Process and Productivity in Visual Analytics: Reflections on E-Discovery

Sean M. McNee and Ben Arnette

Attenex Corporation

ABSTRACT

Because visual analytics is not used in a vacuum, there are no cut-and-dry metrics which can accurately evaluate visual analytic tools. These tools are used inside of existing business processes, thus metrics to evaluate these tools must measure the productivity of information workers on the data-centric critical path of these business processes. In this position paper, we will argue for process-centric visual analytic metrics grounded in the concept of information worker productivity. We will place our discussion the context of legal e-discovery, the business process within which Attenex operates and within which we have demonstrated that visual analytic tools can increase productivity dramatically. After discussing how productivity metrics for visual analytics helped e-discovery, we make the argument that they can help any data-intensive business process and discuss both how to create these metrics and apply them successfully.

CR Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval] - Information Filtering, Search Process. I.6.9 [Visualization] - Information Visualization

Additional Keywords: process, productivity, visual analytics, information visualization, metrics, e-discovery, Attenex.

1 INTRODUCTION

The amount of new information stored on paper, film, and electronic media doubled between 1999 and 2002 [Lyman 2003]. Information workers need to understand, organize, and make decisions on these vast quantities of data. Interactive information visualizations help workers can shape and control the flow of information they receive. This process of visual analytics is widely used in many formal and informal settings.

How effective are current visual analytic tools at helping information workers make high quality decisions efficiently? We posit that any metric which only evaluates the tool, or the tool used in isolation by a single user, will not truly judge how effective that tool is. In order to be realistic, effective, and meaningful, a proposed visual analytic metric must be a process-centric metric, evaluating the level of productivity of the information worker while using the tool.

In this paper we discuss process and productivity in visual analytics through the domain of e-discovery. After reviewing visual analytics, we will describe our business context. We will then define and discuss our visual analytics metrics. Next we will show how visual analytics is mapped to e-discovery processes. Finally, we propose a methodology to create visual analytic productivity metrics for any data-intensive business process.

925 Fourth Avenue, Suite 1700, Seattle, WA 98104 USA
Email: smcnee@attenex.com, barnette@attenex.com

Submission to the Vis-2007 Visual Analytics Workshop

2 VISUAL ANALYTICS IN A NUTSHELL

Visual analytics is the science of analytical reasoning supported by interactive information visualizations [Thomas and Cook 2005]. The analytic reasoning process—the core of critical thinking—is the multi-step process of gathering, processing, integrating, and disseminating information to generate profound insights and make effective decisions.

In visual analytics this process is augmented or enhanced by the use of interactive information visualizations. The power comes from taking advantage of the high bandwidth between the eyes and the brain [Card et al. 2001]. These visualizations allow people to explore or 'play' with the data during analysis. Not only do visualizations help users review larger amounts of data, they also lead both to serendipitous discovery of connections in the data and help users develop a "visual intuition" about the nature of the data [Tufte 2001].

More specifically, visual analytics allows users to *shape* and *control* the information flow. *Shape* refers to the amount of information being used at each step of the analytic reasoning process. To shape the information flow means to alter the amount or kind of information used at each step. For example, 'zooming in' on a small subset of data, or 'zooming out' for a overview of a large amount of data. The terms 'converging' and 'diverging' have also been used to discuss the shaping of information flow.

The analytic reasoning process is fluid—users constantly create, challenge, and review hypotheses. Visualizations allow users to *control* this process, choosing where they are in the process and where to go next. It is a data-centric decision, with users organizing, comparing, and evaluating only the data needed for each hypothesis. This transforms the process into an *analytic discourse* where the user and the visualizations work together to discover profound insights and make effective decisions.

3 E-DISCOVERY

Our context is e-discovery: legal electronic-document discovery for litigation, regulatory requests, and investigations. "Discovery" means locating and organizing all relevant documents for a specific legal matter. At trial, copies of these documents are given to (a.k.a. "produced to") both the court and opposing council. In essence, by understanding the documents, you are "discovering" what the case is really about.

Once served with a complaint, corporations are legally obligated to find all related information. Privileged attorney-client communications and related "work products" are excluded. For example, if an attorney asks an analyst to perform a specific work task, all documentation related to that task is privileged. In the United States, the Federal Rules of Civil Procedure strictly define this process [109th U.S. Congress 2006].

