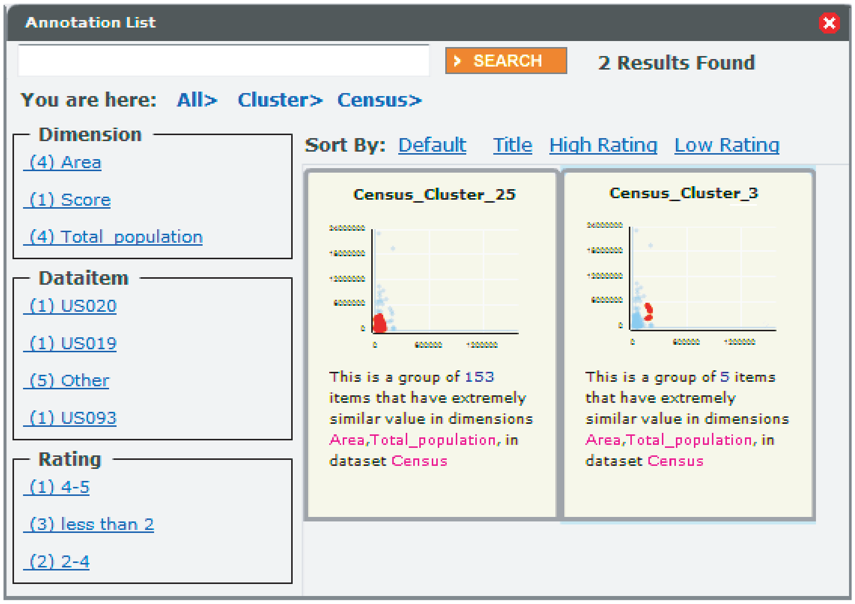
6.1 Introduction

The purpose of capturing discoveries is to use them for making decisions. During a complex visual analytics process, successful decision making requires the ability to retrieve useful information from large collections of discoveries. To this, users would like to pose queries containing rich context from different aspects, and find relevant discoveries to those queries. For instance, a user should be able to search for discoveries about a specific dataset, involving multiple data dimensions, or created by a collaborator. Such flexible queries are difficult to conduct in visual analytics systems that manually generate annotations since they usually contain unformalized and unstructured information difficult to search.

Moreover, when users want to continue an analysis performed in the past, they need to browse and review the relevant discoveries, either generated by their own or collaborators. Often, they need substantial flexibility to access the discoveries, such as browsing and exploring them in any visualizations where the relevant context, such as data items or dimensions, can be observed. Such flexibility, however, is not adequately supported in most visual analytics systems because of the lack of accuracy and formalization of manually generated annotations.

In this chapter, we present two techniques, faceted discovery search and scented discovery browsing, to address the aforementioned challenges. The efficiency of the approaches is achieved by utilizing the context-rich information semi-automatically captured by Click2Annotate (Chapter 5). In particular, the faceted discovery search allows users to create their own custom queries by combining the rich context information of discoveries. The scented discovery approach utilizes the context of discoveries to make the annotations transparent to data and visualizations [52].

6.2 Faceted Discovery Search



This is a group of 153 items that have extremely similar value in dimensions Area,Total_population, in dataset Census

This is a group of 5 items that have extremely similar value in dimensions Area,Total_population, in dataset Census

Figure 6.1: The faceted discovery search interface.

Faceted search [81], a popular searching approach used in mass online markets, has shown its efficiency and flexibility in finding items using custom navigation based on various perspectives, rather than through a specific path. Since the annotations generated by Click2Annotate can also be aggregated based on multiple attributes, faceted search can be applied to help users retrieve discoveries according to their specific analysis interest. In particular, a set of common attributes shared by multiple templates, including author, time, rate, title, discovery type, dataset, dimensions, and tags, are used as faceted filters for searching discoveries in ManyInsights. Users can search discoveries in any order using these filters through the *faceted search interface* provided by ManyInsights (see Figure 6.1).

For example, Figure 6.1 shows how a user retrieves a annotated cluster using the *faceted search interface*. First, she uses the pattern type "cluster" to filter out discoveries that are not clusters. Second, she narrows down the results using the dataset name "census". The search results dynamically roll over the screen from left to right. Besides, keyword search is also provided in the interface.

Inspired by the document card approach [82], we use an annotation card to represent each discovery, showing a preview of the content in the retrieval interface. The annotation card summarizes the discovery using a visualization thumbnail and a short sentence that captures the essential information of the discovery. It allows the user to quickly capture the main content of the discovery. The user can sort the search results by different criteria, such as rate and title. When the user clicks on an annotation card, the annotation will be presented in full detail in an annotation pop-up window. Once the user finds interesting discoveries, she can either export them to XML files for sharing or further group and associate them in the discovery network.

6.3 Scented Discovery Browsing

We propose a scented discovery browsing approach (see Figure 6.2). If a user turns on the scented browsing mode, discovery flags are attached to the visualizations, not only the views where the discoveries were captured, but also other views where the relevant data items/dimensions of the discoveries can be observed. Users can retrieve a discovery from any view where its flag is displayed by clicking on the flag. Compared to existing approaches that require users to manually mark discoveries on the visualizations [54], our approach has several benefits.

First, based on the pre-defined essential information for different pattern types,

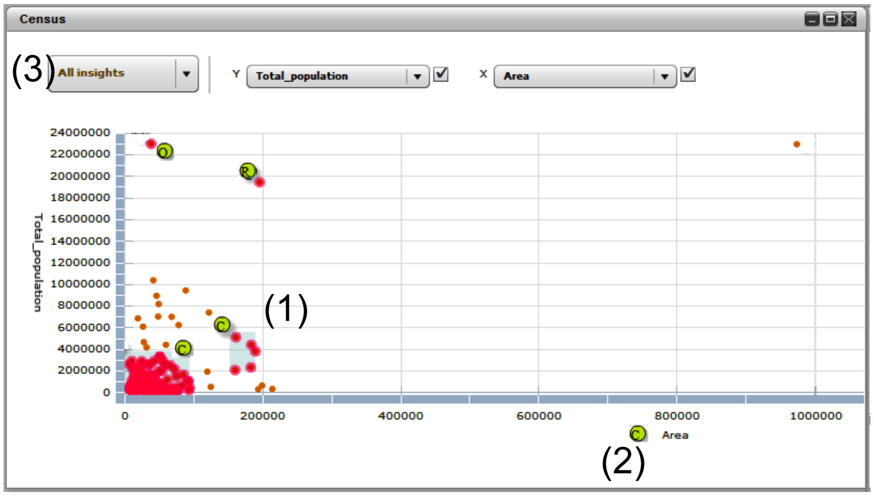


Figure 6.2: The scented discovery browsing.

our approach automatically marks different types of discoveries in different ways to avoid cluttering the display. For example, Figure 6.2 shows a scatterplot with multiple annotated discoveries flagged. In this figure, the flags of data item-oriented discoveries are attached to their data items (see Figure 6.2(1)) while the flags of dimension-oriented discoveries are attached to their dimensions (see Figure 6.2(2)). To reduce the displayed objects, users have options to expand and collapse an annotation in pop-up window by clicking on the corresponding flag. In systems with manually generated annotations, users have to draw marks carefully to achieve similar effects.

Second, discoveries can be flagged in any display where the relevant data items/dimensions of them can be observed, not only the visualization where the discoveries were discovered. Thus it is an “annotate once, appear anywhere” approach [83]. For example, an discovery of dimension correlation can be marked in any of the visualizations where any of the dimensions involved is displayed. This feature allows the users to access relevant discoveries anywhere during their visual exploration process, without going back to the previous views or re-annotating them in the new view. Moreover, by making the annotations transparent to the visualizations, the users can use the annotated dimensions and data items as a focus point from which to find related annotations and visualizations. Thus the visual exploration becomes more convenient and flexible. Such an “annotate once, appear anywhere” approach is not easily supported by manual annotation approaches because of the lack of accuracy and formalization of manually generated annotations.

In addition, the scented discovery browsing approach can work together with the faceted discovery retrieval approach as described in Section 6.2. In particular, users can interactively select the discoveries they want to flag using criteria such as the discovery types and the dimensions involved (see Figure 6.2(3)). In this way the users can display only flags of discoveries of interest in the display to reduce clutter. Again, this benefit is brought by the accuracy and formalization of the automatically generated semantic-rich annotations.

6.4 Conclusion

In this chapter, we presented two techniques to support discovery retrieval and browsing in a visual analytics process. The effectiveness and efficiency of the techniques are achieved by utilizing the rich context of discoveries captured in annotations. As a group of interesting discoveries is retrieved, users need to further organize and associate them to explore their correlations. In the next chapter, we describe a suite of visual analytics approaches to explore the correlations among the retrieved discoveries.

CHAPTER 7: DISCOVERY CORRELATION EXPLORATION

Effective decision making requires users to connect interrelated discoveries to make sense out of them, and to form hypotheses about how these discoveries are correlated [64]. Exploring discovery correlations is a challenging task due to the scale of complex visual analytics, and becomes more difficult for the collaborative visual analysis since users need to associate not only their personal discoveries, but also the discoveries found by others. In this chapter, we present a novel visual analytics approach that automatically organizes, summarizes, and associate discoveries to explore their correlations for hypothesis generation and evaluation. We also present a case study and a user study to demonstrate the effectiveness of the approach.

7.1 Introduction

Decision making involves iterative information forging and sense making loops. Users need to continuously gather information from data, dynamically retrieve and organize the information based on their drifting analysis needs, and associate the interrelated discoveries to form hypotheses. Organizing and associating discoveries allows the users to derive unknown information or new hypotheses, and guides their exploration towards a direction that might lead to more discoveries relevant to the information or hypotheses [64]. In collaborative visual analytics, collaborative workers often experience an initial analysis stage, where they need to browse, organize, and associate each other's individual discoveries to reach a common ground [15]. Effective common ground construction minimizes the need to verbally confirm actions among collaborators, and reduces the cost of collaborative effort [16].

Organzine discoveries and revealing their correlations are essential steps for driving effective decisions and facilitating collaboration. However, users often face significant challenges in conducting these tasks during a complex visual analytics process, where vast amounts of discoveries could be generated from diverse collaborative workers and many datasets. Browsing, organizing, and associating information in such high volume and great variety are challenging tasks. To make it worse, it is often difficult for users to collaboratively manage their discoveries through effective communication methods such as face-to-face discussion. Thus, there is an urgent need for effective and efficient visual analytics approaches for organizing and associating discoveries, especially the following tasks:

Requirement 1 - Generating Organizational Overview: Exploring discovery correlation usually starts from forming an overview of the discoveries that have been recorded and gathered [15]. The overview presents the overall structure, key aspects, and evolution of the discoveries to help the users gauge the context and determine future direction [84]. Existing visual analytics systems provide limited capability with manual inspection and organization, which hinders users' efforts on quickly forming a mental map of existing analysis. New approaches for overview generation that satisfy the following requirements must be developed:

- Collecting information effectively and efficiently: Rich semantic information about discoveries is needed for automated discovery organization, retrieval, and association according to varying user interests. Collecting such information should not impose extra burden to users, i.e., their ongoing visual exploration process should not be disturbed or diverted.
- Employing automatic discovery analysis: Manual discovery association and grouping are not realistic for a fast growing pool of discoveries in dynamic knowledge construction and collaboration process. Development and integration of automatic discovery analysis techniques, such as automated discovery correlation, clustering, and summarization, is direly needed in the framework for fast and operative overview generation.

- Supporting dynamic overview construction: In a complex visual analytics process, analysts usually have diverse information needs. Dynamic overview construction should be supported in the framework so that the analysts can explore the discovery space according to their specific needs. Moreover, the approach should allow the users to dynamically manipulate visualization results according to their changing interests and developing understanding.
- Providing a rich set of views and interactions: Multiple coordinated views should be provided to allow users to examine discoveries from different aspects. For example, temporal visualization helps users track and employ temporal evolution of discoveries, so that they can keep awareness of timing and preserve historical contents of discoveries [54, 15]. Furthermore, proximity-based projections, where closely related discoveries are visually presented as clusters, facilitate users' ability to browse many discoveries at a glance. In addition, interactions should be provided so that users can effectively navigate within the discovery space, retrieve discoveries of interest, and manage overview construction.

Requirement 2 - Supporting Sensemaking: After the users identify interesting discoveries from the overview, they need to closely explore these discoveries for forming hypotheses. The following tasks are important in the sensemaking process:

- Comparison: Scalable comparison among discoveries should be supported. The comparison can be among discoveries generated by different analysts, from different datasets, during different time periods, or discoveries contributing to conflicting or relevant hypotheses. Overlapping information identified from comparison helps analysts to associate discoveries captured in different analysis steps or by different analysts and to acquire additional evidence for developing hypotheses [15]. It also helps analysts to retrieve contextual information, to examine the historical evolution of the reasoning process, and to evaluate conflicting hypotheses.

- Revisiting and refining: Sensemaking is an iterative process. The system should allow analysts to revisit the sources of existing discoveries and refine them without disturbing the ongoing analytic process. This function is important in promoting new discoveries and hypotheses.
- Result outreach: A crucial function of sensemaking is to ensure that analysis results can be preserved and shared among groups [15]. Hence, the approach should provide solutions for this task.

In this chapter, we propose a novel visual analytics toolkit to address these critical tasks. The toolkit automatically retrieves important semantic information about discoveries, such as what they are (e.g., clusters, outliers, ranks, etc.), relevant data information (e.g., dimension names, data item names, etc.), and meta information (e.g., authors, timestamps, etc.), from semi-automatically generated, formalized discovery annotations (Chapter 5). A rich set of views and interactions built upon automatic discovery analysis is then provided. This allows users to browse semantics and identify clusters from a large collection of discoveries, to track their temporal evolutions, to retrieve and compare groups of discoveries, and to preserve and share results for hypothesis generation and collaboration. We demonstrate the effectiveness of the toolkit through a case study and a user study.

7.2 Approach

The semi-automatic discovery annotation approach, named Click2Annotate (see Chapter), is a basis of the proposed toolkit. The pipeline of the toolkit consists of the following steps:

- The system collects semantic information of discoveries when users semi-automatically annotate discovery using Click2Annotate (see Chapter). The contents are stored in a discovery database, upon which correlations between the discoveries can be calculated using the approach presented in Section 7.3.
- After the users retrieve a collection of discoveries, the discoveries are automati-

cally clustered based on the correlations. The clustering results are represented in an automatically generated dynamic organization view (see Section 7.4.2). The view intends to support the requirements of generating overview (R1). It provides multiple organizational metaphors, coordinated analysis components, and animations to help the users to explore discovery clusters, their major semantics, and their temporal evolution.

