Designing for Social Data Analysis

Martin Wattenberg and Jesse Kriss

Abstract—The NameVoyager, a Web-based visualization of historical trends in baby naming, has proven remarkably popular. We describe design decisions behind the application and lessons learned in creating an application that makes do-it-yourself data mining popular. The prime lesson, it is hypothesized, is that an information visualization tool may be fruitfully viewed not as a tool but as part of an online social environment. In other words, to design a successful exploratory data analysis tool, one good strategy is to create a system that enables "social" data analysis. We end by discussing the design of an extension of the NameVoyager to a more complex data set, in which the principles of social data analysis played a guiding role.

Index Terms—Design study, time-varying data visualization, human-computer interaction, social data analysis.

1 Introduction

In February of 2005, Laura Wattenberg, the wife of the first author, published a guide to American baby names called *The Baby Name Wizard* [16]. To help call attention to the book, a Web-based visualization applet, the NameVoyager [10], was launched. The NameVoyager lets users interactively explore name data—specifically, historical name popularity figures. The gambit succeeded and without any advertising the applet drew more than 500,000 site visits in the first two weeks after launch. Eight months afterward, it is maintaining an average of 6,000 visits a day. Perhaps more important is that evidence suggests many people are engaging deeply with the visualization, spending considerable time and discovering for themselves facts and insights about name trends. This paper, an extension of [17], analyzes possible reasons for and lessons learned from the applet's popularity.

The broad popularity and effectiveness of the NameVoyager is especially interesting because it is, in essence, an exploratory data analysis application for a data set of 6,000 time series. In many situations, ranging from education to retirement planning, it is important to encourage users to interact with complex data sets. Understanding the factors that led a statistical exploration program to become a minor fad may shed light on the broader problem of encouraging users to engage in their own personal data mining expeditions.

An important piece of the puzzle is the public nature of a Web-based application. As of October 2005, Google reports more than 43,000 references to the NameVoyager, many of which turn out to be lengthy sequences of comments on blogs and discussion sites. These comments provide clues as to how and why users are spending time with the applet. This data is in no way a scientific survey, but it does represent a large body of field usage information in which patterns emerge.

In hundreds of spontaneous comments, users are seen to be engaged in extended exploratory data analysis, identifying trends and anomalies and forming conjectures. These self-reports also lead to an observation about the NameVoyager: Usage patterns are strongly social and seem more closely related to those of online multiplayer games than to a conventional single-user statistical tool. Indeed, users seem to fall neatly into Richard Bartle's well-known categorization of online game players [4] as explorers, achievers, socializers, or killers. This stands in contrast to the traditional view of information visualization as a task-oriented problem-solving activity. We hypothesize that the broad popularity of the NameVoyager stems from features that not only give it a game-like sense of fun, but that make it especially suitable for "social" data analysis. We then suggest some general design principles which may encourage this type of usage of visualizations.

As an example of these design principles in action, as well as the general utility of the visualization technique, we describe the BookVoyager, an extension of the NameVoyager that was created for a well-known technical book publisher. The BookVoyager allows a user to explore a large set of time series with a hierarchical organization and incorporates several features to encourage social data analysis.

2 The NameVoyager

2.1 Data

The NameVoyager is based on a data set derived from public Social Security Administration (SSA) information that tracks baby name trends in the US. For each decade since 1900, the SSA publishes lists of the most popular 1,000 boys and girls names, along with the exact number of babies given these names. These lists were downloaded, collated, cleaned, and normalized by the author of the *Baby Name Wizard* book to produce a data set containing popularity time series for roughly 6,000 distinct names.

These time series turn out to be meaningful in many ways. A graph of the popularity of a given name reveals a great deal about its overall cultural connotations; names whose popularity is correlated over time tend to seem subjectively similar to Americans. (For more information, see *The Baby Name Wizard*.)

- *The authors are with IBM Research, Visual Communication Lab, 1 Rogers Street, Cambridge, MA 02142. E-mail: {mwatten, jesse.kriss}@us.ibm.com.*

Manuscript received 22 Oct. 2005; revised 30 Dec. 2005; accepted 19 Jan. 2006; published online 10 May 2006.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-0138-1005.

1077-2626/06/$20.00 © 2006 IEEE Published by the IEEE Computer Society

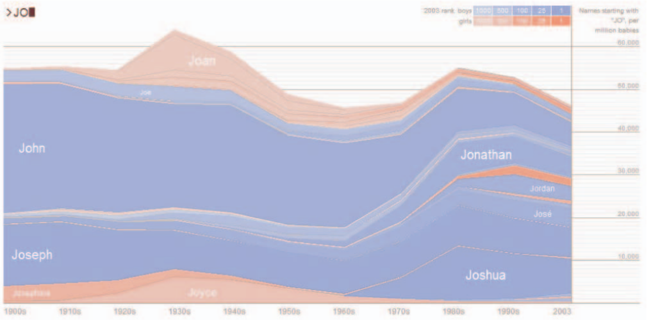

Fig. 1. The NameVoyager.

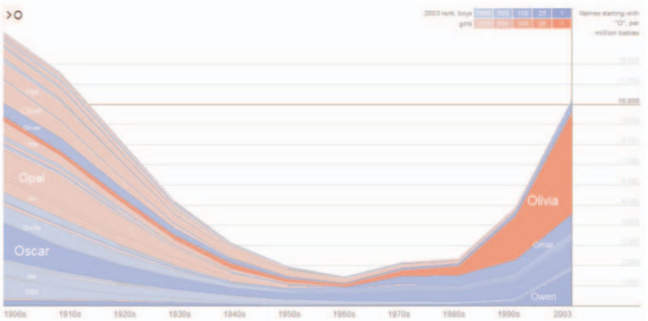

Fig. 2. Names beginning with O.