Our company, Attenex, combines software, experts, and best practices to help corporations and their law firms establish electronic discovery solutions that reduce the risk, complexity and cost of litigation, regulatory requests, and internal investigations. Our software helps corporations perform e-discovery on time and under budget. Before we discuss how we do that, let's dig deeper into e-discovery.

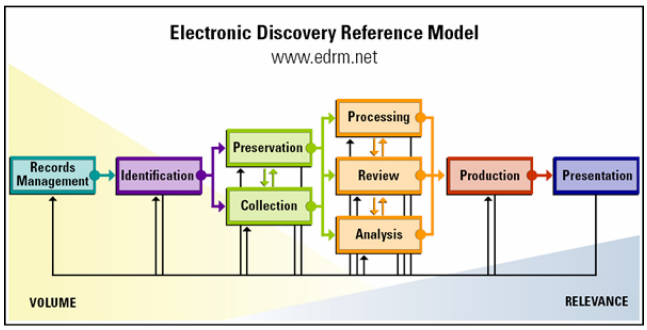

Figure 1: The Electronic Discovery Reference Model

3.1 The E-Discovery Reference Model

Traditionally, the discovery process was performed by teams of lawyers and paralegals reviewing boxes of paper by hand. This approach does not scale. E-discovery describes a process that maintains the digital integrity of information throughout the entire legal matter.

To address industry best practices in managing digital files, the Electronic Discovery Reference Model Project (EDRM) was created in May of 2005. It helped develop a "common, flexible and extensible framework for the development, selection, evaluation and use of electronic discovery products and services" [Socha and Gelbmann 2005].

The EDRM reference model is shown in Figure 1. The model flows from left to right, starting with available electronic information ("Records Management") and ends with the documents presented in court ("Presentation").

Once served, a corporation needs to identify and collect all relevant electronic information ("Identification", "Preservation", and "Collection"). After collection, the information needs to be processed and reviewed to determine what is relevant and what is privileged ("Processing", "Review", "Analysis"). These three orange steps are the most time consuming and error-prone part of the process; they also happen to be the most critical.

Once finalized, the relevant documents need to be prepared to be delivered to opposing council and the court ("Production"). This process includes ordering and organizing the documents as well as redacting non-related content (e.g. attorney-client privileged information, medical patient information, or other confidential information not related to the current case).

3.2 Document Review

We want to focus on document review (a.k.a. the three orange boxes in EDRM). It is here when individual documents are reviewed and a decision is made about relevance and/or privilege. It is the heart and soul of e-discovery—the critical path on which all documents travel.

This is where the analytic reasoning process comes into play. Our information worker is a legal worker, such as lawyer or paralegal. These workers are the scarce resource, and their time is extremely expensive. There may be dozens or hundreds working in parallel for any given case, but each one is critical to the success of the case.

While 'Review' looks like one orange box, it is multi-step process itself, usually including some variation on the following:

1. *"Meet and Greet":* The Federal Rules for Civil Procedure state that both sides must meet within 100 days of the suit being filed to agree on an e-discovery review strategy [109th U.S. Congress 2006, Srivastava 2007]
2. *First Pass Review:* Determine which documents are relevant and which are privileged; it is mostly culling what is grossly irrelevant.
3. *Quality Control:* The First Pass review is rechecked for accuracy and consistency
4. *Second Pass Review (Case-building):* Determine how documents are relevant and prioritize documents, build the legal strategy for the case
5. *Final Quality Control:* Verify privileged documents
6. *Prepare for Production:* Organize and redact documents

Several things can affect the review. Decisions made by individual reviewers are propagated along the process. Incorrect decisions during the First Pass can delay further action. Other regulations also may affect the review. For example, medical records must follow HIPPA rules; financial records must follow Sarbanes-Oxley regulations.[1] Finally, the opposing council many change the scope of their information request; new and/or different documents may be needed.

3.3 Measuring Success: Productivity as a Metric

E-discovery is big business, with billions of dollars spent every year [Murphy 2006]. With so much at stake, how do you measure success in e-discovery? There are many potential measures.

Perhaps winning the case is the metric. Performing e-discovery superbly cannot win a case; it just prepares the legal team argue the facts of the case. Performing e-discovery poorly, however, can lose the case [Schuman 2006].