- The users iteratively refine the discovery clusters on the dynamic organization view to reflect their aims and concentrations. They can dynamically adjust the weights of different contents in the discovery correlation calculation (see Section 7.3) to change the clusters, or refine the search with instant visual feedbacks.
- The users select discoveries of interest from the dynamic view. The selected discoveries are compared and examined in detail in the region graph (see Section 7.4.5), which intends to support the requirements of sensemaking (R2).
- The users preserve and share their key findings and hypotheses for future use by other collaborators (see Section 7.4.7).

7.3 Discovery Correlation Calculation

Using the Click2Annotate technique (Chapter 5), the following information is recorded for each discovery: (1) data contents, such as dataset names, types of insights (e.g., clusters, outliers, rank, correlation, and etc.), relevant dimensions and data items, and essential characteristics of the insights (such as the mean of clusters); (2) user-generated semantic information, such as hypotheses associated with the insights and tags given by users; (3) meta information, such as the name of the author who annotated an insight and a timestamp recording when the insight was annotated.

The rich annotated information is used for modeling and calculating the complex interrelationships of discoveries, which forms the foundation of our approach. The correlation between two discoveries is calculated as a weighted sum of the following similarity measures. The measures are normalized to the range 0 to 1 and their weights can be interactively adjusted by users:

- Closeness in data space: We use data similarity (Sim_{data}) to capture the closeness of two discoveries in the data space. It is calculated using Exact Transformation Measure (ETM) (please refer to [60] for details). ETM is based on transform cost and can handle subsets with different data populations efficiently, which makes it suitable in our application to calculate the closeness between discoveries of different types (e.g., an outlier and a large cluster). If two discoveries are not in the same dataset, their data similarity is set to 0.
- Shared dimensions and data items: We use dimension similarity (Sim_{dim}) and data item similarity (Sim_{item}) to capture the relationships between two discoveries involving the same dimensions and data items, respectively. Note that the two discoveries can be about different datasets that share dimensions and data items. By considering each dimension as a weighted keyword and an discovery as a document, we use an improved cosine similarity measure [85] to calculate Sim_{dim}:

  $$Sim_{dim}(I_i, I_j) = \frac{\sum_{k=1}^{K}(\log_2 \frac{N}{n_k} W_k)(\log_2 \frac{N}{n_k} W_k)}{\sqrt{\sum_{k=1}^{K}(\log_2 \frac{N}{n_k} W_k)^2 . \sum_{k=1}^{n}(\log_2 \frac{N}{n_k} W_k)^2}} \tag{7.1}$$

  where I_i and I_j are two discoveries, W_k is the importance of a shared dimension k, K is the total number of shared dimensions, N is the total number of discoveries, and n_k is the number of discoveries sharing the dimension k inside N. Data item similarity (Sim_{item}) is computed in a similar way.

  Each dimension or data item is assigned an importance in the above calculation. We allow users to interactively set the importance of individual dimensions and data items according to their exploration focus. For example, by assigning a high importance to a dimension of interest, discoveries containing this dimension are considered closely related.

- Shared pattern type: People may want to examine and compare discoveries of the same type, such as all discoveries about ranks, at the same time. Type similarity (Sim_{type}) allows users to conduct this task. $Sim_{type}(I_i,I_j)$ is 1 if discovery I_i and I_j have the same type and 0 otherwise.
- Shared hypotheses: In collaborative visual analytics, users often use discoveries as evidence to support or refute their hypotheses [15]. Discoveries that are associated with the same hypothesis may have semantic relationships. Hypothesis similarity $Sim_{hypo}(I_i,I_j)$ is 1 if discovery I_i and I_j are associated with the same hypothesis and 0 otherwise.
- Shared tags: Tags with descriptive text can be attached to a discovery to express user interest, record their evaluations, and convey the semantic properties of that discovery [57]. They are manually generated by users and shared among discoveries in ManyInsights. Sharing tags indicates semantic relationships between discoveries. Tag similarity (Sim_{tag}) is used to capture such relationships. In particular, each tag is considered a keyword and Sim_{tag} is also calculated using Equation 7.1 with user-specified importance.
- Author: Users often want to examine discoveries from the same author. We define author similarity $Sim_{author}(I_i,I_j)$ as 1 if discoveries I_i and I_j are created by the same user and 0 otherwise.

We calculate discovery correlation ($Cor_{discovery}$) between any pair of discoveries I_i and I_j using the similarity measures described above:

$$\begin{aligned} Cor_{discovery}(I_i, I_j) &= w_{data}Sim_{data}(I_i, I_j) + w_{dim}Sim_{dim}(I_i, I_j) \\ &\quad + w_{item}Sim_{item}(I_i, I_j) + w_{type}Sim_{type}(I_i, I_j) \\ &\quad + \ldots + w_{author}Sim_{author}(I_i, I_j) \end{aligned}$$

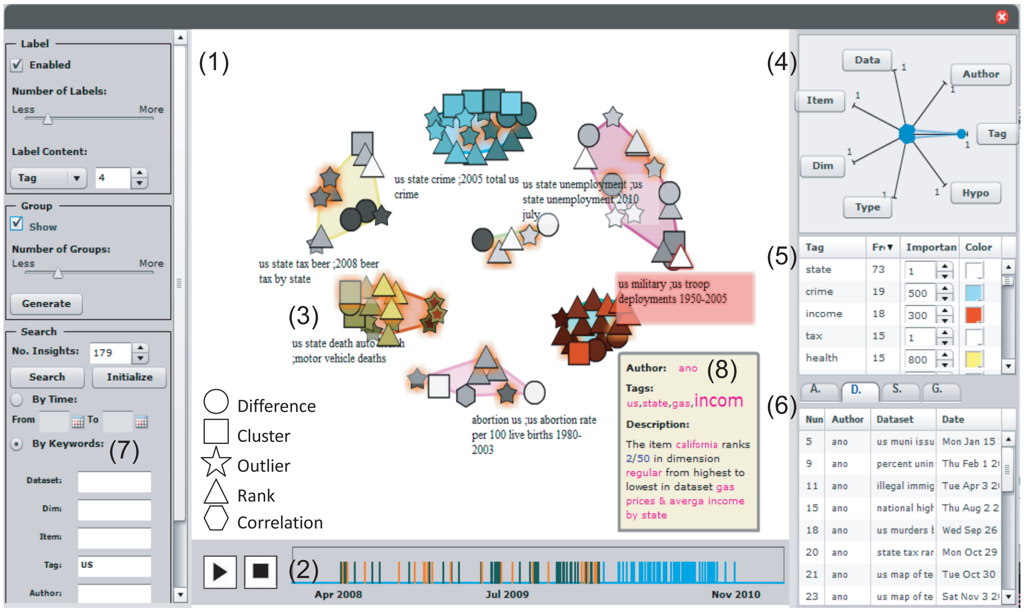
where w_{data}, $w_{dim} \ldots w_{author}$ are user-controllable weights. The sum of all weights equals 1. By adjusting the weights, users can organize and associate discoveries according to a variety of interests. For example, if the users are interested in authors, they can set w_{author} to 1 and other weights to 0. Then the discoveries will be grouped by their authors in the visualization. Section 7.4.2 presents how users interactively adjust the weights and receive instant feedbacks in detail.

Among the similarity measures, the most computationally heavy ones are the data closeness, dimension, data item, and tag similarity calculations. They are either $O(N^2)$ or $O(N^3)$ approaches. However, once the calculation is performed, an individual measure is stored and only re-calculated when users adjust the importance (e.g. changing the importance of a dimension in dimension similarity). Hence, the modification of weights in discovery correlation calculation can be performed efficiently.

7.4 Visualization

Multiple coordinated views are provided in our system to support the requirements of discovery cor. To generate overviews (R1), the dynamic discovery clustering display, the content cloud, the timeline, and the discovery table are provided. The dynamic discovery clustering display (see Figure 7.1(1)) reveals the correlations among the discoveries by placing related discoveries close to each other. It can also reveal the temporal evolution of the discoveries through controllable animations. The content cloud visually summarizes the most significant semantic contents of an discovery group (see Figure 7.2(a)(b)). The timeline allows users to examine the discoveries along a time axis (see Figure 7.1(2)). The discovery table allows users to access the discoveries in a familiar table metaphor(see Figure 7.1(6)).

Our system also allows users to compare groups of discoveries, examine them in detail, and preserve and share the findings they derived (R2). The region graph (see Figure 7.3) presents the relationships among a group of discoveries in detail. It also allows the users to compare two groups of discoveries for shared or distinct informa-


(1)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
Data
Author
Item
Tag
Dim
Hypo
Type
1
us state crime ;2005 total us crime
us state unemployment ;us state unemployment 2010 july
us state tax beer ;2008 beer tax by state
us military ;us troop deployments 1950-2005
us state death auto death ;motor vehicle deaths
abortion us ;us abortion rate per 100 live births 1980-2003
Difference
Cluster
Outlier
Rank
Correlation
Apr 2008
Jul 2009
Nov 2010
Generate
Search
Initialize
A.
D.
S.
G.

Label

☑ Enabled

Number of Labels: Less — More

Label Content: Tag ▼ 4

Group

☑ Show

Number of Groups: Less — More

Search

No. Insights: 179

○ By Time: From To

● By Keywords:

Dataset:

Dim:

Item:

Tag: US

Author:

Tag	Fr▼	Importan	Color
state	73	1	☐
crime	19	500	🟦
income	18	300	🟥
tax	15	1	☐
health	15	800	🟨

Nun	Author	Dataset	Date
5	ano	us muni issu	Mon Jan 15
9	ano	percent unin	Thu Feb 1 2
11	ano	illegal immig	Tue Apr 3 2
15	ano	national high	Thu Aug 2 2
18	ano	us murders l	Wed Sep 26
20	ano	state tax rar	Mon Oct 29
21	ano	us map of te	Tue Oct 30
23	ano	us map of te	Sat Nov 3 2

Author: ano

Tags: us,state,gas,incom

Description:

The item california ranks 2/50 in dimension regular from highest to lowest in dataset gas prices & averga income by state

Figure 7.1: The dynamic discovery clustering display interface. The left part includes label and group controls and search interfaces. The center part is the dynamic discovery clustering display and the timeline (bottom). The right part includes weight controller, keyword tables, and discovery tables. In the dynamic discovery clustering display, each discovery is represented by a shaped particle, colored according to the keywords it contains ("health" - yellow, "income" - red, and "crime" - blue). Discoveries containing the keyword "Texas" are selected and highlighted by orange halos. The weight of tag similarity is set to 1 and others are set to 0.

tion (see Figure 7.4). The users can further examine a discovery in an annotation window (see Figure 7.2(c)) or revisit the data in a multidimensional display for more discoveries (see Figure 7.5). They can also preserve and share their exploration results by creating new discoveries. In the following sections, we present the views and interactions in detail.

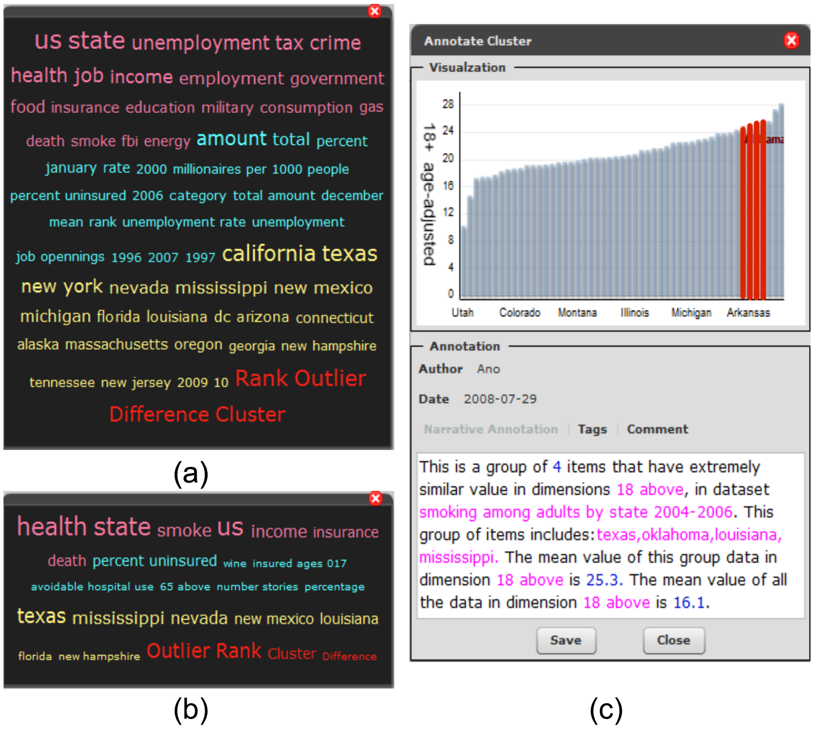
7.4.1 Content Cloud

Figure 7.2(a) shows a content cloud of 179 discoveries. Each tag is a frequently shared keyword in the discovery annotations. Keywords from the same type of contents are grouped together with the same color. For example, the blue keywords are all from dimension names. The size of a keyword indicates its frequency of occurrence in all discoveries or the importance of the keyword assigned by users. In a cloud of all discoveries in an overview (see Figure 7.2(a)), the tags are ordered according to descending frequencies. In a cloud of a group of discoveries selected from the overview interface (see Figure 7.2(b)), the tags are ordered by their TF-IDF weights [86] to emphasize salient features of the group. The TF-IDF weights are calculated using an improved TF-IDF weighting algorithm described in [87]. In Figure 7.2(b), keyword “smoke” is ranked high since it is significant in this group, despite that its global frequency is low.

Interactions: The content cloud provides users a convenient way to start exploring a set of unknown discoveries. In particular, by clicking a keyword in a cloud, users can select all discoveries with this keyword in their contents. They can also set the colors or importance of keywords to control the dynamic discovery clustering display from the content cloud.

