Toward Effective Insight Management in Visual Analytics Systems

Yang Chen* Jing Yang† William Ribarsky‡

Department of Computer Science
UNC Charlotte

ABSTRACT

Although significant progress has been made toward effective insight discovery in visual sense making approaches, there is a lack of effective and efficient approaches to manage the large amounts of insights discovered. In this paper, we propose a systematic approach to leverage this problem around the concept of facts. Facts refer to patterns, relationships, or anomalies extracted from data under analysis. They are the direct products of visual exploration and permit construction of insights together with user's mental model and evaluation. Different from the mental model, the type of facts that can be discovered from data is predictable and application-independent. Thus it is possible to develop a general Fact Management Framework (FMF) to allow visualization users to effectively and efficiently annotate, browse, retrieve, associate, and exchange facts. Since facts are essential components of insights, it will be feasible to extend FMF to effective insight management in a variety of visual analytics approaches. Toward this goal, we first construct a fact taxonomy that categorizes various facts in multidimensional data and captures their essential attributes through extensive literature survey and user studies. We then propose a conceptual framework of fact management based upon this fact taxonomy. A concrete scenario of visual sense making on real data sets illustrates how this FMF will work.

Keywords: Visual Analytics, Decision Making, Taxonomy, Knowledge Management, Multidimensional Visualization.

Index Terms: H.5.0 [Information Interfaces and Presentation]: General;

1 INTRODUCTION

Recently, a burst of visual analytics approaches have been developed to support analytical reasoning facilitated by interactive visual interfaces. In many of these approaches, insights are captured from interactive visual exploration and used for supporting high level hypothesis generation and evaluation toward problem solving and decision making. A significant challenge faced by these approaches is that large amounts of insights are often involved in the sense making process. As a consequence, Insight Management (IM), such as insight recording, association, retrieval, and exchange, becomes essential for effective visual analytics approaches.

Although quite a few efforts have been made towards managing insights in visual analytics systems, most existing approaches suffer from the following problems: (1) they require manual insight annotation, such as manually posting discoveries [27] or attaching hand-drawn marks to the visualization views [15]. Manual annotation is time-consuming and thus reduces users' interests in annotating insights. Moreover, it can be incomplete, imprecise, and hard to understand. It leads to difficulties in the following IM activities such as insight retrieval and exchange; (2) most existing approaches require users to manually detect relationships among insights, such as to manually create a knowledge view [26] or inference network [7]. We argue that such manual approaches do not scale to sense making processes where a large amount of insights, long analysis time, or multiple analysts are involved; (3) it is hard to search and reuse recorded insights in existing approaches, especially when applying to distributed data sets in an asynchronous collaboration environment. It is often hard to construct a query to effectively fetch stored insights since different users may use various terms to express similar meanings when manually annotating insights. Users often need to examine insights one by one to find the ones of their interests. It is often hard for users to understand insights captured and annotated by others since the annotation process is not well regulated; (4) insight exchange in collaborative visual analytics is not effectively supported since most approaches heavily rely on users to manually search and understand insights provided by their collaborators.

Effectiveness and efficiency of insight management will be hard to achieve without solving the above challenges. To address these problems, a close look must be taken at what is an insight. Insights have been defined and discussed in many papers. Amar et al. [3] equated insights with user tasks, such as finding extreme values or detecting outliers. North [20] pointed out that such a definition is informal and he summarized five important characteristics for insights: complex, deep, qualitative, unexpected, and relevant. Gersh et al. [10] proposed that an insight "is about something and it is based on something" and has three basic components: a set of information items, a collecting specification that describes how the information items were gathered, and descriptive annotation to express the insight. Yi et al. [33] argued that insights are more likely a by-product of exploration without an initial destination. From the above previous work, we observe that an insight is a complex concept that is associated with not only data under analysis, but also objective and subjective evaluations about the significance and confidence of the data based on real-world knowledge which is stored in user's mental model [18]. We thus propose a three-components model to describe an insight: a fact extracted from data under analysis, such as an outlier, a pattern, or a relationship, a mental model upon which the fact is evaluated, and objective and subjective evaluations of the fact. In a typical case, an analyst discovers a fact as a result of a user task during an interactive visual exploration process. She then evaluates the fact against her mental model to see if it is a significant and reliable piece of evidence that can be used in the sense making process. The fact, the mental model applied, and the evaluations construct an insight for the sense making process.