Other traditional business metrics include the total cost of performing e-discovery and the total time taken. While useful at a global level, these metrics are hard to optimize; there are many ways to speed up a review process!

We need to focus and return to the business process itself. This process is data-centric and we can follow the information flow. Let's return to the critical point: the legal worker reviewing documents. For this worker, the important metric is *productivity*. If we can reliably measure and increase the productivity of each legal worker, we can reduce costs and increase quality of the entire e-discovery process. The better performing the critical path is, the better performing the entire process will be.

Our philosophy for choosing productivity as a metric comes from the work of Adrian Slywotzky. He states that improving productivity is core to sustaining successful business. Moreover, successful businesses move from burdening talent with low-value work to gaining high talent leverage [Slyworzky 2002]. That is, they get smart people to work at their potential for sustainable periods of time, and not deal with menial tasks.

The trick is in carefully defining the productivity metric. It has to be measurable and verifiable. Also, such a metric must support existing business processes. At Attenex, we define productivity in e-discovery as *document decisions per hour*. As a baseline, consider a linear review of the same material. We want to increase decision velocity while maintaining decision quality. Thus, the faster reviewers can make these decisions, the more productive they will be. By measuring this value during a review, we not only know how productive each reviewer is, we can also estimate both total financial cost and time to finish before the review ends.

[1] HIPPA is a US regulation related to the privacy of a patient's medical information, for details see http://www.hhs.gov/ocr/hipaa/. The Sarbanes-Oxley Act of 2002 is a US Federal Law related to how financial institutions conduct business and record financial transaction information, for details, see http://www.sec.gov/about/laws.shtml.

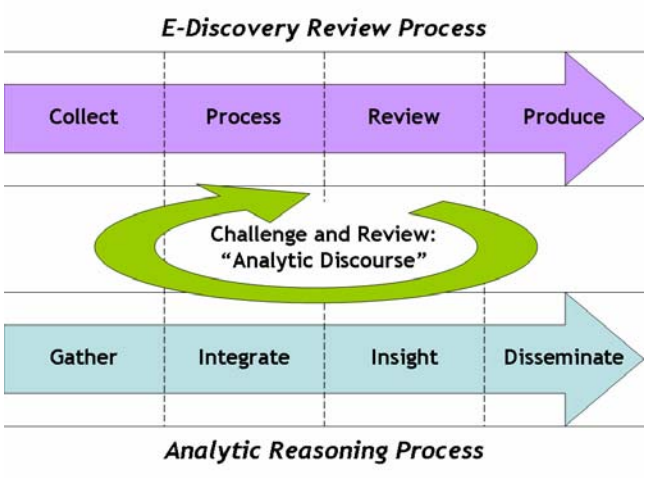

Figure 2: Mapping E-Discovery to Visual Analytics

4 Applying Visual Analytics to E-Discovery

The steps in an e-discovery review align with steps in the analytic reasoning process (see Figure 2). Our multi-step "review" box is aligned with the "insight" box for analytic reasoning. At both points, this is the critical step where a decision is made based on the evidence at hand.

Our goal is to achieve analytic discourse, where legal workers and the software work together to make decisions. This discourse happens at different levels. First, it happens between the software and a single worker as the worker explores relationships in the data to make more effective documents decisions. It also happens between reviewers: their decisions, along with any tags or comments are passed on to the next level of review. Finally, these decisions might be saved and used in future legal matters (once a document is determined to be privileged, it remains so independent of the legal matter).

The e-discovery process is also unique in several ways. Both the potential document pool and the time frame are strictly limited. The documents need to be related to the complaint; usually this limits to documents within a specific data range and/or from specific people. Moreover, the Federal Rules for Civil Procedure state that cases cannot drag; each side only has so much time within which to perform document review.

Another critical difference is that the legal workers might be legal experts but may be new to this case and new to the e-discovery software used. For each case, the workers are trained both on the facts of the case and the software. Yet this training cannot go on forever; there are documents to review!