7.4.2 Dynamic Clustering View

The dynamic discovery clustering display reveals correlations among discoveries by placing related discoveries close to each other. It also reveals the temporal evolution of the annotation activities through controllable animations. Figure 7.1(1) shows 90 discoveries in the dynamic discovery clustering display. Discoveries are represented as particles with a variety of shapes indicating their pattern types (see the shape legend in Figure 7.1(1))). The luminance of the particles indicates the age of the discoveries (the darker, the older). Discoveries are automatically clustered according to their correlations (refer to [85] for details of the underlying force-based dynamic system).


us state unemployment tax crime
health job income employment government
food insurance education military consumption gas
death smoke fbi energy amount total percent
january rate 2000 millionaires per 1000 people
percent uninsured 2006 category total amount december
mean rank unemployment rate unemployment
job opennings 1996 2007 1997 california texas
new york nevada mississippi new mexico
michigan florida louisiana dc arizona connecticut
alaska massachusetts oregon georgia new hampshire
tennessee new jersey 2009 10 Rank Outlier
Difference Cluster
(a)
health state smoke us income insurance
death percent uninsured wine insured ages 017
avoidable hospital use 65 above number stories percentage
texas mississippi nevada new mexico louisiana
florida new hampshire Outlier Rank Cluster Difference
(b)
Annotate Cluster
Visualzation
18+ age-adjusted
0
4
8
12
16
20
24
28
Utah
Colorado
Montana
Illinois
Michigan
Arkansas
Alaba
Annotation
Narrative Annotation
Tags
Comment
Save
Close
(c)

Author Ano

Date 2008-07-29

This is a group of 4 items that have extremely similar value in dimensions 18 above, in dataset smoking among adults by state 2004-2006. This group of items includes:texas,oklahoma,louisiana, mississippi. The mean value of this group data in dimension 18 above is 25.3. The mean value of all the data in dimension 18 above is 16.1.

Figure 7.2: (a) A content cloud shows the most significant contents of 179 discoveries. (b) A content cloud shows the most significant contents of a group of discoveries selected from the 179 discoveries. Content colors: tag - pink, dimension - blue, data item - yellow, and type - red. (c) An annotation window, showing the visualization and contents of a discovery in detail.

Labels are automatically generated to convey the semantics of the discovery clusters (see Figure 7.1(3) for an example). Users can interactively control which types of discovery contents to be included in the labels. To avoid lengthy labels, we only use the top-N most frequent keywords.

Dynamic clustering: Users can dynamically cluster the discoveries in this view to reflect their current exploration interest. For example, by setting the tag weight to 1 through the star glyph (see Figure7.1(4)), discoveries are grouped by their tags. Users can also interactively adjust the importance of keywords from the keyword table [1] (see Figure 7.1(5)) to cluster the discoveries by keywords, as shown in Figure 7.1(1).

Animation: Users can play animations to examine temporal evolution of the discoveries. During the animation, discoveries are continuously injected into the display in chronological order. The layout gradually evolves to reveal how clusters are formed and evolving over time. Users can use play and stop buttons to pause and resume the animation. They can jump to a particular moment using the timeline (see Section 7.4.3).

Tracking keywords: To track discoveries with keywords of interest, users can assign colors to them from the keyword table (see Figure 7.1(5)). An discovery can have multiple colors if it contains multiple keywords of interest (see Figure 7.6(a-1)).

Grouping and tracking discoveries: The system can automatically divide the discoveries into groups according to a user-defined dissimilarity threshold. In Figure 7.1(1), the automatically generated groups are highlighted in different background colors. The groups are stored in a group table for further operations, such as comparison, viewing content clouds, and etc.. During an animation, users can highlight groups using colored halos for tracking their evolution (see Figure 7.6(a-1)).

7.4.3 Timeline

According to the experimental evidence reported in [34], users can easily understand the development of discoveries by organizing them in chronological order. Following this, the timeline represents discoveries as bars along a time axis (see Figure 7.1(2)), whose distribution reflects the temporal distribution of the discoveries.

Interaction: The timeline is coordinated with the dynamic discovery clustering display. Clicking a bar will navigate to that particular moment in an animation. Bars in blue are discoveries yet to be displayed in the dynamic discovery clustering

[1] Keywords stored in this table are discovery contents, such as dimension names, data item names, tags, authors, and etc..

display.

7.4.4 Discovery Table

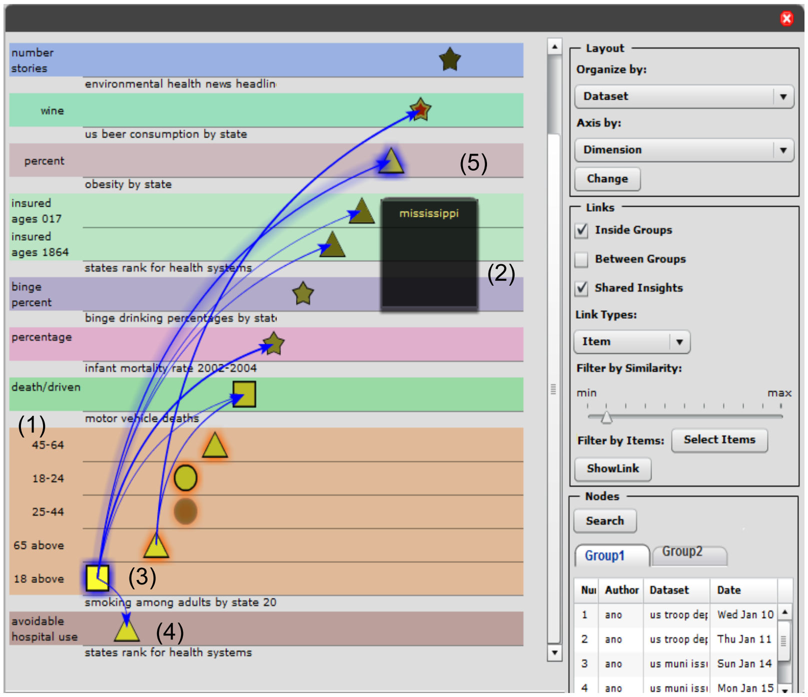
Users can examine a group of discoveries, all displayed discoveries, or the entire discovery collection from the discovery table (see Figure 7.1(6)). They can sort the discoveries by different contents and manually construct discoveries groups.

7.4.5 Region Graph

A region graph, inspired by the substrate graph [88], presents the relationships among a group of discoveries in detail. It also allows the users to compare two groups of discoveries for shared or distinct information. The region graph can have one or two columns, each for a discovery group to be examined. In Figure 7.3, the details and the relations among discoveries in the same group are examined. In Figure 7.4, two groups of discoveries are compared and associated. Nodes displayed in the region graphs represent discoveries, which have the same visual representations (e.g., shapes and colors) with the discovery particles in the dynamic discovery clustering display.

Layouts: In the region graph, discoveries are represented as particles in the same way as the dynamic discovery clustering display. They are placed in nonoverlapping, user-defined content substrates based on their contents. For example, in Figure 7.3, each substrate is a rectangle with a distinct color. It represents a dataset whose name is displayed underneath it. A substrate is evenly divided into rows, each of which presents a dimension of the dataset appearing in the discoveries. The labels of the dimensions are placed on the left of the rows. Only datasets and dimensions appearing in the discoveries are displayed. Users can also map other contents to rows (e.g., tags and authors). Each discovery is displayed in one or more rows according to the dataset and the dimensions it is related to. Its horizontal position is tied to its age. The oldest discoveries are on the right and the newest ones are on the left. When a discovery is displayed in multiple rows (it happens when the discovery is related to multiple datasets or multiple dimensions), the topmost particle is drawn in solid and the others are drawn with a blurring effect.

Users can learn the basic semantics of a discovery by its spatial location and shape without reading its annotation. They can also easily identify how discoveries are temporally related. Another advantage is that the proportionally-sized regions indicate the relative cardinality of each region. For example, in Figure 7.3, users can quickly identify that the dataset "smoking among adults by state" (see Figure 7.3(1)) has more dimensions involved in the discoveries than other datasets.


number
stories
environmental health news headlin
wine
us beer consumption by state
percent
(5)
obesity by state
insured
ages 017
insured
ages 1864
mississippi
states rank for health systems
(2)
binge
percent
binge drinking percentages by stat
percentage
infant mortality rate 2002-2004
death/driven
motor vehicle deaths
(1)
45-64
18-24
25-44
65 above
18 above
(3)
smoking among adults by state 20
avoidable
hospital use
(4)
states rank for health systems

Layout

Organize by: Dataset ▾

Axis by: Dimension ▾

Links

☑ Inside Groups

□ Between Groups

☑ Shared Insights

Link Types: Item ▾

Filter by Similarity:

Nu	Author	Dataset	Date
1	ano	us troop dep	Wed Jan 10
2	ano	us troop dep	Thu Jan 11
3	ano	us muni issu	Sun Jan 14
4	ano	us muni issu	Mon Jan 15

Figure 7.3: The region graph. The left part shows discoveries and their relationships. The right part includes layout controls, link visibility controls, and discovery tables. Nodes with "health" are yellow and nodes with "income" are red. Data item links are blue.

Links: The region graph represents discovery relationships using directed links between nodes. Discoveries could have multiple relationships, such as shared tags and shared data items. They are distinguished using colors of the links. To reduce clutter, users can interactively turn on/off a type of relationship. The thickness of a link indicates the corresponding similarity measure. Users can hover their mouse over a link to examine the relationship in detail. For example, in Figure 7.3(2), two nodes are connected with a data item link since they share the same data item "Mississippi".

Alignment for two groups: To compare two discovery groups, the region graph horizontally places them in two columns (see Figure 7.4). To help users identify shared regions and rows (they indicate shared contents), we consider two goals when laying out the graphs: (1) any pairwise shared regions/rows should be placed closely to each other; and (2) all shared regions/rows should be grouped and placed in prominent positions (e.g., the topmost position). To achieve these goals, the following iterative, greedy algorithm is used (we assume that both columns use dataset-dimension layouts in the description):

Step 1: we denote a pairwise shared regions that represent the same dataset between two columns as PR_{common}. We denote the difference of the height between PR_{common} as $Diff_{pr}$. Identify all PR_{common} between two columns and put them into a sorted queue (denoted as Q_{pr}) where the one with the smallest $Diff_{pr}$ is first. For each column, the rest of the regions (denoted as $R_{uncommon}$) are placed into a sorted queue (denoted as Q_r) where the one with smallest height is first.

Step 2: take the PR_{common} at the front of Q_{pr}. Place each region of PR_{common} at the topmost position of the corresponding column. For each column, compute the total height of regions that have been placed. Then compute the difference of total height between two columns (denoted as $Diff_{tr}$). For the column with the smaller total height, take the $R_{uncommon}$ at the front of its Q_r and place it at the topmost position and update $Diff_{tr}$. Repeat this step until $Diff_{tr}$ reaches the minimum value.

Step 3: repeat *Step 2* until the Q_{pr} is empty.

Step 4: for each column, take the $R_{uncommon}$ at the front of its Q_r and place it at the topmost position of the column. Repeat this step until the Q_r is empty.

Step 5: for each PR_{common}, sort their rows by identifying overlaping rows and place them on the topmost position of the region.

Figure 7.4 shows a result of the graph alignment. Links crossing two graphs represent discovery relationships between the groups. If a discovery is contained in both groups, the corresponding nodes in each column are connected by red, undirected dot links (see Figure 7.4(1)).

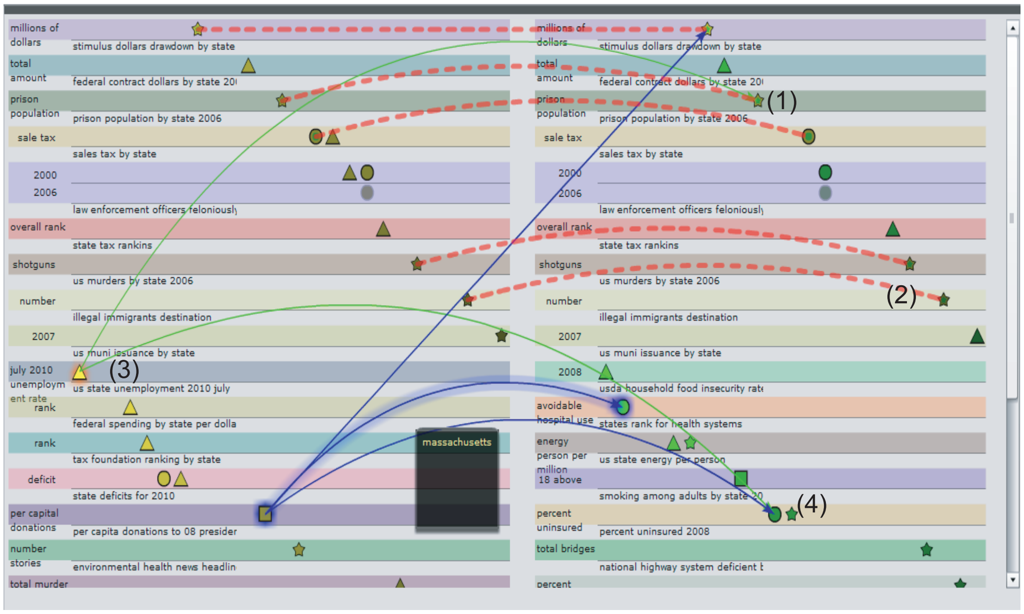

millions of dollars
stimulus dollars drawdown by state
total amount
federal contract dollars by state 20
prison population
prison population by state 2006
sale tax
sales tax by state
2000
2006
law enforcement officers feloniousl
overall rank
state tax rankins
shotguns
us murders by state 2006
number
illegal immigrants destination
2007
us muni issuance by state
july 2010 unemployment rate
(3)
us state unemployment 2010 july
rank
federal spending by state per dolla
rank
tax foundation ranking by state
deficit
state deficits for 2010
per capital donations
per capita donations to 08 presider
number stories
environmental health news headlin
total murder
massachusetts
millions of dollars
stimulus dollars drawdown by state
total amount
federal contract dollars by state 20
(1)
prison population
prison population by state 2006
sale tax
sales tax by state
2000
2006
law enforcement officers feloniousl
overall rank
state tax rankins
shotguns
us murders by state 2006
number
(2)
illegal immigrants destination
2007
us muni issuance by state
2008
usda household food insecurity rate
avoidable hospital use
states rank for health systems
energy person per
us state energy per person
million 18 above
smoking among adults by state 20
percent uninsured
(4)
percent uninsured 2008
total bridges
national highway system deficient b
percent

Figure 7.4: Compare and associate two discovery groups using a region graph. Each column represents a discovery group. Shared regions are indicated by the same colors and labels. Shared discoveries are connected by red, undirected dot links. Nodes with "Texas" are green and nodes with "California" are yellow. Data item links are blue and tag links are green.