Among the three components, the mental model is hard to be handled using a general approach since it varies a lot among data sets, applications, and analysts. On the other hand, the type of facts that can be discovered from data is predictable and independent from data sets, applications, and analysts. In addition, facts are direct products of user interactions in the visual exploration process and thus their management can be tightly integrated into the visualization system. We argue that a general *Fact Management Framework* (FMF) can be developed to allow visualization users to effectively and efficiently detect, annotate, associate, retrieve,

*e-mail: ychen61@uncc.edu
†email: jyang13@uncc.edu
‡ribarsky@uncc.edu

and exchange facts using automatic or semi-automatic approaches. Since facts are the fundamental components of insights and bridge the visual exploration process and IM, it will be feasible to extend the general FMF to IM in various visual analytics applications by adding real-world knowledge from mental model and evaluation management upon the FMF and thus lead to effective and efficient IM.

In this paper, we present our preliminary work towards such a general FMF. First, we construct fact taxonomy, using multidimensional data as an example. This taxonomy categorizes a large variety of facts that can be discovered from multidimensional data and characterizes the essential information of the facts that can be used for enhancing automation in fact and insight management activities such as annotation, indexing, retrieval, association, and exchange. Using the taxonomy as a common language, we propose a conceptual framework of fact management that includes semi-automatic fact annotation, retrieval, association, and exchange among other fact management activities. A concrete scenario of visual sense making on real data sets is then given to illustrate the management activities in the proposed FMF.

The major contributions of this paper are:

- the major challenges faced by existing insight management approaches are recognized and a solution toward addressing the challenges, namely the fact management framework, is proposed and illustrated using a concrete scenario;
- a fact taxonomy for multidimensional data is constructed and presented to provide a solid foundation for the proposed FMF work. The taxonomy is constructed through extensive literature survey, user studies, and field studies;
- a concrete scenario is given to illustrate how to use the proposed fact taxonomy and FMF.

The rest of the paper is organized as follows: section 2 discusses related work; section 3 illustrates our approaches to constructing the fact taxonomy; section 4 presents the constructed taxonomy; section 5 proposes the conceptual framework of fact management; section 6 provides a concrete scenario on using the FMF; section 7 presents our conclusion and future work.

2 Related Work

2.1 Insight Management

Robinson [22] provides experimental evidence that effective management of visual analytic results, such as annotating, organizing, and sharing user's finding results, is one of critical aspects of collaborative synthesis. There exist a few visualization systems that allow users to capture, store, retrieve, and share discovered insights. For example, Many-Eyes [27] provides a discussion forum where users can share their findings or free thoughts on visualizations by posting comments. A URL bookmarking mechanism is used to point back from the comments to the associated views so that users can revisit and evaluate their findings. A drawback of this forum is that it does not provide sufficient aids to help users create, associate, organize, and retrieve comments. Ellis and Groth [8] use annotations to share discoveries in their collaborative data visualization environment. Unfortunately, users need to create the annotations manually and find insights of their interests manually. Shrinivasan and van Wijk [26] enable 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. However, this approach also requires manual note taking and grouping. Systems such as Sandbox [30] and Name Voyagers [15] try to leverage user's efforts in insight recording and retrieval by allowing users to jot down their observations and opinions into visualization views in collaborative annotations, but it is difficult to express the interrelations among the annotations within different visualization views in these systems.

Recently, a few initial efforts have been made to take advantage of automatic analysis and visual exploration techniques to manage insights. The work closest to our approach so far is the Nugget Management System proposed by Yang et al. [32]. It allows users to extract, refine, and record nuggets (subpart of multivariate data) with the help of automatic analysis techniques. Statistical information, such as the number of data records included and average values on each dimension can be automatically computed and attached to a discovered nugget in addition to manual annotation given by users. Currently this system only supports the discovery and annotation of clusters in multivariate data. A more extensive range of insight types and insight management activities is yet to be considered. HARVEST [12] automatically manages user defined concepts and their evidences. However, user has to manually extract the concepts and evidences from unstructured information.

Many efforts have been made toward domain-specific insight management tools. For example, Xiao et al. [31] propose a knowledge representation approach to save and reuse discoveries from network traffic data. Sandbox [30] allows users to use automatic process model templates to collect and organize evidences discovered from document data. COPLINK [6] allows users to capture, analyze and share law enforcement entities and visually explore the relationships among multiple law enforcement. Schneider et al. [24] propose a novel method to gather, organize, and share useful information about entities in a terrorism knowledge base. Bier et al. [4] develop a document browser that supports organizing, collaboratively recommending, and sharing entities captured from documents. All the above systems are designed for specific domain applications and are hard to be used in other applications.