2.2 Visualization Method

The method used to visualize the data is straightforward: Given a set of name popularity time series, a set of stacked graphs is produced, as shown in Fig. 1. Such stacked graphs are common in print information design and have recently been used in several information visualization projects such as *ThemeRiver* [7] and *Artifacts of the Presence Era* [15]. The x-axis corresponds to date, and the y-axis to total frequency for all names currently in view, in terms of occurrences per million babies. Each stripe represents a name, and the thickness of a stripe is proportional to its frequency of use at the given time step.

In keeping with contemporary American custom, the stripes are colored pink for girls and blue for boys. The brightness of each stripe varies according to the most recent popularity data, so that currently popular names are darkest and stand out the most. The idea behind this color scheme is twofold. First, names that are currently popular are more likely to be of interest to viewers—many people will probably want to know statistics on Jennifer, but few are looking for Cloyd. Second, the fact that the brightness varies provides a way to distinguish neighboring name stripes without relying on visually heavy borders.

2.3 Interaction

The NameVoyager follows Shneiderman's mantra of "overview first, zoom and filter, details on demand" [12]. When the applet starts, the viewer sees a set of stripes representing all names in the database. Filtering this data is achieved via an extremely simple mechanism. A user may type in letters, forming a prefix; the applet will then visualize data on only those names beginning with that prefix.

The applet reacts directly with each keystroke, so it is not necessary for the user to press return or to click a submit button. Not only does this instant interaction save the user some work, but it helps demonstrate how to mine the data. A user might not think that searching the data set by prefix would be interesting, but seeing the striking patterns for single letters like O or K could encourage further exploration. In addition, the applet moves smoothly between states, so that when a letter is typed, an animated transition helps preserve context.

Fig. 1 shows an example: typing "JO" will yield a graph with prominent stripes for popular names such as John, Jonathan, Joseph, and Joyce, along with many thinner stripes for less popular names like Josette. Because the initial letters of a name contribute strongly to its sound, names that start with the same letters often have similar graph patterns. As a result, the simple mechanism of filtering by prefix is effective in highlighting interesting name trends. Typing "O" produces the graph in Fig. 2, with an easily identifiable pattern of popularity of O names at the beginning and end of the 1900s, but a significant dip mid-century. Typing "LAT" highlights a trend in the African-American community in the 1970s, comprising names such as LaToya, LaTanya, LaTisha, and so on, as in Fig. 3. Name stripes are ordered alphabetically on the screen from top to bottom to aid in identifying such prefix-based cultural clusters.

To learn the details of a name, a viewer can use the mouse. Hovering over a name stripe will produce a pop-up box with numerical details for a given name at a given point in time. Clicking on a name stripe produces a graph of the popularity of that name alone.

This interaction technique may be compared to dynamic query systems such as starfield displays [2] or TimeSearcher [8]. The keyboard interaction may be viewed as an alternative to the Alphaslider of [1]. A key distinction between the graphical display of the NameVoyager and the visualization used in TimeSearcher is the NameVoyager's use of a graph that sums all the time series. This technique seems likely to be of use in many other situations where summing is a natural operation, such as investigating product sales data.

2.4 Technical Implementation

The NameVoyager is a Java applet, written using JDK 1.1 so that it may run in a wide variety of browsers. All the name data (a 60K zip file) is loaded at startup and parsed into Java objects, so that it may be accessed rapidly.

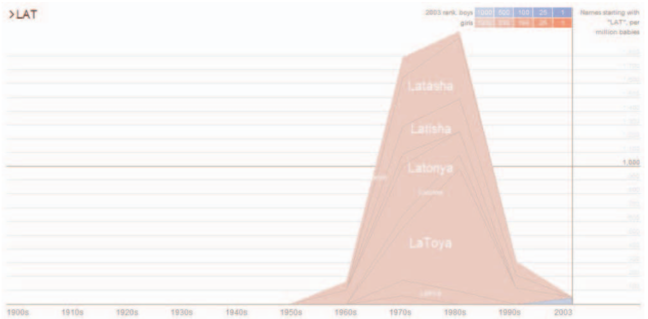

Fig. 3. Names beginning with LAT.

To make the animated transitions run smoothly, not all 6,000 stripes are drawn; instead, a simple level-of-detail calculation is performed so that only stripes wider than 2 pixels are rendered to the screen. As a result, in practice, the applet only draws about 200 or fewer stripes per frame. In an initial version of the applet, this culling of names caused prominent and irritating white stripes in the graph, where the white background would "show through" the undrawn stripes. Replacing the white background with a neutral gray, halfway between the blue and pink tones of the name stripes, was a simple and effective remedy: The background was still visible but barely noticeable.

3 Results

3.1 Traffic and Web Comments

As mentioned in the introduction, the NameVoyager received a remarkable number of visits within weeks of launch. The applet has been downloaded more than 900,000 times as of mid-April. It has also been extensively discussed on the Web, in blogs, discussion forums, and similar sites.

This intense level of conversation is further evidence that users were engaging deeply with the applet and of its widespread popularity. Although Internet fads are common, the detailed and often technical debate that surrounded the NameVoyager provides evidence that its popularity went beyond that of a "cool toy." It is not uncommon to find discussions in the comments section of a blog that contain dozens of posts. Such long discussions occur even when it is not related to the topic of the Web site; for instance, one of the most extensive sets of comments was found on a forum in a well-known libertarian magazine site.