Changing layout: Users can change the contents mapped to the regions and rows through a control panel, and hence organize the discoveries in different ways. They can also manually adjust the order of vertical placement of the regions for a graph to place regions of interest in prominent positions.

Filtering links: Users have multiple options to control the visibility of links to reduce clutter.

Visualizing data: Users can select dimensions from the region graph to open a multidimensional display (see Figure 7.5). Within the display the related discoveries will be highlighted, with flags indicating their types (see Chapter 6). In this way, users can explore the visualization for new discoveries or examine existing discoveries for refinement. This function is important in promoting new discoveries and hypotheses.

7.4.6 Other Interactions

Search: Users can search discoveries by a variety of discovery contents (see Figure 7.1(7)).

Manual selection: Users can manually select discoveries from any view by clicking a discovery.

Annotation card: Users can hover their mouse over a discovery to examine its annotation card (see Figure 7.1(8)). It provides a preview of the discovery by summarizing its essential information using descriptive language (see Chapter 6). Keywords in the annotation card are highlighted. The keywords with high importance are enlarged.

7.4.7 Preserving and Sharing Results

In collaborative visual analytics, collaborators need to preserve and share their analytic results, in term of organized discoveries and hypotheses, in a shared work space. Users with different interests may want to organize discoveries in different ways. In addition, a static view is not suitable for the dynamic nature of visual analytics. Therefore, rather than providing a shared work space for persevering and sharing analytic results, we allow users to record their results through a special type of discoveries, namely the hypothesis discoveries. A hypothesis discovery contains a tag given by users. The tag is also assigned to discoveries related to it as their hypothesis contents. A screenshot of the common ground view is attached to the discoveries and users can also make free notes. To review the work of other users, a user can search for hypothesis discoveries. Furthermore, they can organize the discoveries by their hypotheses (see Section 7.3).

7.5 Use Case

To demonstrate how the system can be used to explore a large number of discoveries generated by a diverse set of real users from real datasets, we imported 239 discoveries of 102 datasets from Many Eyes [89] to ManyInsights. Many Eyes [89] is a popular web-based collaborative visual analytics system, where users visually explore datasets contributed by themselves or others. They share their discoveries by posting comments linked to specific visualization displays. A large number of discoveries are reported daily by users from a variety of domains [13]. These discoveries come from a wide range of datasets.

All the discoveries we imported were generated from multidimensional visualization displays. They all contain one of two popular tags: "US" and "world". Most of them were generated from users' comments linked to visualization displays. We reviewed the comments and displays to extract their semantic contents, and manually generated a formalized annotation for each discovery.

Consider Mary and Tom, two graduate students majoring in sociology in different universities, are both investigating the quality of living by state in America. To acquire information, they search Many Eyes for comments on related data. A large number of comments are returned and it is time consuming to read them one by one. Therefore, they use our system to explore these discoveries.

Identifying Clusters: One day, Mary logs into the system and searches discoveries with the tag "US" (see Figure 7.1(7)). 179 discoveries are returned and displayed in the overview interface. From the content cloud (see Figure 7.2(a)), she immediately identifies three tags - "health", "income", and "crime", which are important factors related to the quality of living. To cluster the discoveries by these factors, she increases the weight of the tag similarity to 0.9 and sets the importance of these three tags to a value much higher than the remaining tags. She also assigns colors to discoveries with the three tags, for example, yellow for discoveries with the tag "health". She starts

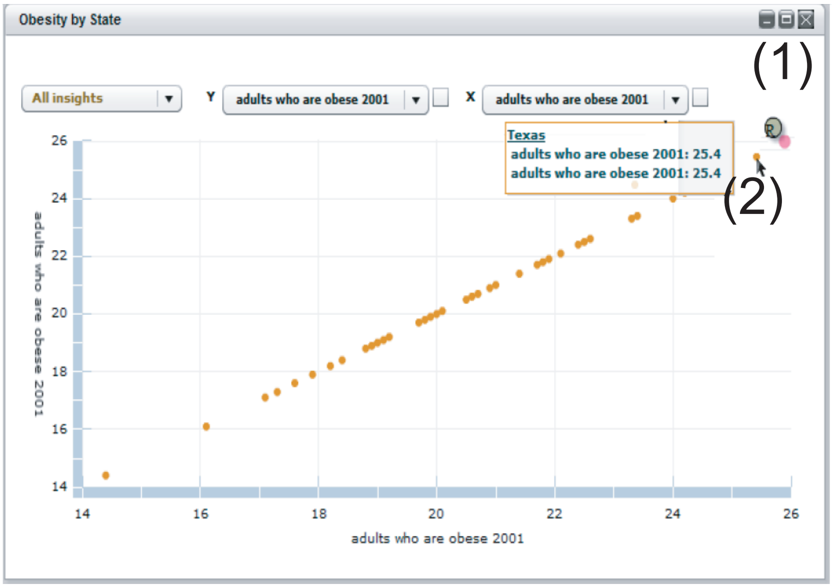

Obesity by State
(1)
All insights
Y
adults who are obese 2001
X
adults who are obese 2001
Texas
adults who are obese 2001: 25.4
adults who are obese 2001: 25.4
(2)
26
24
22
20
18
16
14
adults who are obese 2001
14
16
18
20
22
24
26
adults who are obese 2001

Figure 7.5: Revisit the visualization of the dataset "obesity by state". An annotated discovery is displayed in the visualization.

an animation in the dynamic discovery clustering display, and soon notices several clusters (see Figure 7.1(1)). She is interested in the group of discoveries with the tag "health" (see Figure 7.1(3)). She pauses the animation and selects this group for further exploration.

Examining a cluster in detail: Mary opens a content cloud for this group, as shown in Figure 7.2(b). The data item "Texas" in this figure catches her attention since it is significant in this cloud. Mary then examines the group in a region graph (see Figure 7.3). She quickly locates a dataset, named "smoking among adults by state" (see Figure 7.3(3)), involving more dimensions and discoveries than others. She thinks that the smoking population is highly related to the quality of living, so she focuses her exploration on this dataset. Based on the shapes and the annotation cards of the nodes, she quickly learns the essential content of discoveries in this dataset. To further investigate how discoveries about smoking are related to other discoveries, she displays their links to others. An discovery about a cluster (see Figure 7.3(3)) catches her attention since it has many links to other discoveries. She clicks the discovery to explore it in the annotation window (see Figure 7.2(c)). By reading the annotation, she realizes that “Mississippi”, “Texas”, and other two states have similar high values in percentage of smoking people. Mary finds many interesting links from this discovery to other discoveries. For example, “Texas” is highly ranked in percentage of smoking people (see Figure 7.3(3)) and is ranked low in health systems (see Figure 7.3(4)).

Hypothesis generation and validation: Following the link between (3) and (5) in Figure 7.3, Mary observes that “Mississippi” is also highly ranked in percentage of obesity (see Figure 7.3(5)). Mary thus makes a hypothesis that “Texas” is also high in obesity. To validate her hypothesis, she selects the dataset “obesity by state” and the dimension “percentage” from the graph to create a visualization (see Figure 7.5). From the visualization, she easily discovers that “Texas” (see Figure 7.5(2)) is ranked the second highest in obesity percentage just behind “Mississippi” (see Figure 7.5(1)). Therefore, her hypothesis is confirmed. Mary makes an annotation for this new discovery.

Mary then highlights all discoveries with “Texas” in the dynamic discovery clustering display (see Figure 7.1(1)). She observes that several other discovery clusters also contain discoveries about “Texas”. By examining them she collects more discoveries showing low quality of living in “Texas”. She saves this result in a new hypothesis discovery “living quality in Texas is low” and posts it in ManyDiscoveries. The hypothesis is also attached to the relevant discoveries Mary found.

Comparing discovery groups: Later on, Tom logs into the system and organizes the discoveries by their hypotheses in the dynamic discovery clustering display. He finds the group of discoveries showing low quality of living in “Texas”, which was discovered previously by Mary. Tom has investigated the quality of living in “California”, so he is interested in continuing Mary’s investigation to compare discoveries about “Texas” and “California”. He does so by making a discovery group for “Texas” and another for “California” and compares them in the region graph. He quickly locates several shared discoveries, datasets, and dimensions on the top positions of the region graph (see Figure 7.4). For example, he finds that “Texas” and “California” are extremely high in prison population (see Figure 7.4(1)) and illegal immigration population (see Figure 7.4(2)). Tom then selects two interesting discoveries in the “California” group (left column) and examines their links to the “Texas” group (right column). Here, he chooses to display the tag and data item links and filters out links that contain general terms such as “US” and “Texas”. While examining this display, he realizes that the discoveries about “unemployment population” (see Figure 7.4(3)) are related to discoveries about “uninsured population” (see Figure 7.4(4)) and “prison population” (see Figure 7.4(1)) because they share the same tags.

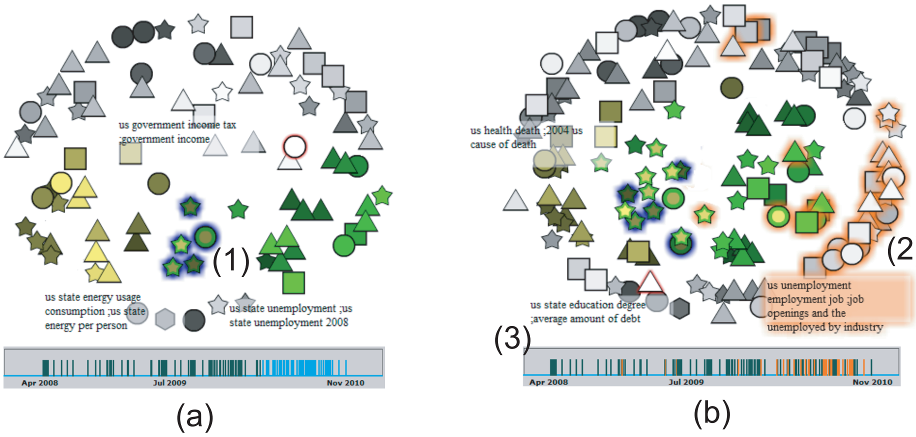

us government income tax ;government income
us state energy usage consumption ;us state energy per person
us state unemployment ;us state unemployment 2008
(1)
Apr 2008
Jul 2009
Nov 2010
(a)
us health death ;2004 us cause of death
us state education degree ;average amount of debt
us unemployment employment job ;job openings and the unemployed by industry
(2)
(3)
Apr 2008
Jul 2009
Nov 2010
(b)

Figure 7.6: (a) A total of 97 discoveries are displayed in the dynamic discovery clustering display (Jan. 2010). (b) 169 discoveries in the dynamic discovery clustering display (Nov. 2010). Data item keyword colors: “Texas” - green and “California” - yellow. Content weights: data item - 0.6 and tag - 0.4. (3) is a cluster of discoveries with both keywords “Texas” and “California”.

Tracking discovery evolution: Tom examines the temporal trends by playing the animation in the dynamic discovery clustering display. Figure 7.6(a) and Figure 7.6(b) show two screenshots during the animation, where the discoveries about both “Texas” and “California” Tom has explored are highlighted in blue halos. During the animation, Tom notices that the highlighted group shown in Figure 7.6(a-1) gets much bigger by adding several outliers about “Texas” and “California”. He also notices a newly formed discovery group (indicated by lighter colors, see Figure 7.6(b-2)). He quickly learns that this group is about “unemployment” and “job” issues according to the label and the content cloud. He highlights discoveries with these keywords. From the timeline view (see Figure 7.6(b-3)), he finds that most highlighted discoveries (they are in orange) were developed after Jul. 2009. Based on this pattern, Tom thinks that more and more people are concerned about how “unemployment” affects their quality of living after the economic crisis. Therefore, “unemployment” and “job” should be important topics for his further investigation of quality of living.

7.6 User Study

A controlled experiment has been conducted to evaluate the effectiveness and efficiency of the region graph. The study was a 3×8 (system types×data) between-subjects design. We compared the region graph with two baseline tools in individual analysis and asynchronous collaboration analysis situations. Participants were asked to use tools to associate and compare discoveries in these two situations. We hypothesized that in both of the situations, subjects could spend the least time and make the highest accuracy using the region graph, and would express a preference for the region graph over the baseline tools.

7.6.1 Baseline Tools

Two baseline tools were compared with the region graph (see Figure 7.7 (a)). The first baseline tool was used to simulate the manual association approaches used in online visualization systems such as Many Eyes [13] and sense.us [54]. It requires users to manually inspect, associate, and compare discoveries in a faceted discovery search interface (Chapter 6). Users can search discoveries using keyword search and faceted search and browse their semantics in annotation cards. When comparing two discovery groups, two search interfaces were provided.

The second baseline tool automatically associates discoveries and represents their

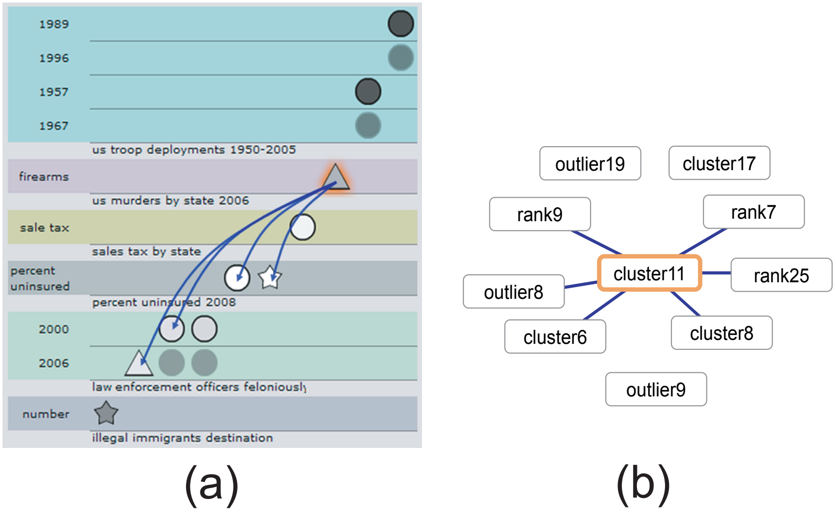


Figure 7.7: The visualizations used in the user study. The same discovery annotation is selected and highlighted in two visualizations. (a) A region graph. (b) A node-link diagram graph.

correlations in a node-link diagram (see Figure 7.7 (b)), similar to the existing structured organization tools (e.g., [51] and [60]). The layout was based on force-directed layout [90]. Each label shows an Id of a discovery. Users can click on a node to explore the discovery in an annotation card. The tool also supports basic graph interactions such as link selection and node filtering. When comparing two discovery groups, two graphs were constructed.