2.2 Taxonomy

Researchers have made efforts on classifying facts in specific application domains. For example, Rester et al. [21] classify facts in psychotherapeutic treatment documents. Saraiya et al. [23] categorize findings from users' short term and long term exploration of microarray data. Xiao et al. [31] provide a list of network behaviors that are likely to be observed in network traffic analysis. To the best of our knowledge, our fact taxonomy is among the first general taxonomies that categorize facts for multidimensional data independent of any application domains.

Since users often gain insights by performing analytic tasks, our fact taxonomy is closely tied to existing taxonomic work on user tasks. Among existing task taxonomies, there are Shneiderman's task by data type taxonomy [25], Wehrend and Lewis' congitive task taxonomy [28], Zhou and Feiner's low level visualization system tasks [34], Lee et al.'s graph exploration tasks [17], Amar and Stasko's low level analytic task taxonomy for multidimensional data with analytic goals [3], and Gotz and Zhou's action taxonomy [11] considered. Our fact taxonomy for multidimensional data is strongly tied to the above work, with a different focus on characterizing the facts resulted from the analytic tasks. In addition, our taxonomy is different in that it is constructed for serving an explicit goal of effective insight management that is not considered by the other work.

3 Taxonomy Construction

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

1. *Completeness:* the taxonomy should cover the majority of facts that can be discovered using various visualization tools

and from multidimensional data sets of various sizes and dimensionalities in different application domains.

2. *Unambiguous:* the taxonomy should accurately and clearly distinguish different types of facts.

3. *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 facts, and the users who discover the facts.

4. *Utility:* the taxonomy should be feasible to use in fact and insight management.

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

1. *a literature survey on existing visualization taxonomy work and visualization techniques was conducted to generate an initial fact categorization;*

2. *the initial categorization was evaluated and refined through an experiment and a user study using real insights from real users;*

3. *interviews of domain experts were conducted to further evaluate the categorization and to learn the attributes of facts that are essential in their insight management tasks.*

4. *a literature survey on existing statistical and data mining work was conducted to summarize essential attributes for each category of facts.*

3.1 Literature Survey for Categorizing Facts

We constructed an initial fact 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 fact taxonomy among all taxonomy work since there is a strong tie between facts and visual analytic tasks: users often discover facts from visualizations by performing visual analytic tasks, i.e., visual analytic tasks are the analytical processes and facts 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 facts that can be discovered using the techniques under discussion in them.

After this turn of literature review, we constructed an initial fact categorization that captures the results of most tasks considered in the task taxonomies and covers most facts 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 3.2 for more details), two other categories, namely *compound fact* and *meta fact*, were added. We define rows in a multidimensional data sets as items and columns in it as dimensions. Most categories of facts exist in both the item space and the dimension space. For each category we gave a formal definition in Table 3, along with examples extracted from real user insights posed in Many-Eyes [27].

3.2 Experiment and User Study for Evaluating Fact 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 facts. As a consequence, the initial fact categorization needs to be evaluated and refined with facts from a diversity of real users, real data sets, and real tasks. Toward this goal, we sampled facts discovered by users of Many-Eyes [1] and conducted an experiment and a user study.

Many-Eyes [1] 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 insights are reported daily as comments by a large number of users ranging from scientists, managers, to sports fans [27]. These insights 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 insights can be examined since visualization is attached to each comment. We thus considered Many-Eyes comments as a good source of facts 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 facts embedded in them. For duplicative facts that have same data elements and same categories, we just picked out one of them. Facts about data types other than multidimensional data were also removed. As the result, we got a sample containing 215 facts which were collected from 56 multidimensional data sets. Some data sets contained temporal and geographical dimensions.

3.2.1 Experiment for Completeness Testing

An experiment was conducted to examine if the initial categorization covered the majority of the facts contained in the Many-Eyes sample. In particular, we reviewed all 215 facts and tried to fit them into the fact categorization. For example, the fact "big drop in males becoming eye doctors in the past ten years" was classified into the *trend* category and the fact 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 facts falling into each category.

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

- **Compound facts:** there were 46 facts that were facts about facts. For example, the fact "it's interesting how different the second letter distribution is from the first letter distribution" contains a *difference* fact about two *distribution* facts.