Given this context, Slywotzky's imperatives, and our definition of productivity, we developed our software suite, Attenex Patterns, as a solution to the e-discovery document analysis and review problem. We have chosen a visual analytics approach to this problem, allowing legal workers to perform a non-linear review of documents by letting them review documents in concept-related clusters. Our software:

- Indexes and performs a semantic analysis on large quantities of electronic documents, including email, MS Office documents, and Adobe PDF files
- Finds and removes duplicate documents across a corpus as well as near-duplicate email threads
- Dynamically generates visualizations in the document concept space so that legal workers can quickly focus on the subset of documents relevant to the current legal matter
- Integrates multiple documents views along with decision making tools customized for the e-discovery domain

As shown in Figure 3, Attenex Patterns creates clusters of similar documents, where each document is a dot. Through this interface, a reviewer can see a document in the context of other similar documents. We have also integrated other visualizations, including a timeline view and a social networking view for emails. Changes made in any one view are automatically propagated to the other views. By exploiting locality, the reviewer makes better decisions faster for groups of related documents instead of documents in isolation.

The most significant part of our tool is the integration of e-discovery-specific decision features. By right-clicking on any document or cluster, a reviewer can "mark" a document as belong to a specific predetermined category, for example, whether a document is relevant or privileged. There are also user-defined tags and comments a reviewer can associated with each document.

In our interface, document exploration is directly linked with decision making; in other words, "play" has a purpose. As legal workers search for meaningful clusters, they can immediately mark and make notes as to why those documents are important. This frees them to continue exploring the space at will, marking as they go. At no point to we dictate how they explore—that is up to the individual reviewer. We just want them to be productive.

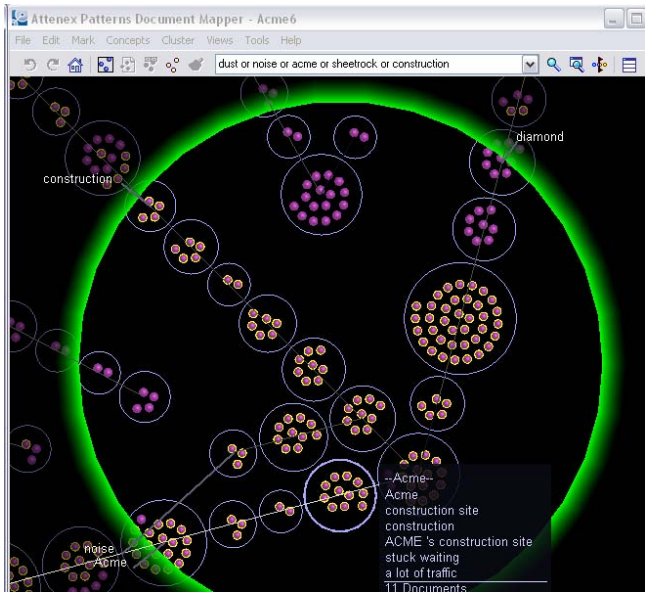

Figure 3: Clustering similar documents in Attenex Patterns

5 Process and Productivity in Visual Analytics

The Attenex productivity metric is linked to individual legal workers, determining how productive each is at reviewing documents. We can make both absolute comparisons (to a linear review) as well as make relative comparisons between legal workers. Moreover, we can use this metric to make predictive models as to when a review will finish and at what cost.

While our specific metric was created with e-discovery in mind, we believe the principles on which it was created translate back to visual analytics in general. It focuses on two core concepts: *process* and *productivity*.

While visual analytics itself contains the analytic reasoning process, this is not the process we care about. Instead, we look to

the existing business process within which visual analytics will be embedded. Is it an emergency response process? Is it fraud detection? Or is it generic business data analysis? In all cases, these processes have three important characteristics: they have existing business metrics by which their success is judged, there is localized and aggregated decision-making, and they have data-centric critical paths.

Wherever human decision-makers exist along this path is where the question of productivity should be asked. It is along these paths where visual analytics will have the most effect. The kinds of questions to ask are as follows:

1. What is this person's role?
2. How does the process define success in this position?
3. How does the person in the position define success?
4. How can this person be more productive?

The last question is subtle. It is obvious that we want people on the critical path to be more productive. The trick is that we cannot settle for incremental improvements in productivity—we need radical improvements. The question really is:

4. How can this person be *10 to 100 times* more productive than they are today? How can visual analytics help them achieve this goal?