7.6.2 Analytic Settings and Tasks

We considered two analytic situations: individual analysis situation which intended to simulate individual dynamic knowledge construction process, and asynchronous collaboration situation which intended to simulate collaborative analysis. In the individual analysis situation, we assumed that subjects were about to review and explore the discoveries created by their own so that they had been familiar with both the datasets and the discoveries being explored. In the asynchronous collaboration situation, we assumed that subjects were about to explore the discoveries that had been created by their collaborators. Therefore, they had no prior knowledge about the discoveries.

In each analytic situation, two evaluation sessions were conducted: an association session and a comparison session. Each session contained three task groups with four tasks each. The tasks included the general network exploration tasks described in [88] as well as tasks specific to discovery exploration.

Association session: subjects were asked to explore a single discovery group. The session had three task groups:

- Basic association explored the basic structure and relationships in the discovery group. For example, given a discovery, find all the associated discoveries, find the strongest links, and find the links with a given attribute.
- Attribute based association made a deeper understanding by considering the attributes of the discoveries. For example, find the links of the discoveries of the given datasets, dimensions, and types or find the proportion of the links from a discovery that goes to others with the given dimensions.
- Attribute aggregation obtained an aggregation of the attributes for the discovery group, which helped to make sense of the datasets and analysis process for adjusting future exploration directions. For example, find the datasets or dimensions associated with the most links.

Comparison session: subjects were asked to compare two discovery groups. The session had three task groups:

- Basic comparison compared the basic structure and relationships for the two discovery groups. For example, find the shared discoveries between the groups or the links across the groups.
- Attribute based comparison compared the two groups regarding to specific discovery attributes. For example, find the shared discoveries with given datasets and dimensions.
- Attribute aggregation compared the aggregation of attributes between the two discovery groups. For example, find the dimensions associated with the most/least

links across the two discovery groups or finds the datasets containing the most shared discoveries.

7.6.3 Datasets and Discoveries

To generate data for the user study, we collected 174 real users' discoveries from Many Eyes [13]. The discoveries were captured from users' comments linked to multidimensional visualizations. We reviewed each comment, extracted its essential information about the discovery, and manually generated a formalized annotation for it (Chapter 5). The collected discoveries were generated from 49 datasets containing the tag "U.S." and involved at least one of the states in U.S.. Therefore, most of the collected discoveries can be associated by their tags and data items.

For each task session, we manually generated two discovery groups: a small group (15 discoveries) and a large group (30 discoveries). The discoveries in each group were randomly selected from semantically correlated datasets that shared at least two tags. For example, discoveries generated from the datasets "Prison Population by State", "Overall Violent Crime Rate", and "Murder by State" were grouped in the same group since all the datasets contained the tags "U.S." and "criminal". In the comparison session, each discovery group was further divided into two sub-groups. Pairs of the sub-groups had similar sizes and shared at least one discoveries. Figure 7.1 summarizes the number of discoveries and datasets used in the user study.

7.6.4 Analysis Condition and Procedure

A total of 15 subjects (12 males and 3 females) participated in the study. The subjects included 9 graduate students majoring in computer science and statistics and 6 data analysts working in various domains such as financial analysis and bioinformatics. All the subjects were self-identified as having data analysis and visualization experience and had strong English reading ability. Before the user study, the subjects were evenly divided into three groups: one group used the region graph, one group used the search interface, and one group used the simple system. The same discover-

Table 7.1: The design of the region graph evaluation.

Situation	Session	Discovery Size	Dataset Size
Individual exploration	Association	Small: 15	3
		Large: 30	7
	Comparison	Small: 15 (sub1: 8, sub2: 7)	5
		Large: 30 (sub1: 16, sub2: 14)	8
Asynchronous collaboration	Association	Small: 15	5
		Large: 30	9
	Comparison	Small: 15 (sub1: 9, sub2: 8)	5
		Large: 30 (sub1: 19, sub2: 17)	7

ies and tasks were used for all the groups. The subjects took the experiment one by one on the same laptop following the same process.

At the beginning of the study, a 40-minute tutorial was provided by an instructor to explain the discovery annotations and show examples of how to associate the discoveries using the tool. The subjects were also instructed to explore the datasets of the discoveries using parallel coordinates and scatter plot. The individual analysis was conducted right after the tutorial, followed by the asynchronous collaboration. At the beginning of the individual analysis, the subjects were asked to browse and review the discoveries for 30 minutes to understand them. In each situation, the association session was first conducted, followed by the comparison session. In each task group, there were first practical tasks, second experimental tasks (the small-size discovery group was first provided, followed by the large-size discovery group), and survey questions. In each task, we recorded the task results and completion time.

7.6.5 Results

We recorded two types of results: quantitative results (completion time and correctness) which were captured on the tools and the subjective preferences which were captured via survey questions. They were reported in the following sections.

7.6.5.1 Task Completion Time and Correctness

The comparisons of average completion time of the three task groups in individual analysis situation are shown in Figure 7.8(a) for small-size discovery groups and in Figure 7.8(b) for large-size discovery groups. The same comparisons for asynchronous collaboration situation are shown in Figure 7.9(a) and Figure 7.9(b).

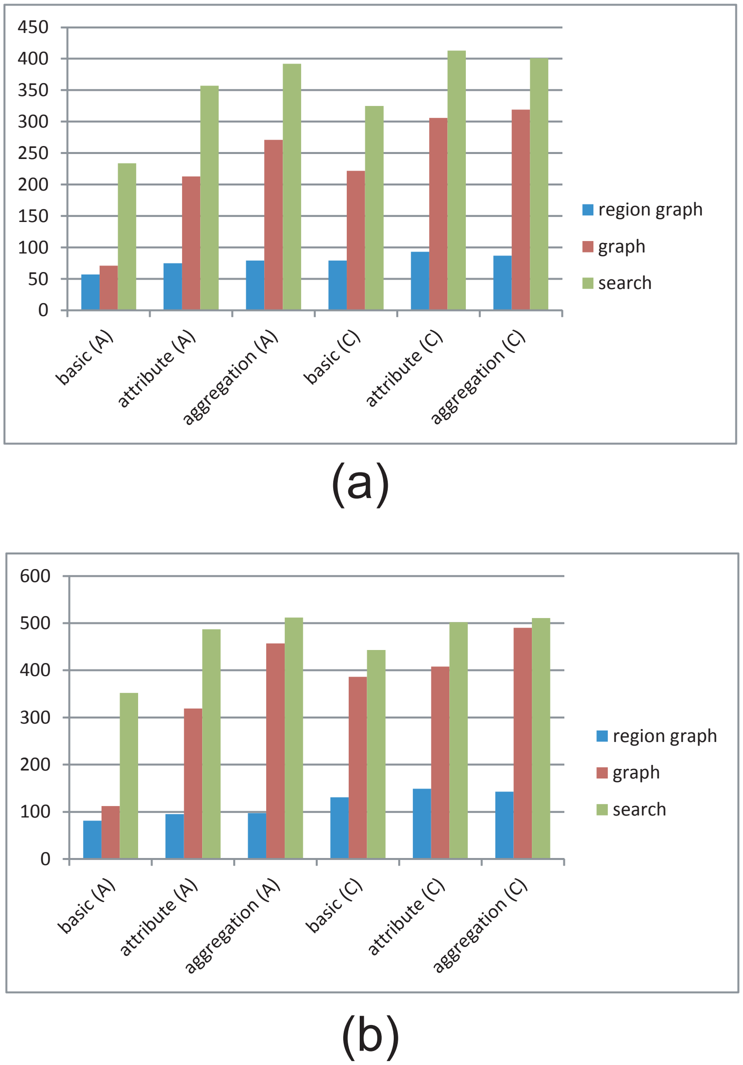


Figure 7.8: The results of individual analysis situation. The task groups with "A" are in the association session. The task groups with "C" are in the comparison session. (a) Using small discovery groups. (b) Using large discovery groups.

In all the situations, the region graph achieved less completion time than the search interface and node-link diagram graph, which confirmed the first hypothesis.

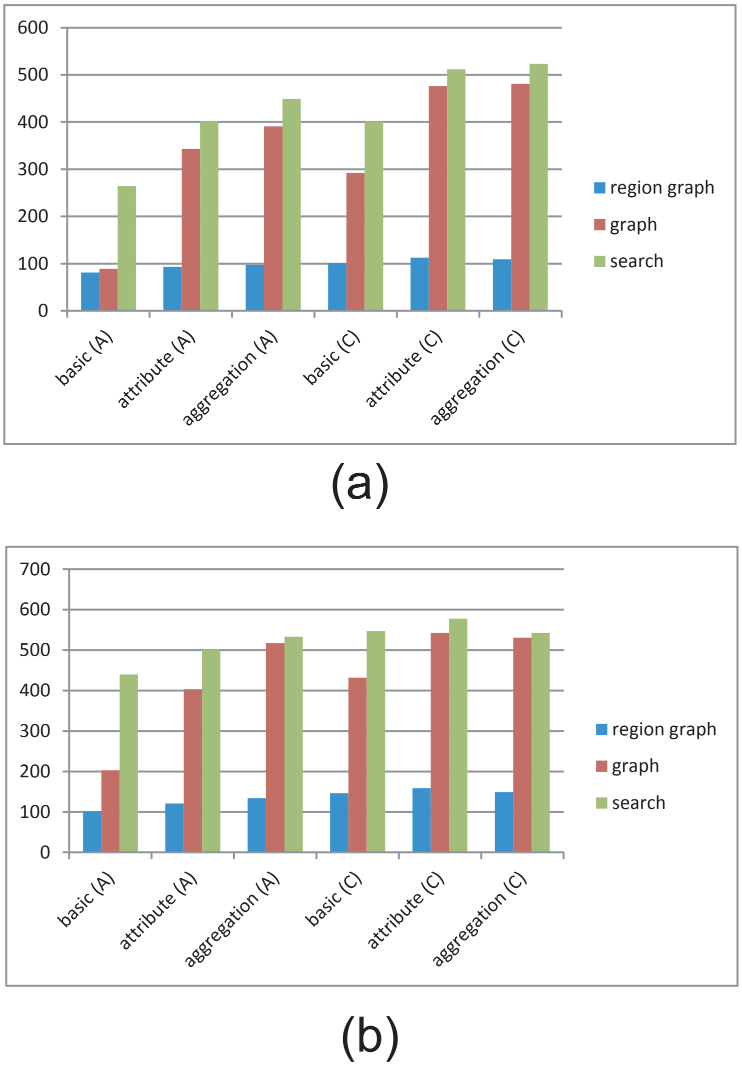


Figure 7.9: The results of asynchronous collaboration situation. The task groups with "A" are in the association session. The task groups with "C" are in the comparison session. (a) Using small discovery groups. (b) Using large discovery groups.

In particular, significant differences can be observed from:

- Average completion time for attribute based tasks and aggregation tasks. The region graph showed the strength in exploring the discovery attributes, which were achieved by the new features such as a novel node layout and representation techniques.
- Average task completion time for large discovery group. The region graph showed its strength in supporting large scale exploration.

- Average task completion time for comparison sessions. The region graph provided advanced visual interface to support the visual comparison between different discovery groups.
- Average task completion time for asynchronous collaboration. The new feature of the region graph allowed the subjects to more quickly understand and reveal the hidden information in their unfamiliar discoveries.

The average accuracy rate for all the sessions in individual analysis was 84.5% for the region graph, 68.2% for graph, and 47.4 % for the faceted discovery search. The average accuracy rate for all the sessions in asynchronous collaboration was 80.5% for the region graph, 64.8% for graph, and 40.5 % for the faceted discovery search. The results suggested that subjects with the region graph could associate and compare discoveries with highest accurcy. The second hypothesis was validated.

Table 7.2: The average ratings of survey questions.

Situation	Session	Task Group	Region Graph	Graph	Search
Individual exploration	Association	Basic	5.2	4.4	2.4
		Attribute	5.6	3.0	1.8
		Aggregation	5.4	2.4	1.8
	Comparison	Basic	5.6	2.8	1.4
		Attribute	5.2	1.6	0.6
		Aggregation	4.8	1.4	0.8
Asynchronous collaboration	Association	Basic	5.0	4.0	2.0
		Attribute	5.0	1.6	1.6
		Aggregation	5.2	1.4	0.8
	Comparison	Basic	5.2	2.2	0.4
		Attribute	4.8	1.4	0.6
		Aggregation	4.6	1.2	0.8

7.6.5.2 Subjective Preferences

At the end of each task group, the subjects were asked to answer survey questions with a 7-point Likert scale (0=strongly disagree, 6=strongly agree) to rate the usefulness and confusedness of the tools. General survey questions were also provided at the end of the user study. A total of 32 questions were provided. The average score for the region graph was 5.1, 2.4 for the node-link diagram graph, and only 1.2 for the simple search interface. Overall, the subjects found the region graph more helpful and less confusing than the other two tools. Table 7.2 summarizes the comparisons of the average score for each task group. Significant differences are observed in attribute based tasks and aggregation tasks in all the analysis sessions and situations. The results prove the benefits of the region graph and confirms our third hypothesis.

7.7 Conclusion

The visual analytics toolkit presented in this chapter is among the first efforts on supporting flexible discovery correlation exploration in visual analytics. The case study and user study suggested that the toolkit greatly reduces human efforts and enhances the visual sense making process by allowing analysts to quickly construct exploratory overviews for a large amount of evolving discoveries, flexibly study their relations and patterns, as well as effectively share and exchange discoveries. We argue that such features are essential and should be supported by all visual analytics systems. In addition, our approach, namely the semi-automatic annotation combined with semi-automatic discovery correlation, is general enough to be extended to other data types, such as geospatial data and graph data. The approach is independent from the visualization platforms where the discoveries are captured and thus it can be used in a wide range of visual analytics applications.

In the future, we plan to extend the toolkit to support the exploration of discoveries generated from miscellaneous data sources and different visualization tools. Such a generalized approach accommodating various datasets and scenarios will benefit a diverse range of communities across scientific and social domains.