The comments also provide an unusual and informative window into the user experience, and we quote them extensively below. Comments that have been posted to the Web are clearly not a scientific sample, since only the most enthusiastic users will comment. Nonetheless, examining these comments suggests some interesting hypotheses regarding the source of popularity of the NameVoyager.

3.2 The Target Audience and the Surprise Factor

As one might expect, there are many positive comments from people in the target audience for the visualization-users who have a strong interest in names and therefore might be interested in buying the book. Two examples (all quotes in this paper are taken from public Web sites) illustrate this:

"*This is perfect, as baby names weigh heavily on my mind these days.*"

"*Useful fodder for historical fiction, too, if you're looking for typical names for a given age and time period.*"

A surprising observation is that many people outside the target audience found themselves enjoying the applet. The surprise here is not the authors', but the users' themselves. Some sample quotes:

"*Surprisingly addictive.*"

"*This rules, even though it's about baby names.*"

"*Cool...by the way, I don't like babies or children.*"

This "surprise factor" is a reason for optimism. It is common to want users to explore a set of data that they may have little inherent interest in. A good example is the amount of effort and money that American companies spend to encourage their employees to understand 401(k) plans. It is therefore worthwhile to look for clues to what made the NameVoyager appeal to people who profess boredom with the topic of baby names.

4 Social Data Analysis

One of the most consistent themes seen in comments about the NameVoyager is that exploring the data has become a social activity. Many people mention group usage, for instance:

"*I happened upon it at work today and it affected the productivity of our entire department.*"

Of special interest, however, is that when a group of people uses the applet, they often do so in a social, collaborative fashion, engaging in a dialogue as they mine the data. This is true even for loosely knit groups of Web users. For example, here are some quotes from the comments section of one blog:

"*For a challenge, try finding a name that was popular at the beginning of the sample (around 1900), went out of style, then came back into vogue recently.*"

Another person responds, "*Take a look at Grace, #18 in the 1900s, #13 in 2003, and down in the 200s and 300s during mid-century.*"

A third writes,

"*1900's comeback: Porter. Another one, with a mini-peak in trough: Caroline,*" and then adds,

"*More challenges: which is the steadiest popular name? Victor?*" and "*Which letter has gone down most consistently? W? Observation: Note the recent upsurge in Y; basically all due to Hispanic (and some Middle Eastern) names.*"

The original poster responds, "*You're right, W has gone most consistently down, although F is pretty close (if it weren't for Faith...)*"

These quotes, which are just a small part of the full exchange, illustrate two points. First, they show how a group of people is using the NameVoyager as a stimulus to conversation and repartee.

They also reveal an effective style of data analysis: This group of people is diving very deeply into the data set! They are setting each other pattern-finding challenges, noting outlying data points, and making guesses about causal relations. Each person seems to be building on the findings of the others, making the group as a whole extremely effective at mining the data—and having fun at the same time. Strange or surprising pieces of information serve as a kind of trophy for the finder. We refer to this process of data mining through dialogue, one-upmanship, and repartee as social data analysis. It is a version of exploratory data analysis that relies on social interaction as source of inspiration and motivation.

We hypothesize that viewing exploratory data analysis as a social activity may explain much of the reaction to the

NameVoyager. Its popularity among people who do not find the data intrinsically interesting, for instance, could partly be due to the fact that these users are enjoying the social activity surrounding the applet. In the next sections, to better understand the social structure of this type of exploratory data analysis, we consider the different roles that users may play.

4.1 Roles in Social Data Analysis

As in any social system, it seems that people using the NameVoyager have a wide range of styles of interaction with each other. Comments on the Web suggest that there are four distinct types of users. Interestingly, these types seem to align closely with a taxonomy developed by Richard Bartle [4] in the context of an early class of online social environment called a MUD.

Bartle suggested that denizens of such online multiplayer environments typically fall into one of four types: achievers, socializers, explorers, and killers. Below, we describe how each of these roles seems to correspond to a particular type of NameVoyager user. While this is only a preliminary classification, it may be of use to designers in thinking about how people use data visualization in social contexts, and also provides additional evidence that use of the NameVoyager takes place in a complex social environment.

4.2 Achievers

The context of the NameVoyager is a site designed to help expectant parents name their babies, so the stated "goal" of the applet is to find a good name. As described in Section 3.2, many people do exactly that:

> *"We want something slightly retro, nice, and not too popular, and this visualization gives us all that."*

Such users correspond to the Achievers in Bartle's classification: people who try to "achieve within the game's context."

4.3 Socializers

A second class of NameVoyager users consists of people whose main concern is their interactions with others and who place their data exploration in a personal social context. These people, corresponding to Bartle's "Socializers," use screenshots and data from the applet as a catalyst for conversation and storytelling about themselves and their friends and family. A common sight on a blog is a person posting a screenshot of the graph of their own name's popularity, or a friend's, with humorous comments. A typical quote of this type is:

> *"Runes name doesnt show up at all...but my name has suddenly gotten popular...I HAD IT FIRST! heh."*

Often, people talk about family members as they speculate about names and see the changing popularity numbers as a kind of personal plotline:

> *"My grandmother was named Coral and from what I can tell the name appeared out of nowhere in 1880...is it from a celebrity or something?"*

> *"I got: 'No names starting with LINUS were in the top 1,000 names in any decade.' Translation: Your son's name will NEVER be cool."*

> *"Woo! Emily (being me) was number 1 in 2003! go me!"*

Such relationship-oriented and storytelling behavior in the context of information visualization has been observed before in depictions of e-mail archives [14].