- **Facts about meta data:** there were 17 facts about data itself such as missing values or errors in the data sets, appearance or disappearance of dimensions, and meanings of labels. For example, the fact that "a change happened between 1999 and 2000 when a bunch of new categories showed up" was about the appearance of new dimensions. The fact 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 fact* and *meta fact* to fit those facts in. In addition, we decomposed each compound fact into multiple elementary facts and counted them not only in the *compound fact* category, but also in the elementary fact categories. Table 2 shows the final result. In this table, categories are sorted according to the total number of related facts in the Many-Eyes sample.

Table 1: Result of comments classification

Knowledge type	Number of comments	Percentage
Trend	55	25.6%
Compound fact	46	21.4%
Outliers	41	19.1%
Difference	31	14.4%
Association	27	12.6%
Extreme	25	11.6%
Meta fact	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%

3.2.2 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 facts into the fact 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 fact category was explained and fact 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 facts that were randomly sampled from the Many-Eyes facts one after another. The classification results and the time spent for each fact 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 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 fact was 223 seconds and 360 seconds respectively. It indicated that the subjects were able to make the classification without much effort.

3.3 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 facts sampled from specific application domains, and (2) to determine which information about facts 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 facts collected in their analytic tasks. The participant was also asked to provide a list of attributes about the facts that were important for their analytic tasks. Towards the end of the interview, our existing fact categorization was explained and the participant was asked to classify his/her reported facts into existing categories. When the participant encountered any facts that did not fit, the facts were placed in a list for future analysis. Extensive field notes were taken during the interview. Some participants provided screenshots or hyperlinks to example facts after the interview.

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

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

- *Content*: This category includes information characterizing the content of facts, such as sizes and averages of clusters, values of anomalies, and names of correlated dimensions.

- *Context*: This category includes information capturing the context of facts. For example, the distribution of the whole data sets provides a context to an *outliers* fact. The significance of most facts can only be evaluated among their contexts. Quality is a special context attribute of facts. Many participants suggested that quality information is important since it helps them index, retrieve, and filter facts.

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

3.4 Literature Survey for Summarizing Fact Attributes

A literature survey has been conducted on statistics and data mining textbooks [9, 14, 5, 19] to learn what information should be captured as content and context attributes for different categories of facts. The attribute lists collected from the domain expert interviews were also referenced. The essential content and context attributes for each category are listed in Table 3.

4 Resulting Fact Taxonomy

The constructed fact taxonomy, which includes the categorization, formal definition, examples, content attributes, and context attributes, is presented in Table 3. Table 3 also shows tasks related to each fact category for users' reference. In the following section, we will present how a fact management framework is constructed based on the fact taxonomy.

5 FACT MANAGEMENT FRAMEWORK

In this section, we propose the conceptual framework of fact management toward the following utility goals:

1. To keep found things found [16], i.e., to allow users to capture, annotate, retrieve, and inspect discovered facts;
2. To reveal relationships among detected facts and allow users to interactively explore the relationships;
3. To guide users to discover facts from massive data sets according to their exploratory goals and existing knowledge, such as to help a user discover facts for or against a given hypothesis, or related to previously found facts;
4. To aid collaborative workers in sharing and exchanging facts.

Toward these goals, a set of fact management activities are proposed in our FMF based on the fact taxonomy. They are described in detail as follows:

Semi-Automatic Fact Annotation: Effective fact annotation summarizes high level knowledge of the facts, such as their categories, contents, and contexts. Annotations allow users to organize, browse, retrieve, associate, and exchange facts using keywords in them. We propose a semi-automatic insight annotation approach based on the fact taxonomy. The fact taxonomy will suggest what should be included in an annotation for a certain type of facts. In particular, after a fact is distinguished by visualization through interactions (such as brushing) and its fact 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 fact category listed in the taxonomy. The automatically extracted information will be used to annotate the fact. Users will be allowed to interactively improve the automatically generated annotations for more flexibility.

Fact Organization, Indexing, Browsing, and Retrieval: When the system automatically generates annotations, the same vocabulary will be used for all facts and thus the facts will be easily organized, indexed, browsed, and retrieved using keywords in their annotations, as if the way that tags are used in YouTube [2]. For example, we can allow users to search facts by example facts or by keywords, or visually browse and explore facts using document visualization techniques by treating fact annotations as documents.

Fact Network: A fact network can be automatically constructed according to correlations among fact annotations and manually modified by users. For example, facts can be automatically associated according to the dimensions or the data elements they contain. The fact network captures the relationships among discovered facts and allows users to compare, associate, and retrieve related facts. Graph visualization techniques can be applied to help users interactively navigate in the fact network and retrieve facts from it using graph interactions such as extension search and association search.