CHAPTER 8: CASE STUDY

Although the individual components of ManyInsights, such as the Clicck2Annotate and the region graph, have been evaluated through formal user studies, it is also necessary to understand how these components work together to benefit complex analytic tasks. In this chapter, we present two long-term case studies of ManyInsights conducted by domain experts with real datasets and real analytic tasks. The case studies provided an in-depth understanding of how the proposed discovery management framework and targeted techniques facilitate exploratory data analysis.

8.1 Introduction

In the previous chapters, we have shown the effectiveness and efficiency of the individual components of ManyInsights through controlled user studies. During these studies, the participants were asked to perform predefined tasks (e.g., annotating discoveries) on preselected data. The performance time and accuracy of the participants' response were recorded and analyzed. However, our controlled studies had inherent problems such as the lack of real-use context [32]. In contrast, a real-world analysis scenario with less guide and more in-depth could provide stronger endorsement for our discovery management approaches. Moreover, it is necessary to provide a full picture of the discovery management framework, understanding how the individual discovery management components work together to support the visual reasoning process.

Bearing these needs in mind, we conducted two long-term studies to examine the uses of ManyInsights in two real-world analysis scenarios: a single user analysis scenario and an asynchronous collaborative analysis scenarios. The first scenario intended to understand how the system can impacts the dynamic knowledge construction process, while the second scenario intended to understand how the system can benefit the common ground construction in collaboration. The studies were essentially longitudinal studies with real analytic tasks and real datasets. During the studies, we worked closely with the domain experts to understand their discoveries and analysis strategies. Our observations and interviews provided strong support for the usefulness of ManyInsights and its underlying discovery management framework. The limitation of the system is also discussed.

8.2 Individual Analysis Case Study

We have conducted a long-term case study with a domain expert using real datasets and real analytic tasks. The study was focused on the domain expert's discover management activities in a long-term data exploration process. More specifically, the study intended to answer the following questions:

- How do the proposed discovery management framework impact domain experts' visual analytics process?
- How do the different discovery management functions supported in ManyInsights benefit domain experts' long term analytic tasks?
- What improvements are further anticipated for ManyInsights?

A researcher with over 6 year research experience on environmental policy participated in the study. He was interested in analyzing energy-related carbon dioxide emissions in U.S., so he used ManyInsights to perform an 8-week data analysis on relevant datasets.

8.2.1 Problem, Tasks, and Datasets

The carbon dioxide emissions from energy production (e.g., electricity and transportation) are primarily responsible for the global anthropogenic climate change. Therefore, understanding and reducing energy-related carbon dioxide emissions has become a critical global issue. In the case study, the researcher conducted two specific analytic tasks to analyze energy-related carbon dioxide emissions in U.S.. The first task was to identify which states have the highest CO2 per capita emissions and why

Table 8.1: A partial list of datasets used in the case study.

Dataset	**Example dimensions**
U.S. per capita carbon dioxide emission (2005)	Per capita emissions
U.S. census (2005)	Population, Income per capita, Age, Educational attainment, Housing units, Area, Density
U.S. transportation fuel (2005)	Highway use, Non-highway use, Total use
U.S. transportation fuel use and emission (2005)	Transportation fuel emission, Fuel consumption
U.S. average electric power emissions (2005)	Electricity emission, Generation
U.S. electricity consumption by sector (2005)	Residential, Commercial, Industrial
U.S. average household emission by state (2005)	Household fuel emission, Residential
U.S. electricity generation by state (2005)	Source, Generation
U.S. annual heating degree days: (2005)	Apr., Jan.
U.S. average electricity price per kWh (2005)	Residential, Commercial, Industrial
U.S. average gasoline price per gallon (2005)	Gasoline prices by formulation, Grade, Sales type

they are higher than other states. The second task was to provide recommendations to reduce the emissions for the states with high CO2 emissions.

The researcher came up with his own data which included 22 multidimensional datasets. The number of their dimensions ranged from 4 to 32. Table 8.1 provides a partial list of these datasets as well as their key dimensions.

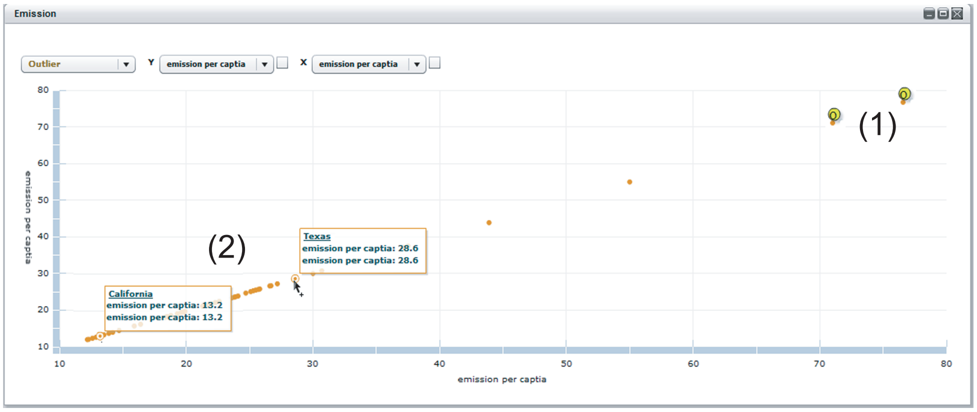
8.2.2 Method

The main methods of the case study were participatory observations and interviews. During the 8-week case study, weekly meetings were conducted between the researcher and an instructor, each of which included a 2-hour data exploration session. A training session was conducted before the data exploration session in the first meeting, in which the instructor introduced ManyInsights to the researcher and taught him how to use it. In each data exploration session, the researcher was asked to use ManyInsights to conduct the two tasks. Four visualizations (parallel coordinates, scatter plot, bar chart, and pie chart) were used for data exploration based on the researcher's request. The instructor observed the process and provided instructions when the researcher encountered any problems. The analytical artifacts generated by the researcher, such as insight annotations, hypotheses, and screenshots of important visualizations, were collected, such as insight annotations, hypotheses, and screenshots of important visualizations. After each data exploration session, the instructor interviewed the researcher to collect his feedback regarding the system and to understand his analysis process and findings.

8.2.3 Observed Analysis Procedure

The researcher began by exploring the *2005 U.S. per capita CO2 emissions* dataset. He identified several states with extremely high per capita emission, such as "Alaska" and "Wyoming" (see Figure 8.1 (1)). He used the outlier and rank templates to record them. The researcher also noticed a significant difference for per capita emission between "California" and "Texas", the two largest states in U.S. (see Figure 8.1 (2)). He thought this was an interesting pattern and used the difference template to annotate the discovery.

Next, the researcher focused on the visual exploration of three datasets, namely *transportation fuel use and emission*, *electric power emissions*, and *average household emission*. They contained important energy consuming information. For each dataset, the researcher identified the states that ranked the highest and the lowest in a variety of dimensions and annotated them accordingly. The captured discoveries were then visually explored in the region graph. The researcher quickly identified several dimensions of interest from the energy consuming datasets, such as "transportation fuel emission" and "household fuel emission". The discoveries about these dimensions included states that also appeared in the discoveries about high per capita overall emission. The researcher called these dimensions the key emission categories.



Outlier

Y emission per captia □

X emission per captia □

Emission
80
70
60
50
40
30
20
10
emission per captia
(1)
(2)
Texas
emission per captia: 28.6
emission per captia: 28.6
California
emission per captia: 13.2
emission per captia: 13.2
10
20
30
40
50
60
70
80
emission per captia

Figure 8.1: The visualization of the *2005 U.S. per capita CO2 emissions* dataset. Several interesting patterns were identified from the visualization. (a) "Alaska" and "Wyoming" had extremely high per capita emissions. (b) "California" and "Texas" had big variation in per capita emission.

After identifying the key emission categories, the researcher explored more datasets related to each category to investigate the factors that caused the emission. In this process, he focused on discoveries about dimension correlations and explored these discoveries using the region graph, as shown in Fig. 8.2. The region graph helped the researcher to develop a global picture of the factors from multiple datasets. For example, the researcher captured several strong correlations in the *transportation fuel use and emission* dataset (e.g., "fuel consumption" and "transportation emission")', the *census* dataset (e.g., "population density" and "per capital fuel consumption"), and the *transportation fuel* dataset (e.g., "fuel price" and "fuel consumption"). By associating these discoveries in the region graph (see Fig. 8.2) and examining the relationships in detail, the researcher concluded that low population density area and low fuel price may cause more highway driving and fuel use, which would account for higher transportation emissions. He commented that the region graph clearly summarized the dimension relationships and allowed him to reach conclusions quickly.

As the data exploration continued, many discoveries were captured and annotated. The researcher extensively used the faceted discovery search and the dynamic discovery clustering display to keep the awareness of these previous analysis results and guide the current exploration. More specifically, when exploring a new dataset, the researcher frequently used the faceted discovery search to identify the dimensions, data items, and tags most frequently captured in the previous analysis sessions. This important information was then used to aid the analysis of the current data for new hypotheses. He also grouped discoveries in the dynamic discovery clustering display. He often assigned high importance to the popular items identified from the cluster labels in clustering. In this way, the researcher could easily inspect the discoveries related to these important items and revisit their visualizations for new discoveries.

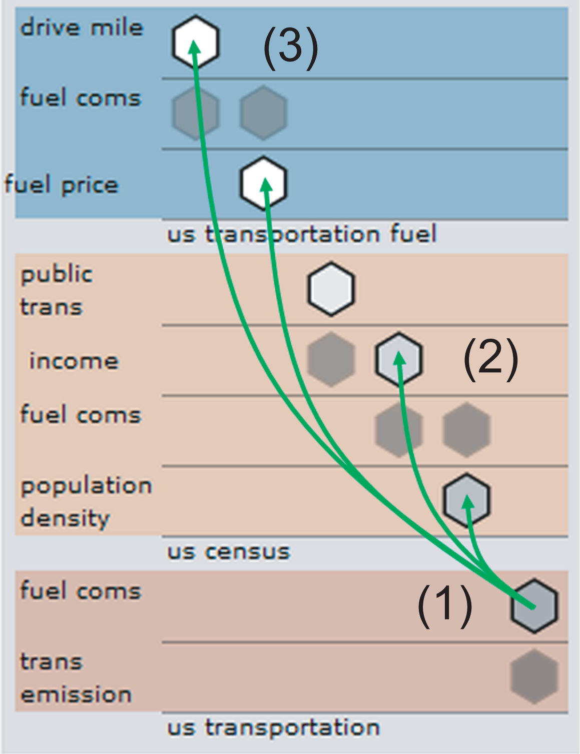


Figure 8.2: Explore the dimension correlations in the region graph.

Toward the end of the study, the researcher utilized the dynamic discovery clustering display and the region graph to review the captured discoveries and find evidence that could explain the high emissions of the states. The dynamic discovery clustering

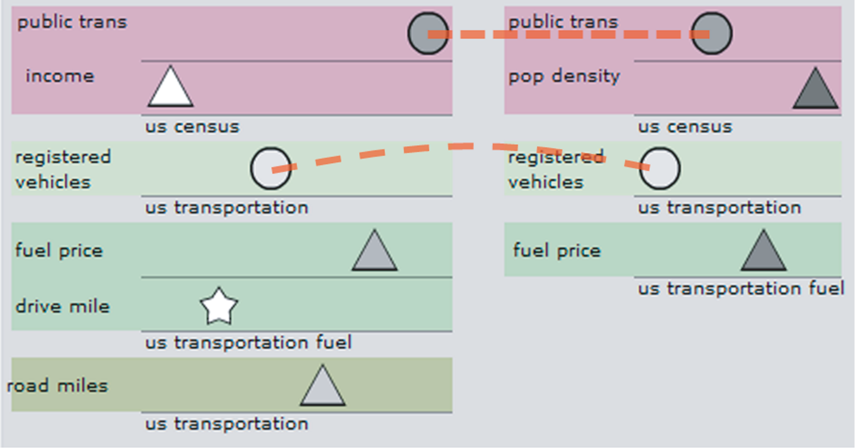


Figure 8.3: Compare two discovery groups about “Texas” (left) and “California” (right) in the region graph. All the discoveries contain the keyword “transportation”.

display allowed the researcher to explore the vast amount of discoveries in a divide-and-conquer manner. More specifically, the researcher first grouped the discoveries by the states they involved. After several clusters were observed, he adjusted the attribute importance to find subsets that contained interesting dimensions or tags within each cluster. By partitioning the clusters into smaller groups, the researcher could flexibly explore and compare them in the region graph, in which the differences between states in various dimensions could be easily identified. Fig. 8.3 shows an example where a subset about “Texas” and a subset about “California” were compared side-by-side. All the discoveries were related to transportation. By exploring the links and revisiting the discoveries in the visualization, the researcher easily identified the big difference between “Texas” and “California” in “registered vehicles” and “public transportation”. He also quickly captured the difference of “fuel price” in “Texas” and “California”. As a result, the researcher concluded that these factors could explain why “Texas” had much higher transportation emission than “California” even though they had similar population.

To conduct the second task, the researcher first reviewed all the *correlation* discoveries in the region graph and identified controllable factors among them. For example, "average gas price" and "share of public transportation" were important factors affecting transportation emission and could be controlled by policies. The researcher grouped all the *correlation* discoveries that contained the controllable factors and associated them with the discoveries of states with high emission in the region graph. In this way, the researcher quickly determined the controllable emission factors for these states and made the recommendation accordingly. For example, if a state with high transportation emission had very low fuel price, the researcher would suggest increasing the fuel price to reduce the transportation emission for this state. In this case study, the researcher annotated 147 discoveries and created 15 hypotheses.

8.2.4 Feedback

Overall, the researcher reported that he had enjoyed the case study. He also showed enthusiasm for ManyInsights. He commented that the discovery management functions provided in ManyInsights incorporated well into his natural analysis flow and that they helped drive him to perform in-depth analyses. He particularly liked the ease with which he was able to conduct semi-automatic discovery annotation, grouping, and association in a single system. He commented that previously he had to use multiple tools, such as a text editor, tables, and organization charts to manually record and manage discoveries. It was time-consuming to transform and share the results among these tools. ManyInsights freed him from these tedious tasks so that he could spend more time on analyzing important discoveries, detecting the hidden relationships, and conducting reasoning tasks. Moreover, the researcher was impressed by the interactivity and visual interfaces of ManyInsights, such as visually grouping, associating, and interactively browsing discoveries. He thought they were very useful for integrating the discoveries and drawing hypotheses.