4.4 Explorers

Many users of the NameVoyager seemed to delight in unearthing odd names or unusual clusters. One person posted a screenshot created after typing "ETH": It showed the name Ethel being gradually and completely eclipsed by the trendy name Ethan. Another found the dramatic cluster of names starting with "LAT" (Latisha, Latoya, etc.) described in Section 2.3. A well-known pundit used the NameVoyager to comment on the changing statistical distribution of names over the past century.

These users were certainly not using the NameVoyager to name children, but rather were mining for nuggets of information that they could show to others as trophies of their expedition. They are directly analogous to Bartle's Explorers, people who want to learn as much as possible about the environment and who delight in discovering odd or unexpected features.

4.5 Killers

The last type in Bartle's taxonomy is the Killer, someone who enjoys imposing themselves on others and causing distress. One might think that there would be no Killers in the gentle world of baby names, but one would be wrong. A common theme is that certain users take pleasure in singling out names for ridicule. For these people, the NameVoyager is a delightful source of fresh targets:

> *"It is also damn entertaining to me (and the real reason why I am writing this) that I can type in Lexus and find that people actually name their kids Lexus."*

(Lest there be any doubt about the pugilistic nature of the quote's author, note that it was found on a site called www.youandwhosearmy.com.)

> *"Britney, Brittney, Britany, Brittany, Brittani, Britannie, Britni. Enough already."*

5 Design Hypotheses for Social Data Analysis

The evidence above suggests that a large part of the power and popularity of the NameVoyager derives from the fact that it encourages a social style of data analysis. What leads users to approach data analysis as a social activity? Certain factors are obvious. The NameVoyager is easily accessible on the Web so that a large group of people can see it. The interaction design, referred to on the Web with such terms as "cool," "fantastic," and "whizzy," means that applet is something that people may be eager to associate themselves with, like a fashionable piece of clothing.

These factors, however, would apply to anything trendy on the Web, whether it is a funny Flash animation or witty personality quiz. Are there any aspects of the NameVoyager's popularity that are specific to information visualization? We present three hypotheses below.

5.1 Common Ground But Unique Perspective

The first hypothesis is that a combination of common ground with unique individual perspectives will encourage social data analysis.

In the case of the NameVoyager, the common ground is a shared understanding of the cultural connotations of names. Although people may differ in their tastes, most Americans would agree on the likely ethnicity of a Rodrigo or a LaTanya, or the likely age of an Ethel versus a Heather. Similarly, many names relate to celebrities, pop culture icons, or historical figures.

This common ground is what makes conversation about the data possible and interesting. Some sample quotes:

> "*Look what the Simpsons did to the name Bart.*"

> "*Roosevelt has two spikes right about where you'd expect them.*"

> "*I love the fact that Xander and Willow show up on the list in the 90s, thereby confirming the existence of Buffy fans as hardcore as me.*"

The authors of these comments are sharing results of their data mining because they know that their readers will understand the cultural references. The fact that the data is presented as a timeline over a standard period, 1900 to present, also provides a common context on which users overlay personal and cultural knowledge.

At the same time, we hypothesize that it is helpful for each person to have a naturally unique perspective on the data. This individual viewpoint can serve as a kind of icebreaker in the conversation. It also means that, because each person is approaching the data in a different way, a group may collectively explore more pieces of the data. Evidence for this hypothesis comes from [9], which described a system that encouraged community participation by highlighting unique pieces of knowledge that an individual might have. A well-respected educational method known as the Jigsaw Classroom [3] uses a similar technique.

In the case of the NameVoyager, each person has one obvious point of entry: their own name. Names of relatives and close friends are also common conversation starters. Some sample comments illustrate this:

> "*I was appalled to note that my name is now in the top 100, while it was about 700th when I was born...*"

> "*My given name peaked in 1900 (or earlier) and has been on the slide ever since. Seems to be off the radar now. Elmer is more popular these days!*"

> "*It also confirmed my suspicion that our eight-month-old son's name, Jackson, was rapidly gaining in popularity. Dangit, and we thought he would avoid having 4 kids in kindergarten with the same name!!!*"

Thus, usage of the NameVoyager follows a pattern in which people look at different aspects of the data set, but have an expectation that their particular findings will be interesting and understandable to others. We term this pattern the "common ground but unique perspective" principle.

Applying this principle in other situations may require some flexibility in the data set, but it may also be possible to guide people without modifying the data. For instance, imagine a visualization tool designed to help people understand different stock market investment strategies. Using well-known companies or events as landmarks could provide common ground. At the same time, there are several unique perspectives that people might take: for instance, looking at how their own company's stock has performed or how the market as a whole did at significant points in their life. It is possible that the visualization could be tailored to bring out these perspectives.

5.2 Expressive Spectator Interface

In many cases, a group of two or more people used the NameVoyager together. This is to be expected in the case of two parents-to-be trying to find a name they both like, but also seemed to occur in other contexts; as one person wrote,

> "*We spent hours typing in the names of everyone we know.*"

When a group uses a single-input software tool like the NameVoyager, there are two distinct user roles. At any given moment, one person will be active, controlling the input, while others in the group will act as spectators. (These spectators may of course be active in other ways, talking with each other and making suggestions to the user controlling the input.) Traditionally, interface designers have focused on the active participant, but recently it has been suggested that designing for the spectator role creates important special considerations [11]. A natural hypothesis is that a social data analysis tool should support spectators as well as active participants.