Guided Fact Discovery: *Fact notification services* can be provided to automatically keep track of facts registered by users so that the users do not need to keep them in mind. The users will be notified if a new fact is discovered that is related to a registered fact. In addition, when a user meets a potential fact, such as a brushed data cluster, that is related to a registered fact, the system will automatically notify the user about the situation. *Fact recommendation services* automatically or semi-automatically recommend views containing potential facts of interest to users according to registered facts or user requirements during the visual exploration process. The guided fact discovery process can be tightly coupled with automatic data analysis techniques. The fact taxonomy provides a standard language among the users, the system, and the automatic analysis techniques for describing facts that users look for.

Fact Exchange: Standard fact exchange requests can be generated based upon the fact taxonomy to allow efficient fact exchange in collaborative visual analytics. For example, when a user wants to get information from other users, she first requests the system to automatically generate a form listing attributes of a fact in a desired fact category. 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 use guided fact discovery techniques to complete the form and send it back to her.

Since almost all the above activities are associated with the fact taxonomy, we believe that our fact taxonomy construction work will provide a solid basis for the development of the whole FMF.

Table 2: Insight Annotation Form

Attribute	**Value**
1) Creation date	3/20/2008
2) Annotation author	Mary
3) Name of data set	USA emissions per capita by state
4) Visualization	Bubble chart
5) Number of data items	1
Identifier of data item1	Wyoming
6) Number of dimensions	1
Name of dimension 1	emissions per capita
Value on dimension 1	126
7) Rank	1
8) Total number of data items	50
9) Free annotation	Wyoming might make big contribution to weather warming due to the high emission value!

6 A SCENARIO OF FACT MANAGEMENT

In this section, we provide a scenario of how fact management works in visual analytic activities. 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 system. The data sets used in this scenario are real data sets uploaded to Many-Eyes for an ongoing discussion of a similar topic.

First, Mary uploads the data set "USA emissions per capita by state" to the workspace and creates a bubble chart views to visualize it. From this view, Mary discovers the fact that the Wyoming has the highest emissions per person among all the states. According to this fact, Mary suspects that Wyoming might contribute more to global warming than the other states and she decides to record the fact. She selects the *rank* category for the fact and the system automatically creates an annotation form (see Table 2) with all lines except line 9 filled. Mary manually writes her hypothesis in line 9 and stores the fact into a fact database. Mary wants to keep track of this fact, so she registers it so that the system will automatically notify her if a new fact related to it is discovered.

A few days later, Tom wants to know which states make big contributions to weather warming. He submits a search to the fact database for facts in the *ranking* category and with the keyword "emissions" in dimension names. The system returns him some facts, including the fact Mary discovered.

Tom reviews Mary's fact. 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 data set "USA overall emissions by state" and creates a bar chart on it. From the bar chart he discovers the fact that Texas, Florida, Ohio and New York have much higher overall emissions than Wyoming. He thus records this fact as a *difference* fact. Since "Wyoming" is involved in this fact, Mary automatically gets a notification about it from the system. After

Mary reviews this new fact, she discusses it with Tom and makes the conclusion that her previous insight is wrong.

7 Conclusion

In this paper, we build a fact taxonomy that categorizes various facts about multidimensional data and captures their essential attributes. Based on this taxonomy, we propose a conceptual framework of fact management that includes a rich set of fact management activities such as fact annotation, indexing, retrieval, fact network, guided fact discovery, and fact exchange with significant automaticity. This FMF provides a solid basis for effective and efficient insight management in a wide variety of visual analytics applications.

In the future, we will construct fact taxonomies for other data types such as trees and graphs and extend the FMF to those data types. We will also implement and evaluate prototypes of the proposed FMF in a variety of visual analytics environment such as collaboration and heterogeneous data visual sense making, and customize the prototypes for a variety of applications such as bioinformatics and financial intelligence.

Acknowledgements

The authors wish to thank Dr. Martin Wattenberg who gave many valuable suggestions for this work, and developers of Many-Eyes who developed the platform upon which our studies can be conducted. We also would like to thank all the user study participants and Many-Eyes users.

This work was performed with partial support from the National Visualization and Analytics Center (NVAC(TM)), a U.S. Department of Homeland Security Program, under the auspices of the Southeastern Regional Visualization and Analytics Center. NVAC is operated by the Pacific Northwest National Laboratory (PNNL), a U.S. Department of Energy Office of Science laboratory .

References

[1] Many-eyes. http://www.many-eyes.com.

[2] Youtube. http://www.youtube.com.