Regarding to the specific components and functions, the researcher commented that the semi-automatic annotation approach was very useful and the predefined templates could fulfill his annotation needs. The researcher particularly liked the hypothesis generation function. He commented, “Previously, I would have to use the text editor to record the hypotheses and manually associate the findings to the hypotheses. It required much more efforts and I could easily lose track of the associated findings.” Moreover, the researcher pointed out that the tag function was extremely useful, especially for searching and organizing discoveries.

The researcher commented that the faceted discovery search was very intuitive and enjoyable to use. In the training session, he showed a great interest to the interface and grasped it with little instruction. In the analysis process, the researcher was able to examine the most frequent items of each attributes through the interface, which offered great convenience. He commented, “It helps me quickly keep an awareness of the analysis state at the moment, such as which datasets had been explored a lot and which one requires more explorations. Manually obtaining this information could require many efforts and distract me from the ongoing analysis.” The scented discovery browsing was similarly useful, “Every time I revisited a visualization I would first examine the small indicators to check what I had [discovered]. The function led to many unexpected findings and prevented me from making redundant annotations.”

The researcher pointed out that the region graph was incredibly useful and it was among the most frequently used tools during the study. He commented, “Overall, the region graph is a wonderful tool for summarizing large numbers of discoveries and drawing conclusions from them. The layout, the node placement and representation, and the links help me easily interpret the interrelationships and form a comprehensive understanding of the discoveries I captured.” The researcher was thrilled by the feature of simultaneously comparing the insight groups. He said, “The visual comparison is extremely useful for conducting the state-to-state comparison task. It allowed me to identify the differences and similarities quickly and effectively.”

The researcher also suggested potential future improvements. For example, he emphasized the importance of associating discoveries involving dimensions at different levels of a dimension hierarchy. For example, a yearly emission trend might provide important context for analyzing monthly or quarterly emissions. The researcher also desired a dynamic update function for the region graph so that the newly captured discoveries can be dynamically displayed and associated with existing discoveries. Other suggestions included more flexible data management functions such as splitting/merging datasets.

8.3 Asynchronous Collaboration Case Study

An important goal of managing discoveries is to facilitate collaborative analysis. To better understand the use of the proposed approaches in collaboration, we conducted a preliminary user study for asynchronous collaboration using ManyInsights. The study focused on the common ground construction and had the following specific goals: (1) to understand how users construct common ground using the discovery management functions in ManyInsights; and (2) to learn how well the various functions support their efforts in this process. The study was essentially a longitudinal analysis that was focused on small numbers of users (both experts and novice users) over long time periods. In particular, the case study consisted of two sessions, namely an individual analysis session and an asynchronous collaborative session. They were designed to simulate real world collaborative analytic procedures.

8.3.1 Procedure

We first ran a 2-week individual analysis session with 5 graduate students, all of whom were Computer Science majors and had participated in our previous user study with Click2Annotate. They were asked to explore two datasets individually using either scatterplot or parallel coordinates, and annotate their discoveries using the Click2Annotate tool. The first dataset is the NFL football season data (75 dimensions and 32 data items). Participants were instructed to discover discoveries about key factors for a football team to win more games. The second dataset is the fast food nutrition data (9 dimensions and 274 data items). Participants were instructed to discover discoveries that determine the healthiest fast food restaurant. The tools and data were installed in portable laptops so participants could perform the tasks whenever convenient. They tagged discoveries with predefined keywords or those created by themselves. The generated annotations were automatically collected. After the first session, we collected 43 discoveries for the NFL football season data and 67 discoveries for the fast food nutrition data.

Next, the asynchronous collaborative session was conducted. Participants were instructed to use the system separately to review the discoveries created in the first session. Seven graduate students attended this session, including five existing students (experts) and two novice students not participating in the first session. After training, participants were instructed to review the entire discovery collections for each dataset, followed by free exploration in which they queried and reviewed discoveries of interest. There was no time limit but all the participants completed their work within 3 hours. We observed and recorded the screen of the whole process and conducted interviews after the session.

8.3.2 Findings

Our key findings were derived from the asynchronous collaborative session with observations and users' feedback. First, we observed that author information was commonly used to organize discoveries in the initial period of common ground construction. In particular, three participants used authors to group discoveries at the beginning. Thereafter, they used region graphs looking for shared information between authors. Four participants used color encodings to distinguish discoveries generated by different authors. One participant grouped discoveries by dimensions and colored them by authors, when reviewing fast food nutrition data. He then divided five authors into three groups according to their exploration focuses. This finding answers the call for supporting role assignment in collaborative visual analysis [15].

Second, during the free exploration process, we observed that the rich set of views provided by our system allowed experts and novice users to use different exploration strategies. Most experts first searched discoveries generated by themselves and selected them on the timeline. Then, they used the timeline to navigate to a particular moment, created a group for the selected discoveries, and highlighted the group. By tracking the evolution of this group and manipulating the visual structure, they continuously added correlated discoveries into the group. They further investigated the group using either annotation cards or the region graph to browse and relate them. In this process, searching and sorting on discovery or keyword tables are the most frequent actions taken by the experts. In contrast, we noticed that novice users mostly relied on content clouds to manipulate and organize discoveries. They would spend a longer amount of time on the clouds and find important items to guide them in further exploration. However, both novice users and experts used content clouds intensively in exploring individual groups.

The feedbacks from the participants indicated that the system helped them understand and manipulate each other' discoveries, and was useful in complex collaborative tasks. When asked about specific features, participants were greatly impressed by the dynamic organization of discoveries, the abundant interactions (e.g., color coding, annotation card, and multiple selection tools), and the ability of comparing discovery groups with region graphs. One participant with Many Eyes experiences compared our system to Many Eyes, "*I really like the way to visually present and group discoveries. Even more I can change the group at will. In Many Eyes, I have to endlessly search keywords and read hundreds of posts. It is really boring*". Regarding specific tasks, one participant emphasized that grouping discoveries by data similarity was particularly powerful in understanding NFL football data, "*When I originally explored this data, It really messed me up since there is more than 70 dimensions! But after I grouped discoveries and reviewed them, I suddenly got some interesting correlations about dimensions. It really helps*".

8.4 Conclusion

In this chapter, we presented long-term case studies to evaluate the ManyInsights and its underlying framework in real-world exploratory data analysis. The study provided strong support for the usefulness of ManyInsights and its underlying discovery management framework. Two aspects of ManyInsights turned out to be particularly helpful: semi-automatic discovery annotation and flexible correlation exploration. The studies also inspired the improvements and future directions of our work, which are summarized in Chapter 9.

We also recognize the limitation of the case studies: they did not compare performance against other systems or more traditional methods. In the future, we will conduct more concrete comparison to previous methods (manually recording and associating the discoveries etc.). Benchmark datasets, such as the terrorism detection data provided by IEEE VAST contests [91], and synthesized datasets with embedded discoveries and hypotheses will be used for controlled result comparison. Since the discovery management has a very broad application domain, we will involve more experts from different application domains in long-term evaluation of ManyInsihts. Choosing more applications would increase the confidence in the results and provide a deeper understanding of the impacts of the framework. Finally, we will publish ManyInsights online for public tests. We will collect user feedbacks to evaluate its utility, usability, and scalability, and thus refine the system.

CHAPTER 9: CONCLUSION

This final chapter contains concluding remarks about the work presented in this dissertation. First, we review the main contributions of this dissertation. Secondly, we discuss opportunities for future work.

9.1 Review of Dissertation Contributions

This dissertation identifies an emerging gap between existing visual analytics systems and effective decision making: decision making often involves the annotation, browsing, retrieval, organization, association, and sharing of large amounts of discoveries; few of visual analytics systems provide general and scalable solutions to support these discovery management activities. In response, this dissertation contributes a general framework, novel techniques, and a system to bridge this gap.

The key principles of discovery management, introduced in Chapter 1, were *looking forward and looking backward* and *constructing common ground*. They aimed to support dynamic knowledge construction and collaborative visual analytics, respectively. We also identified a set of discovery management activities that are essential for supporting the principles, including discovery annotation, browsing, retrieval, organization, association, and sharing.

To support these activities, we contributed a general discovery management framework in Chapter 3. In this framework, we introduced the pattern as core concept to achieve the effectiveness and efficiency of discovery management. We contributed the core idea of using pattern taxonomy to enhance the automation and effectiveness of different discovery management activities. Based on this idea, we explored a visual exploration paradigm to integrate the discovery management activities with interactive visual exploration. Using this taxonomy and the paradigm, we contributed a variety of discovery management techniques:

Taxonomy: We constructed a pattern taxonomy for multidimensional data (Chapter 4). The taxonomy provides a categorization for the vast number of discoveries in multidimensional datasets and defines their common characteristics. It provides a solid foundation for all the discovery management techniques.

Annotation: We proposed a novel discovery annotation approach Click2Annotate (Chapter 5) which allows users to generate high quality discovery annotations with reduced efforts. We contributed annotation template techniques for automatically retrieving context of discoveries from data and generating highly formalized and semantically rich annotations based on the information. We also contributed multiple interactive techniques to modify and refine the annotations. Finally, we conducted a user study to evaluate Click2Annotate and found that it could enhance annotation efficiency and improve the quality of annotations.

Browsing and Retrieval: We developed two techniques to support flexible discovery browsing and retrieval (Chapter 6). The faceted discovery search which was informed by the faceted search allows users to flexibly search discoveries using their rich context and visually explored their semantics. The scented discovery browsing seamlessly integrates discoveries with interactive data visualization and provides substantial flexibility to browse and explore them.

Correlation Exploration and Sensemaking: We also contributed a visual analytics approach to help users explore the correlations among discoveries. Our approach enables automatic discovery gathering, organization, and association (Chapter 7) and provides a rich set of visualization and interaction techniques to help users review, explore, and compare discoveries in detail. In addition, hypothesis generation techniques was introduced to facilitate sensemaking and common ground construction tasks. We also conducted user studies to evaluate the effectiveness and efficiency of the approach in different analysis environments using different datasets.

The above techniques were implemented in a prototype system, ManyInsights, for managing discoveries in multidimensional data. We contributed two long-term case studies to evaluate ManyInsights using real analytic tasks and and real datasets (Chapter 8). The case studies focused on understanding the collaboration among the discovery management techniques in dynamic knowledge construction process and asynchronous collaborative visual analysis. The results provided strong support for the usefulness of ManyInsights and its underlying discovery management framework.

In conclusion, this dissertation is significant in the fields of information visualization and visual analytics due to the following reasons:

- It provides a general framework that explores taxonomy + exploration paradigm +scalable techniques discovery management solution to bridge the gap between existing visual analytics systems and decision making;
- The proposed taxonomy is among the first taxonomies of patterns in the fields of information visualization and visual analytics. It provides a foundation for future search of discovery management;
- The *looking forward and looking backward* and *constructing common ground* principles break new ground in large-scale data exploration research. It has the potential to be used in a wide range of applications;
- The dissertation contains many standalone, innovative ideas such as semi-automatic discovery annotation and automatic discovery correlation calculation;
- The prototype ManyInsights is among the first efforts towards effective and efficient discovery management in multidimensional data exploration; and
- The long-term case studies suggest new evaluation metrics and methods for conducting experiments for discovery management.

9.2 Future Work

Based on the contributions of the work, this dissertation also promises new research opportunities for current visual analytics research. As suggested by our literature survey in Chapter 2 and domain expert study in Chapter 8, there is a great deal of future work that can be done to enrich and improve the functionality of the discovery management framework. Here we elaborate some of the limitations of this dissertation and corresponding future work:

Division Construction: In this dissertation, we assume that high dimensional datasets have been partitioned to multiple subsets small enough to be explored by existing visualization techniques for detecting patterns. However, in real-world applications, this is often a challenging work, especially when combining the existing subspace construction techniques (e.g., [92]) with the proposed discovery management activities. For example, partitioning a high dimensional data might break a discovery or its context into pieces. Missing any of these pieces might lead users to draw incomplete conclusions or wrong hypotheses. Moreover, current division construction approaches might prevent users from grasping an overview of datasets, resulting the unawareness of the interrelations between discoveries in different generated divisions. To address these problems, new division construction approaches will be explored. The new approaches will combine the advanced data mining techniques, such as feature selection [93], with the proposed pattern taxonomy to automatically partition the dataset according to the characteristics of the discoveries. As a result, the information loss is reduced and the completeness of the discoveries is maintained. It will also employ state-of-art visualization techniques to help users dynamically construct divisions according to their diverse exploration focuses and visually convey the interrelations of discoveries in different divisions.

Guided Pattern Discovery: Recommendation and subscribe/publish mechanisms have been widely used in online systems for online shopping [94] and broadcast services [71]. However, few of these ideas have been used in the area of multidimensional data exploration. We argue that integrating these innovative ideas into the proposed discovery management framework will benefit the dynamic knowledge construction from massive, high dimensional datasets. In particular, we propose guided pattern discovery in discovery management framework which is supported by two specific techniques. Pattern notification services can be provided to automatically keep track of patterns registered by users so that they do not need to keep the discoveries in mind. The users will be notified if a new pattern is discovered that is related to a registered pattern. In addition, when a user meets a potential pattern, such as a brushed data cluster, which is related to a registered pattern, the system will automatically notify the user about the situation. During the visual exploration process, pattern recommendation services automatically or semi-automatically recommend views containing potential patterns of interest to users according to registered patterns or user requirements.

New Application and Evaluation: Finally, the proposed multidimensional exploration system, ManyInsights, will be developed in parallel with a wide variety of domain specific applications, such as health and food, census, and stock analysis. Upon the knowledge of these domains, external ontologies will be introduced to enhance the automation of various discovery management activities. The effectiveness of ManyInsights will be investigated through case studies in the various applications. Since one of the primary goals of the framework is to support collaboration in visual analytics, we will publish ManyInsights online for public tests. We will collect user feedbacks to evaluate its utility, usability, and scalability, and thus refine our system. Eventually, we will promote it to a variety of realistic applications. Although the focus of our current research is discovery management in multidimensional data exploration, we realize that there are urgent needs for supporting dynamic knowledge construction in a wide range of massive data exploration domains, such as text, graph, and geospatial data. Therefore, we will also extend the discovery management framework to support a wider range of data types. For example, pattern taxonomy for graph visualization will be constructed to facilitate effective discovery management in graph analysis.