Does the NameVoyager interface have special properties that create a good spectator experience? Two notable features of the NameVoyager are the smooth animation between states and the unusually prominent text entry area. The animation was initially added for the simple reason that it looked good, while the text area indicates to novice users that they should start typing. These two features, however, also give the NameVoyager an effective spectator interface.

The prominent text area makes it easy for someone peering over the shoulder of a user to see what is being typed. The immediate letter-by-letter changes in the graphs give the display a live-action quality, allowing spectators to see each step of the user's thinking process. The animation emphasizes the results of the typing and links successive states in a coherent progression. This avoids the jarring feeling—familiar to anyone watching television while someone else wields the remote control—of seeing a series of sudden, unexpected changes.

It is interesting to compare the NameVoyager to the PhotoMesa application described in [5]. PhotoMesa is a tool for sorting, annotating, and displaying large collections of photographs. As described in [5], an important user scenario for PhotoMesa involves "family viewing," in particular, the case of an adult operating the program while a child watches. Two features of PhotoMesa support this scenario. First, as in the NameVoyager, transitions between different states are effected by smooth transitions, involving zooms and pans. Second, the standard mouse cursor movement is augmented by a prominent rectangle showing what would happen if the user clicks; while this feature clearly is helpful to the person wielding the mouse, it also can be seen as analogous to the large flashing text entry area of the NameVoyager, making a standard input device more prominent for the benefit of an audience.

Because both input and output are amplified, the interfaces of the NameVoyager and PhotoMesa fall in the "expressive" quadrant of the spectator interface taxonomy discussed by Reeves et al. in [11]. (The other quadrants are termed "suspenseful," "magical," and "secretive.") We suggest that, for information visualization, where clarity and common understanding are critical, the "expressive" style of spectator interface is best—and that there may be features, such as animated transitions, that have larger value for groups than single users.

5.3 Discovery Transfer

The final hypothesis about how the NameVoyager encourages social data exploration is that it allows people to share the state of the visualization at any point in their explorations. Because the interaction model is so simple—just a matter of typing a few letters—it is very easy to guide other people to the same state. And, indeed, many comments on the Web are written in the imperative voice:

"*Take a look at K and see how it exploded in the last decade or two.*"

"*Type in Adolph for example.*"

"*You want some real fun, run 'Hillary'.*"

What people are doing here, by hand, is creating a kind of pointer into the application—that is, making a reference into a particular state following interaction. The ability for users to transfer their discoveries to others may be critical to the conversation surrounding the NameVoyager. Solitary, asynchronous usage can in this way become a shared experience. The ease of "showing off" discoveries also fosters a motivating sense of pride and competitiveness.

Thus, a natural design principle might be that information visualization software ought to provide "application-state pointers" if it is intended to support collaborative analysis. Such pointers could involve special URLs for later reference or some other technology. A good example of an application-state pointer in a commercial visualization tool comes from the Web interface to the Spotfire system [13], which allows users to make comments about an online analysis. When reading a comment, another user can view the exact state of the visualization (slider position, data, etc.) seen by the comment's author.

Note that allowing application-state pointers may impose some subtle constraints. Some graph layout algorithms, for example, involve random numbers or depend on a long history of user manipulations. These algorithms would need to be modified to allow different people to see consistent views.

6 The BookVoyager

Soon after the introduction of the NameVoyager, we were approached by a publishing company that wished to use the same basic technique to visualize and explore data on historical trends in sales for various types of books. With this impetus, we created a more general tool based on the NameVoyager, which is capable of analyzing a common class of time series data. Although this application is still a work in progress and we cannot report results of user studies, we believe our design decisions provide a useful example of how the design principles for social data analysis may be applied in practice.

6.1 Hierarchical Additive Time Series

The book data set shared two characteristics of the baby name statistics. First, it consisted of a large number (several hundred) of time series. Second, the series were naturally "additive"—it makes sense to sum up the sales of subsets of books. (Note that it doesn't always make sense to add up time series; e.g., there would be no point in adding up a set of daily temperatures from different locations.) For these reasons the NameVoyager was a plausible base for a visualization of book publishing trends.

Two features of the data set were different from the name data, however. The sales were organized into a detailed hierarchy of categories and subcategories. A typical series was "digital design topics," categorized within "digital media," which itself was a subcategory of "digital media applications and devices." A second, related difference was that—unlike in the NameVoyager—typing alone was not always the optimal method of zooming and filtering. In the example above, consider the number of characters needed to distinguish "digital media applications and devices" from "digital media."

These two features—a hierarchical organization and the insufficiency of typing as a navigation mechanism—can be observed in a variety of other data sets as well. (Indeed, the high salience of alphabetical order is one of the most unusual aspects of the baby-name data.) It therefore seemed worthwhile to create a general tool for such data, which we called the BookVoyager to reflect its origins in the book sales data and the NameVoyager.

6.2 The Basic BookVoyager Interface

A screenshot of the basic BookVoyager layout can be seen in Fig. 4.

As in the NameVoyager, the bulk of the screen space is devoted to a stacked-graph representation of a set of time series. At the left is a standard tree control that reflects the hierarchical structure of the data. (This standard control, rather than a treemap, was used so that labels would be easily visible.) The most basic user interaction is to use the tree control to winnow the set of time series displayed: As the user clicks on an element of the hierarchy, only items at that level or below are displayed on screen. As in the NameVoyager, as the user selects different nodes, the transitions in the visualization are smooth and animated. In addition to the tree control, the user may type search terms into a prominent text area at the top of the screen. Unlike the NameVoyager, these terms may include regular expressions.