[3] R. Amar. and J. Stasko. Low-level components of analytic activity in information visualization. *Proc. IEEE Symposium on Information Visualization*, pages 111–147, 2004.

[4] E. Bier, S. Card, and J. Bodnar. Entity-based collaboration tools for intelligence analysis. *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pages 99–106, 2008.

[5] M. Breunig, H. Kriegel, R. Ng, and J. Sander. Lof: Identifying density-based local outliers. *Proc. ACM SIGMOD*, pages 93–104, 2000.

[6] H. Chen, J. Schroeder, R. Hauck, L. Ridgeway, H. Atabakhsh, H. Gupta, C. Boarman, K. Rasmussen, and A. Clements. Coplink connect: information and knowledge management for law enforcement. *Decision Support Systems*, 34(3):271–285, 2003.

[7] D. Cluxton, S. Eick, and S. Research. Decide - a hypothesis visualization tool. *Proc. International Conference on Intelligence Analysis*, 2005.

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

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

[10] J. Gersh, B. Lewis, J. Montemayor, C. Piatko, and R. Turner. Supporting insight-based information exploration in intelligence analysis. *Communications of the ACM*, pages 63–68, 2006.

[11] D. Gotz and M. Zhou. Characterizing users' visual analytic activity for insight provenance. *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pages 123–130, 2008.

[12] D. Gotz, M. Zhou, and V. Aggarwal. Interactive visual synthesis of analytic knowledge. *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pages 51–58, 2006.

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

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

[15] 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*, pages 1029–1038, 2007.

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

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

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

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

[20] C. North. Visualization viewpoints: Toward measuring visualization insight. *IEEE Computer Graphics Applications*, 26(3):6–9, 2006.

[21] M. Rester, M. Pohl, S. Wiltner, K. Hinum, S. Miksch, C. Popow, and S. Ohmann. Evaluating an infovis technique using insight reports. *Proc. 11th International Conference on Information Visualization*, pages 693–700, 2007.

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

[23] P. Saraiya, C. North, V. Lam, and K. Duca. An insight-based longitudinal study of visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1511–1522, 2006.

[24] D. Schneider, C. Matuszek, P. Shah, R. Kahlert, D. Baxter, J. Cabral, M. Witbrock, and D. Lenat. Gathering and managing facts for intelligence analysis. *Proc. International Conference on Intelligence Analysis*, 2005.

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

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

[27] F. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.

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

[29] 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*, 12(6):1363–1372, 2006.

[30] 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.

[31] L. Xiao, J. Gerth, and P. Hanrahan. Enhancing visual analysis of network traffic using a knowledge representation. *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pages 107–114, 2006.

[32] D. Yang, Z. Xie, E. Rundensteiner, and M. Ward. Managing discoveries in the visual analytics process. *ACM SIGKDD Explorations Newsletter*, 9(2):22–29, 2007.

[33] 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.

[34] 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*, pages 392–399, 1998.

Table 3: Fact Taxonomy.

Category	Formal definition	Examples	Content	Context	Related Task
Value/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 [3], Compute derived value [3].
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 [29]; shape description such as convex, skinny, and stringy [29]).	N/A	Characterize distribution [3, 28]
Difference	*Difference* is defined on a 4-tuple (X_i, D_j, f, ∂). $\forall\ x_m, x_n \in X_i$, $f_{Dj}(x_m, x_n) \geq \partial$ 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 [34, 28]
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 [3]
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 [28, 34], Sort [3]
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 [34]
Cluster	*Cluster* is defined on a 4-tuple (X_i, D_j, f, ∂). $\forall\ x_m, x_n \in X_i$, $f_{D_j}(x_m, x_n) \leq \partial$, 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 [34, 28, 3]
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 [3]

Category	Formal definition	Examples	Content	Context	Related Task
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 [14]; quality measures such as correlation coefficient [19] for the continuous scale data, or statistics chi square test [13] for categorical data.	Associate [28, 34], Correlate [3]
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 [14]).	Globe trend; trend of other attributes in the same segment.	N/A
Meta fact	*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 fact	*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 [3]

- X Set of all elements. For a fact in the item space, it refers to the set of all items. For a fact in the dimension space, it refers to the set of all dimensions.
- D Set of all attributes. For a fact in the item space, it refers to the set of all dimensions. For a fact in the dimension space, it is the set of all items.
- V Values of elements on their attributes.
- X_i Subset of X
- D_j Subset of D
- x_i A element in X
- d_j A element in D
- f Distance calculation function
- ∂ User defined constant