This kind of radical questioning breaks traditional thinking about the business process itself; it gets people think about the process at different cognitive levels—it gets to the root problem [Stefik and Stefik 2004]. Every business process was created to solve some root problem. Radical thinking gets past the business process, back to this root problem. The metrics should measure progress against this problem. Not only do the metrics become clear, so do the solutions. For example, once you know that the metric is document decisions per hour, it is easy to design visual analytic systems against it. The hard part is knowing that you have found your true productivity metric.

The next step is to ask this same question for the entire business process. How can the entire process be 10 times better than it is today? At Attenex, we believe we have achieved well over 50 times productivity improvements compared to linear document review. We have saved our customers millions of dollars and months of review time. Any data-centric business process can be optimized in a similar fashion as long as the focus is on the productivity of the information workers engaged in the process.

6 DISCUSSION

Aggregation of information workers can amplify the gain in productivity, but only of all workers on the critical path are performing to the best of their ability. The independence of each worker in their decision-making duties will affect the performance of the entire business process. Processes with too many bottlenecks are candidates to be radically revised.

This paper has talked in terms of integrating visual analytic processes into a larger business context, into existing business needs. But they are both just processes. So why not embed them the other way? Why not put business processes inside of visual analytics? This is more common, actually. How do you deal with abnormal information or error cases? How do you delegate specific tasks? How do you escalate? These are common questions, and usually have specific, business answers. These questions also beg to have productivity metrics and radical solutions. One escalated problem could destroy the productivity for many individuals.

7 CONCLUSIONS

Forrester Research has noted that "tools with visual analytics built in can make these legal professionals more efficient by determining whether or not data is relevant, is privileged, or even needs to be produced in response to a discovery request." [Murphy 2006]. Visual analytics changed e-discovery by radically improving the productivity of individual legal workers. The alignment of the visual analytic process to the e-discovery process allowed us to create powerful visual analytic tools, embedded with domain specific process customizations.

The methodology by which Attenex was able achieve these gains is usable to the field of visual analytics as a whole. We recognize this methodology creates as many questions as it solves. The questions, however, are the right ones to ask. Because visual analytics is not used in a vacuum, there are no cut-and-dry metrics we can use for evaluation of visual analytic systems.

Rather, the metrics must measure the productivity of the information workers on the data-centric critical path of the business process. These metrics must relate to goals of the business processes in which the analytics are embedded as well as the root problems which the business processes were designed to solve. Visual analytic solutions designed to optimize these metrics will bring radical gains in productivity to both individual information workers and business processes as a whole.

REFERENCES

[1] 109th U.S. Congress, Committee on the Judiciary, F.J. Sensenbrenner, chair. Federal Rules for Civil Procedure. Washington, D.C.: U.S. Government Printing Office, 2006. Also available at http://www.uscourts.gov/rules/, last accessed September, 2007.

[2] S.K. Card, J.D. Mackinlay, and B. Shneiderman. Readings in Information Visualization. San Diego, CA: Morgan Kaufmann Academic Press, 1999.

[3] P. Lyman, and H.R. Varian, "How Much Information", 2003. Available at http://www.sims.berkeley.edu/how-much-info-2003, last accessed September, 2007.

[4] B. Murphy. "Believe It — eDiscovery Technology Spending To Top $4.8 Billion By 2011", Forrester Research, December 11, 2006.

[5] E. Schuman. "It's 2AM: Do You Know Where Your E-Mail Is?", eWeek Magazine, December 21, 2006. Available at http://www.eweek.com/article2/0,1895,2075504,00.asp, last accessed September, 2007.

[6] A. Slywotzky. The Art of Profitability. New York, NY: Warner Books, 2002.

[7] G.J. Socha, and T. Gelbmann. "The Electronic Discovery Reference Model Project (EDRM)", 2005. Available at http://www.edrm.net/, last accessed September, 2007.

[8] S. Srivastava, "Search Software Gets Boost from New Rules". The Wall Street Journal, May 16, 2007; Page B6.

[9] M. Stefik and B. Stefik. Breakthrough: Stories and Strategies of Radical Innovation. Cambridge, MA: MIT Press, 2004.

[10] J.J. Thomas and K.A. Cook. Illuminating the Path: The Research and Development Agenda for Visual Analytics. Los Alamitos, CA: IEEE Computer Society, 2005.

[11] E.R. Tufte. The Visual Display of Quantitative Information, Second Edition. Cheshire, CT: Graphics Press LLC, 2001.