Fig. 5 illustrates an additional feature that was requested from several users: an option to normalize the displayed data so that the sum of all visible time series was constant. This option, enabled by clicking the "normalize" checkbox, allows users to see relative trends rather than absolute ones. It is particularly important in situations where the absolute

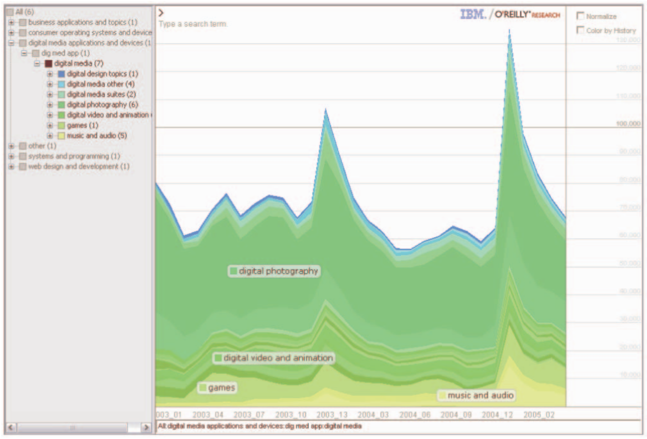

Fig. 4. The BookVoyager.

trends are large and, for some purposes, irrelevant: for instance, in examining book categories whose sales are known to be highly seasonal. (This is the situation in Fig. 5, which shows the same data as in Fig. 4.) When this option is selected or deselected, the resulting change is smoothly animated, again for the purpose of creating an expressive interface.

6.3 Color-Coding to Reflect the Hierarchy

Color-coding is used to make the structure of the hierarchy visible. In the image above, for example, the user has selected the "digital media" category. Each of its seven subcategories has been assigned a distinct hue ranging from blue to yellow. Note that these subcategories themselves have subcategories, so that each contains several time series. An individual time series is assigned a brightness that reflects the overall upward or downward trend, just as in the NameVoyager.

We arrived at this method of color coding after considerable experimentation. Early designs used various monochrome schemes. In one, thick lines were used to delineate the different subcategories. In another, alternating bands of light and dark gray were employed. Both methods suffered from a labeling problem: It was hard for a user to associate the region on the graph that represented a subcategory with the corresponding element in the tree control. To do so required careful reading of the labels associated with areas of the graph, which was often difficult with thin stripes. The second method, alternating light and dark gray, suffered from an additional problem: If a particular subcategory had a stripe on the graph that was extremely thin—so thin as to take less than 1 pixel in height—there would be no visual separation of the surrounding subcategories.

To ensure sufficient visual separation of subcategories, it was necessary at each level to use the full range of hues. Thus, when a user changes levels in the hierarchies, the color coding must change as well. This turns out to be a potentially confusing feature, since it is extremely rare in interactive information graphics for color coding to change as the result of navigation. To help explain the transition to the user, the colors change via smooth interpolation, so that as users smoothly zoom the size scale on a subset of a time series, they see a similar smooth zoom in color space. This transition also helps onlookers understand the change, thus obeying the principle that the application should maintain an expressive audience interface.

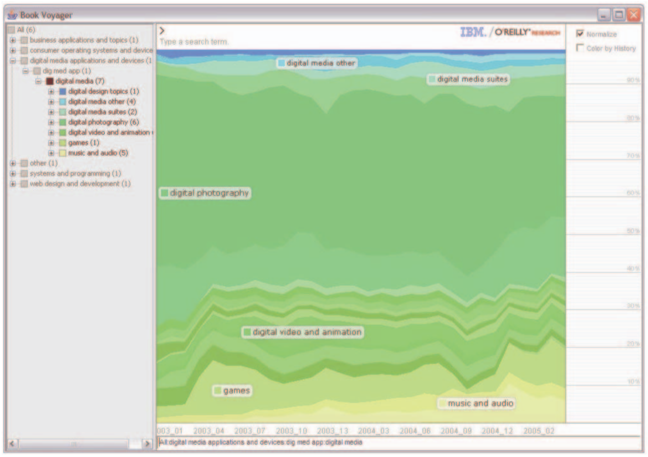

Fig. 5. BookVoyager in "normalized" mode.

6.4 "BookMarks" for Discovery Transfer

A second design principle suggested above was "discovery transfer," or the ability for people to share the application state easily. In the NameVoyager, discovery transfer is effected by typing small strings. The equivalent in the BookVoyager is a text field containing a sequence of characters, called a BookMark, that always reflects the current zoom or search state. If one user is interested in a particular view of the data, he or she can easily copy this sequence of characters onto the clipboard and paste into e-mail or an instant messaging program to send to another user. That user can paste the BookMark into the text field on their copy of the program, which immediately sets the program to the state seen by the first user.

While this gesture may appear somewhat clumsy, it turns out to feel surprisingly natural, possibly because it mimics the common practice of sharing URLs by copying the location in the address bar of a browser. The term BookMark was chosen to reinforce this analogy to familiar addressing systems. In addition, because of the simplicity of the method, it turned out to be relatively easy to implement.

6.5 Empasizing the Road Less Traveled

The third design principle that we have hypothesized to enable social data analysis is that users should bring a unique perspective to the data. In the case of book sales, for certain user groups, this may happen naturally. For example, if several editors are looking at the application, each might naturally look at their own specialties. In other cases, however, it might be useful to assist this process technologically.