REFERENCES

[1] L. Zhang, A. Stoffel, M. Behrisch, S. Mittelstädt, T. Schreck, R. Pompl, S. Weber, H. Last, and D. A. Keim, "Visual analytics for the big data era - a comparative review of state-of-the-art commercial systems," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 173–182, 2012.

[2] J. Thomas and K. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* IEEE Computer Society Press, 2005.

[3] D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," *Proc. Visual Data Mining*, pp. 76–90, 2008.

[4] E. Hetzler and A. Turner, "Analysis experiences using information visualization," *IEEE Computer Graphics and Application*, vol. 24, no. 5, pp. 22–26, 2004.

[5] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: Visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9–20, 2002.

[6] B. Johnson and B. Shneiderman, "Tree-maps: A space-filling approach to the visualization of hierarchical information structures," *Proc. IEEE Visualization*, pp. 275–282, 1991.

[7] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou, "Treejuxtaposer: scalable tree comparison using focus+context with guaranteed visibility," *Proc. ACM SIGGRAPH Conference*, pp. 453–462, 2003.

[8] M. Ward, "Xmdvtool: Integrating multiple methods for visualizing multivariate data," *Proc. IEEE Visualization*, pp. 326–333, 1994.

[9] C. Stolte and P. Hanrahan, "Polaris: A system for query, analysis and visualization of multi-dimensional relational databases," *Proc. IEEE Symposium on Information Visualization*, pp. 5–14, 2000.

[10] J. Heer and M. Agrawala, "Design considerations for collaborative visual analytics," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 171–178, 2007.

[11] P. Isenberg, D. Fisher, M. Morris, K. Inkpen, and M. Czerwinski, "An exploratory study of co-located collaborative visual analytics around a tabletop display," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 179–186, 2010.

[12] H. Chung, S. Yang, N. Massjouni, C. Andrews, R. Kanna, and C. North, "Vizcept: Supporting synchronous collaboration for constructing visualizations in intelligence analysis," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 107–114, 2010.

[13] F. Viégas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon, "Many eyes: a site for visualization at internet scale," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1121–1128, 2007.

[14] M. Wattenberg, "Baby names, visualization, and social data analysis," *Proc. IEEE Symposium on Information Visualization*, pp. 1–7, 2005.

[15] A. Robinson, "Collaborative synthesis of visual analytic results," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 67–74, 2008.

[16] H. Clark and S. Brennan, "Grounding in communication. in perspectives on social shared cognition." *American Psychological Association*, pp. 127–149, 1991.

[17] R. Amar, J. Eagan, and J. Stasko, "Low-level components of analytic activity in information visualization," *Proc. IEEE Symposium on Information Visualization*, pp. 111–147, 2005.

[18] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," *Proc. of ACM SIGMOD Conference on Management of Data*, pp. 94–105, 1998.

[19] C. Yu, S. Bressan, B. C. Ooi, and K.-L. Tan, "Querying high-dimensional data in single-dimensional space," *The International Journal on Very Large Data Bases*, vol. 13, no. 2, pp. 105–119, 2004.

[20] A. Tung, R. Zhang, N. Koudas, and B. C. Ooi, "Similarity search: a matching based approach," *Proc. International Conference on Very Large Data Bases*, pp. 631–642, 2006.

[21] A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multidimensional geometry," *Proc. IEEE Visualization*, pp. 361–378, 1990.

[22] W. Cleveland and M. McGill, *Dynamic Graphics for Statistics.* Wadsworth, Inc, 1988.

[23] F. Bendix, R. Kosara, and H. Hauser, "Parallel sets: Visual analysis of categorical data," *Proc. IEEE Symposium on Information Visualization*, pp. 18–25, 2005.

[24] T. Kapler and W. Wright, "Geotime information visualization," *Proc. IEEE Symposium on Information Visualization*, pp. 25–32, 2004.

[25] C. Weaver, "Building highly-coordinated visualizations in improvise," *Proc. IEEE Symposium on Information Visualization*, pp. 159–166, 2004.

[26] K. Hori, "Do knowledge assets really exist in the world and can we access such knowledge?" *Intuitive Human Interfaces for Organizing and Accesing Intellectual Assets*, pp. 1–13, 2004.

[27] M. Zack, "Managing codified knowledge," *Sloan Management Review*, vol. 40, no. 4, pp. 45–58, 1999.

[28] G. Gavetti and D. Levinthal, "Looking forward and looking backward: Cognitive ad experiential search," *Administrative Science Quarterly*, pp. 113–137, 2000.

[29] K. Weick, K. Sutcliffe, and D. Obstfeld, "Organizing and the process of sense-making," *Organization Science*, vol. 16, no. 4, pp. 409–421, 2005.

[30] G. Convertino, H. Mentis, M. Rosson, J. Carrol, A. Slavkoci, and C. Ganoe, "Articulating common ground in cooperative work: content and process," *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1637–1646, 2008.

[31] G. Convertino, C. Ganoe, W. Schafer, B. Yost, and J. Carrol, "A multiple view approach to support common ground in distributed and synchronous geo-collaboration," *Proc. Coordinated and Multiple Views in Exploratory Visualization*, pp. 121–132, 2005.

[32] P. Saraiya, C. North, V. Lam, and K. Duca, "An insight-based longitudinal study of visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1511–1522, 2006.

[33] Y. Kang, C. Görg, and J. Stasko, "How can visual analytics assist investigative analysis? design implications from an evaluation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 5, pp. 570–583, 2011.

[34] N.Mahyar, A. Sarvghad, and M. Tory, "A closer look at note taking in the co-located collaborative visual analytics process," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 171–178, 2010.

[35] D. Keim and H.-P. Kriegel, "Visualization techniques for mining large databases: A comparison," *IEEE Transactions Knowledge and Data Engineering*, vol. 8, no. 6, pp. 923–938, 1996.

[36] M. Chuah and S. Roth, "On the semantics of interactive visualizations," *Proc. IEEE Symposium on Information Visualization*, pp. 365–372, 1996.

[37] A. Dix and G. Ellis, "Starting simple: adding value to static visualisation through simple interaction," *Proc. Working Conference on Advanced Visual Interfaces*, pp. 124–134, 1998.

[38] M. Ward and J. Yang, "Interaction spaces in data and information visualization," *Proc. IEEE TCVG Symposium on Visualization*, pp. 137–145, 2004.

[39] J. Yi, Y. Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[40] S. Card and J. Mackinlay, “The structure of the information visualization design space,” *Proc. IEEE Symposium on Information Visualization*, pp. 92–99, 1997.

[41] E. Chi, “A taxonomy of visualization techniques using the data state reference model,” *Proc. IEEE Symposium on Information Visualization*, pp. 69–75, 2000.

[42] D. Keim, M. Hao, J. Ladisch, M. Hsu, and U. Dayal, “Pixel bar charts: A new technique for visualizing large multi-attribute data sets without aggregation,” *Proc. IEEE Symposium on Information Visualization*, pp. 113–120, 2001.

[43] D. Gotz and M. Zhou, “Characterizing users’ visual analytic activity for insight provenance,” *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 123–130, 2008.

[44] A. Leontev, *Activity, Consciousness, and Personality.* Prentice-Hall, 1978.

[45] B. Shneiderman, “The eyes have it: a task by data type taxonomy for information visualization,” *Proc. IEEE Symposium on Visual Languages*, pp. 336–343, 1996.

[46] S. Wehrend and C. Lewis, “A problem-oriented classification of visualization technique,” *Proc. 1st Conference on Visualization*, pp. 139–143, 1990.

[47] M. Zhou and S. Feiner, “Automated visual presentation: From heterogenous information to coherent visual discourse,” *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 392–399, 1998.

[48] B. Lee, C. Parr, J. Fekete, and N. Henry, “Task taxonomy for graph visualization,” *Proc. AVI Workshop on Beyond Time and Errors*, pp. 1–5, 2006.

[49] K. Brodlie, A. Poon, H. Wright, L. Brankin, G. Banecki, and A. Gay, “Grasparc: A problem solving environment integrating computation and visualization,” *Proc. IEEE Visualization*, pp. 102–109, 1993.

[50] K. Ma, “Image graphs - a novel approach to visual data exploration,” *Proc. IEEE Visualization*, pp. 81–88, 1999.

[51] Y. Shrinivasan and J. van Wijk, “Supporting the analytical reasoning process in information visualization,” *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1237–1246, 2008.

[52] M. Elias and A. Bezerianos, “Annotating bi visualization dashboards: needs & challenges,” *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1641–1650, 2012.

[53] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, B. Cort, and O. I. Inc, “The sandbox for analysis concepts and methods,” *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, 2006.

[54] J. Heer, F. Viegas, and M. Wattenberg, "Voyagers and voyeurs: Supporting asynchronous collaborative information visualization," *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1029–1038, 2007.

[55] S. Ellis and D. Groth, "A collaborative annotation system for data visualization," *Proc. Working Conference on Advanced Visual Interfaces*, pp. 411–414, 2004.

[56] D. Yang, E. Rundensteiner, and M. Ward, "Analysis guided visual exploration of multivariate data," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 83–90, 2007.

[57] W. Willett, J. Heer, J. Hellerstein, and M. Agrawala, "Commentspace: Structured support for collaborative visual analysis," *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, 2011, to appear.

[58] Y. Shrinivasan, D. Gotz, and J. Lu, "Connecting the dots in visual analysis," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 123–130, 2009.

[59] "Analyst's notebook," http://i2group.com/Analysts-Notebook.

[60] D. Yang, Z. Xie, E. Rundensteiner, and M. Ward, "Managing discoveries in the visual analytics process," *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 2, pp. 22–29, 2007.

[61] T. Chklovski, V. Ratnakar, and Y. Gil, "User interfaces with semi-formal representations: a study of designing argumentation structures," *Proc. ACM Conference on Intelligent User Interfaces*, pp. 130–136, 2005.

[62] D. Billman, G. Gonvertino, J. Shrager, P. Pirolli, and J. Massar, "Collaborative intelligence analysis with cache and its effects on information gathering and cognitive bias," *Proc. HCI Consortium Workshop*, 2006.

[63] J. Stasko, C. Görg, Z. Liu, and K. Singhal, "Jigsaw: Supporting investigative analysis through interactive visualization," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 131–138, 2007.

[64] N. Kadivar, V. Chen, D. Dunsmuir, E. Lee, C. Qian, J. Dill, C. Shaw, and R. Woodbury, "Capturing and supporting the analysis process," 2009, pp. 131–138.

[65] M. Chuah and S. Roth, "Visualizing common ground," *Proc. International Conference on Information Visualization*, pp. 365–372, 2003.

[66] W. Jones, H. Bruce, and S. Dumais, "Keeping found things found on the web," *Proc. ACM CIKM International Conference on Information and Knowledge Management*, pp. 119–126, 2001.

[67] C. North, "Visualization viewpoints: Toward measuring visualization insight," *IEEE Computer Graphics Applications*, vol. 26, no. 3, pp. 6–9, 2006.

[68] J. Yi, Y. Kang, J. Stasko, and J. Jacko, "Understanding and characterizing insights: how do people gain insights using information visualization?" *Proc. conference on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, 2008.

[69] Z. Pousman and J. Stasko, "Casual information visualization: Depictions of data in everyday life," *Proc. IEEE Symposium on Information Visualization*, pp. 1145–1152, 2007.

[70] M. Merrill, "Knowledge objects and mental models," *Proc. International Workshop on Advanced Learning Technologies*, pp. 244–246, 2000.

[71] "Youtube," http://www.youtube.com.

[72] "Flex," http://www.adobe.com/products/flex.html.

[73] D. Keim, "Information visualization and visual data mining," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, 2002.

[74] H. Gauch, Ed., *Multivariate Analysis in Community Ecology.* Cambridge University Press, 1982.

[75] J. Han and M. Kamber, *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, 2006.

[76] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *Proc. ACM SIGMOD Conference on Management of Data*, pp. 93–104, 2000.

[77] D. Moore, *Basic Practice of Statistics.* WH Freeman Company, 2006.

[78] L. Wilkinson, A. Anand, and R. Grossman, "High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1363–1372, 2006.

[79] P. Greenwood and M. Nikulin, *A guide to chi-squared testing.* John Wiley and Sons, 1996.

[80] D. Goncalves and J. Jorge, "In search of personal information: Narrative-based interfaces," *Proc. ACM Conference on Intelligent User Interfaces*, pp. 179–188, 2008.

[81] K. Yee, K. Swearingen, K. Li, and M. Hearst, "Faceted metadata for image search and browsing," *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 401–408, 2003.

[82] H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. Keim, and O. Deussen, "Document cards: A top trumps visualization for documents," *Proc. IEEE Symposium on Information Visualization*, pp. 1145–1152, 2009.

[83] L. Hong and E. Chi, "Annotate once, appear anywhere: Collective foraging for snippets of interest using paragraph fingerprinting," *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1791–1794, 2009.

[84] J. Carroll, M. Rosson, G. Convertino, and C. Ganoe, "Awareness and teamwork in computer-supported collaborations," *Interacting with Computers*, vol. 18, no. 1, pp. 21–46, 2005.

[85] J. Alsakran, Y. Chen, Y. Zhao, J. Yang, and D. Luo, "Streamit: Dynamic visualization and interactive exploration of text streams," *Proc. IEEE Symposium on Pacific Visualization*, 2011.

[86] G. Salton and M. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill Press, 1986.

[87] F. Wei, S. Liu, Y. Song, S. Pan, M. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, "Tiara: a visual exploratory text analytic system," *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 153–162, 2010.

[88] B. Shneiderman and A. Aris, "Network visualization by semantic substrates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 733–740, 2006.

[89] "Many-eyes," http://www.many-eyes.com.

[90] T. Fruchterman and E. Reingold, "Graph drawing by force-directed placement," *Software: Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[91] M. Whiting, C. North, A. Endert, J. Scholtz, J. Haack, C. Varley, and J. Thomas, "Vast contest dataset use in education," *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 115–122, 2009.

[92] J. Yang, M. Ward, E. Rundensteiner, and S. Huang, "Visual hierarchical dimension reduction for exploration of high dimensional datasets," *Proc. IEEE TCVG Symposium on Visualization*, pp. 19–28, 2003.

[93] J. Dy and C. Brodley, "Feature selection for unsupervised learning," *Journal of Machine Learning Research*, vol. 5, pp. 845–889, 2004.

[94] "Amazon," http://www.amazon.com.