To provide such assistance, the application keeps track of which time series have been "visited" or viewed closely. To be precise, we define a visit to mean that either the user fully zoomed in on a time series or it zoomed in on the smallest category containing the time series. (This criterion corresponds to the level of detail that forces a label to appear on the screen.) The application has a checkbox to

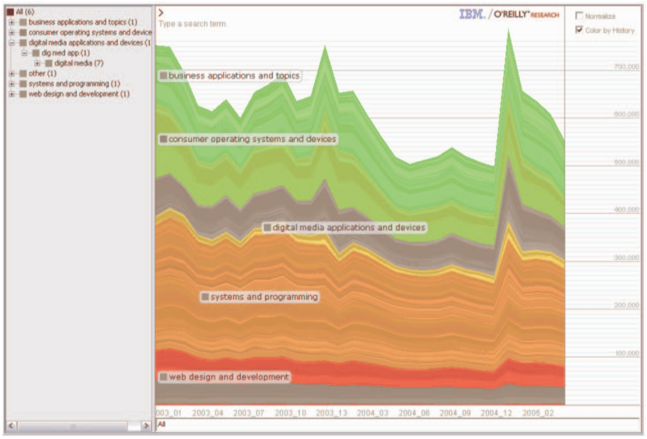
All (6)
business applications and topics (1)
consumer operating systems and device
digital media applications and devices (1
dig med app (1)
digital media (7)
other (1)
systems and programming (1)
web design and development (1)
>
Type a search term.
IBM. / O'REILLY RESEARCH
☐ Normalize
☑ Color by History
business applications and topics
consumer operating systems and devices
digital media applications and devices
systems and programming
web design and development
700,000
600,000
500,000
400,000
300,000
200,000
100,000
003_01
2003_04
2003_07
2003_10
2003_13
2004_03
2004_06
2004_09
2004_12
2005_02
All

Fig. 6. Gray stripes show "visited" data items.

"color by history" that causes any visited series to appear in gray. (See Fig. 6.) The current implementation only tracks this history locally—hence, is most appropriate for a use case in which many people collaborate around a single display over a period of time—but, we are in the process of creating a version that talks to a central server, so that remote users may benefit as well.

The idea behind this feature is to focus attention on unvisited areas. We refer to this as "road-less-traveled navigation": Instead of using previous visits as a cue to importance, as in traditional social navigation interfaces [6], we treat it as a cue to staleness and hope to draw a user's eye to new territory, thus suggesting a unique perspective to each user.

6.6 Evaluation and Future Directions

We have not yet conducted formal studies of the efficacy of the BookVoyager techniques, but initial signs are promising. The first version of the BookVoyager was well-received by the client who initially brought us the book data and was featured by them in several large conference presentations. Perhaps more interestingly, we have been able to apply the BookVoyager to other data. One potential client, for example, approached us with statistics on traffic to a Web site, with historical data divided into a hierarchy based on the site's directory structure. Aside from writing a small adapter to parse their data format, no changes needed to be made to the BookVoyager application. The type of data visualized by the BookVoyager—time-series that have a hierarchical arrangement and that can be meaningfully summed—seems to represent a common pattern. Moreover, the hierarchical arrangement lends itself to summarization and level-of-detail calculations—an additional advantage that future versions of the system may exploit if there are thousands or millions of time series.

A solid evaluation of BookMarks and "visited marks" for data items will require a wider deployment. Based on the early feedback, however, we are investigating ways to embed such "social widgets" into other applications as well.

7 Conclusion

The NameVoyager is a visualization of baby name popularity data, using keyboard-based interaction and smooth animation to allow users to explore a set of 6,000 time series. The applet has proven extremely popular, attracting hundreds of thousands of users in the space of two months. In addition, thousands of comments about the visualization have been written on the Web.

This paper has explored the reaction to the NameVoyager, using these Web comments as evidence. This methodology is somewhat unusual, but the sheer amount of online discussion of the NameVoyager provides a useful source of detailed descriptions from real users and is a fruitful source of hypotheses about how and why the NameVoyager is effective.

Part of the reaction to the NameVoyager comes from people who are naming a baby and who believe the visualization method is effective. In the BookVoyager, we have extended and generalized the visualization and interaction techniques to a broad class of time series data. The BookVoyager interaction is largely based on navigating through a hierarchy and textual search. It would be interesting to include a broader array of query options. While tools exist to allow dynamic queries of time series data [8], none are tuned for the additive, hierarchical time series examined here.

In addition, however, Web comments reveal that the NameVoyager is popular even among people who have no vested interest in looking for names—the applet is somehow appealing to people even when it is not solving an immediate problem. Moreover, users seem to be doing extensive data mining with the application, finding for themselves subtle patterns in the data. These facts make it all the more interesting to understand the NameVoyager's popularity, since it may serve as a model for other situations, especially in education, where the goal is to impart insight into a set of data that may not be immediately relevant to a user.

A central observation made from comments found on the Web is that usage of the NameVoyager often involves a high degree of dialogue between users. It seems, at least in some cases, to be a social activity in which users discuss findings, send each other puzzles, and draw inspiration from one another. We believe this type of activity, which we term social data analysis, is the key to the efficacy and popularity of the applet. The collaborative, distributed nature means that people can join forces and share knowledge; the social aspect, because it is intrinsically enjoyable, may explain the applet's appeal to users who state that they do not like babies or are not interested in baby names.

Understanding the patterns of social data analysis seems like a promising area for future research. This paper uses Bartle's taxonomy of players in multiuser online games as a starting point for understanding the different roles of people interacting with the NameVoyager. A natural area for further investigation would be to test this idea, perhaps through user interviews and questionnaires. Since Bartle's paper first appeared, other analyses of game player types have been made, with slightly varying taxonomies [18]. While we have found that Bartle's taxonomy seems to fit best, it might be fruitful to verify this finding and understand why that particular taxonomy fits.

We have also proposed several design principles for social data analysis, each of which requires validation. It would be interesting, for example, to explore how effective "spectator interfaces" might differ from standard interfaces. Indeed, is there a simple experiment that might show that some feature, such as animated transitions, has no value for a single user but provides a significant benefit for a group?

Similarly, it would be helpful to investigate methods that allow groups to coordinate their investigation. Application-state pointers, we hypothesize, may be one way to do so, but present engineering and algorithmic challenges, as well as more conceptual ones. How should such pointers behave, for instance, in an application where the underlying data is constantly changing? The common ground/unique perspective hypothesis says that it is helpful for users to have unique entry points into a data set. Are there ways to encourage these unique viewpoints?

Given the variety of questions to be asked, we believe exploring further frameworks and design principles related to social data analysis will be a fruitful avenue of investigation.

Acknowledgments

Many thanks are due to the members of the Collaborative User Experience group at IBM for comments on the ideas in this paper, in particular, Steve Rohall and John Patterson. The authors are especially grateful to Fernanda Viégas, whose ideas on the social facets of information visualization strongly influenced this paper. Roger Magoulas deserves credit for suggesting that they extend the NameVoyager into the BookVoyager, and made many helpful design suggestions. Laura Wattenberg provided significant input into the NameVoyager design and helpful comments on this paper.

References

[1] C. Ahlberg and B. Shneiderman, "The AlphaSlider: A Compact and Rapid Selector," *Proc. ACM Conf. Human Factors in Computing Systems,* 1994.

[2] C. Ahlberg and B. Shneiderman, "Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays," *Proc. ACM Conf. Human Factors in Computing Systems,* 1994.

[3] E. Aronson and S. Patnoe, *The Jigsaw Classroom: Building Cooperation in the Classroom,* second ed. New York: Addison Wesley Longman, 1997.

[4] R. Bartle, "Players Who Suit MUDs," *J. MUD Research,* vol. 1, no. 1, http://www.mud.co.uk/richard/hcds.htm, 1996.

[5] B. Bederson, "PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubble Maps," *Proc. UIST 2001, ACM Symp. User Interface Software and Technology,* vol. 3, no. 2, pp. 71-80, 2001.

[6] A. Dieberger, "Supporting Social Navigation on the World-Wide Web," *Int'l J. Human Computer Studies,* vol. 46, 1997.

[7] S. Havre, B. Hetzler, and L. Nowell, "ThemeRiver: Visualizing Theme Changes over Time," *Proc. IEEE Symp. Information Visualization,* 2000.

[8] H. Hochheiser and B. Shneiderman, "Dynamic Query Tools for Time Series Data Sets, Timebox Widgets for Interactive Exploration," *Information Visualization,* vol. 3, no. 1, 2004.

[9] P. Ludford, D. Cosley, D. Frankowski, and L. Terveen, "Think Different: Increasing Online Community Participation Using Uniqueness and Group Dissimilarity," *Proc. ACM Conf. Human Factors in Computing Systems,* 2004.

[10] NameVoyager, http://babynamewizard.com/namevoyager/lnv0105.html, 2006.

[11] S. Reeves, S. Benford, C. O'Malley, and M. Fraser, "Designing the Spectator Experience," *Proc. ACM Conf. Human Factors in Computing Systems,* 2005.

[12] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," *Proc. 1996 IEEE Conf. Visual Languages,* 1996.

[13] Spotfire DecisionSite, Somerville, Mass.: Spotfire, Inc., 2005.

[14] F. Viégas, D. Boyd, D. Nguyen, J. Potter, and J. Donath, "Digital Artifacts for Remembering and Storytelling," *Proc. Hawaii Int'l Conf. System Sciences,* 2002.

[15] F. Viégas, E. Perry, E. Howe, and J. Donath, "Artifacts of the Presence Era: Using Information Visualization to Create an Evocative Souvenir," *Proc. IEEE Symp. Information Visualization,* 2004.

[16] L. Wattenberg, *The Baby Name Wizard.* New York: Broadway, 2005.

[17] M. Wattenberg, "Baby Names, Visualization, and Social Data Analysis," *Proc. IEEE Symp. Information Visualization,* 2005.

[18] N. Yee, "The Psychology of MMORPGs: Emotional Investment, Motivations, Relationship Formation, and Problematic Usage," *Avatars at Work and Play: Collaboration and Interaction in Shared Virtual Environments,* R. Schroeder and A. Axelsson, eds., London: Springer-Verlag, 2005.

Martin Wattenberg received the PhD degree in mathematics from University of California at Berkeley. He invents visualization techniques to help individuals and groups understand complex data. Aside from baby names, he has created software to explore the stock market, music, wikis, search trees, and social networks. His visualization-based artwork has been shown at the Whitney Museum of American Art, the London Institute of Contemporary Arts, and other venues worldwide.

Jesse Kriss received the Master's degree in human-computer interaction from Carnegie Mellon University and the BA degree in music from Carleton College, where he studied composition with Phillip Rhodes and developed a computer-based performance instrument for real-time sampling. Whether working on information visualization, interaction design, or tools for artistic performance, he is interested in shortening the feedback loop between creation and response, so that time can be spent reacting and improving the idea rather than dealing with the tools themselves.