SAND2018-7069C

Entropy and Insight: Using Complexity Reduction to Estimate Understanding

Sidney Holman
Human Factors Department
Sandia National Laboratories
Albuquerque, NM, USA
spholma@sandia.gov

Chris North
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA, USA
north@vt.edu

***Abstract*—A pressing issue in Visual Analytics is the question of how to evaluate and compare the tools we create. If the goal of visualization is to enable insight, it is natural to look at how well the tools facilitate insight in their users; however, with insight being a slippery concept to pin down, we instead focus on a prerequisite to insight: complexity. Building on the work by McNamara, Bauer, Haass, and Matzen, we have developed a method for estimating the complexity reduction in an analyst's workspace [1]. Further, we show that using Shannon's entropy and self-information values to provide a measure of the complexity reduction (resulting from an analyst's actions while sorting the information) can be linked to the knowledge gained by the analyst.**

***Index Terms*—Visual Analytics, Complexity Reduction, Insight, Information Theory, HCI, Sensemaking**

I. Introduction

The lack of a standard or baseline for comparing systems to each other has been a lingering problem in Visual Analytics (VA). This has been widely recognized, and attempts have been made to create a standard; however, they are typically time-consuming to use and plagued by potential validity issues [2]. Compounding the problem is the tendency to create evaluations that are geared towards specific visualization tools rather than addressing underlying user behaviors.

There have been serious efforts made to address this problem; case studies, known solutions, qualitative feedback, longitudinal studies, simple usability heuristics, and any number of other techniques have been attempted [3] [4]. Creating metrics for VA comes with a host of problems. Low-level tasks are specific and easy to measure, but most analytic workflows do not use the same set of tasks, or use them in a way that makes comparison difficult. High-level tasks are, by definition, more generalizable, but more difficult to measure. The process of sense-making may involve many iterations of, as Scholtz puts it, "onion peeling" or "pearl growing" the data, with only small changes made over time [2]. At the same time, the act of analyzing data is a task that spans many higher-level areas of activity.

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

This work builds on the paper by McNamara, Bauer, Haass, and Matzen, applying their ideas to the specific case of an intelligence analyst organizing documents during the sensemaking process. Ultimately, we hope to contribute an expanded understanding of how we can apply a semantically meaningless engineering approach (information theory) to the semantically-focused task of sensemaking and clustering, in a way that links to the understanding users have of the data they are investigating. Such a connection has potential applications in both evaluation of new systems, and in supporting the sensemaking process in real time by providing feedback on how well a workspace has been organized into semantically meaningful groups. In this research, we sought to establish a link between entropy and knowledge gain. In doing so, we hope to work toward a sensemaking baseline standard that could help in designing and using future VA tools.

II. Motivation and Context

Our work is motivated by the question, *Can we estimate insight generation by measuring the reduction in complexity of a workspace over time, using information theoretic measures of entropy and self-information?* As mentioned above, this question stems directly from [1]. However, there are a few other ideas that also form the basis of this paper.

A. The use of virtual space

At the intersection of technological advancement and the distributed cognition notion of external representations is the question of how people make use of virtual workspace. In 2010, [5] looked at how high-resolution displays affected visual analysis tasks. They found that being able to simultaneously place documents in space (physical navigation) while being able to see both the layout and details of the documents resulted in very different strategies. The old visualization mantra of "overview first, zoom and pan, details on demand" was no longer as applicable—groups of documents were positioned in clusters on the workspace to show collections of evidence, and their relationships. This suggests that the grouping of documents in space create a semantic layer that adds meaning to the data, and also serves as a medium for

creating complex structures (like clusters). As for the actual mechanisms involved, the authors note that the large display became a sort of external, easily-accessible memory, in the distributed cognition sense; users with the smaller, regular display were forced to use paper diagrams and notes for the same purpose.

In related work, [6] dug deeper into the meaning underlying the user-created clusters in a large digital workspace. They found that users organized the information into clusters (or other structures) based on a combination of items within the data (for example, entities: key people and places) and their own intuition and external/prior knowledge. At the same time, the follow-up interviews with participants found that the clusters created didn't usually have labels that describe entities or specific features contained in the clusters—which suggests users are pulling primarily from their external knowledge when organizing the workspace. There was virtually no clustering on the basis of a single term or entity; analysis of the clusters revealed that the best way to understand *why* documents were placed together was through "transitive terms". These are terms that occur in subsets of the documents within a cluster, and act as bridges between seemingly unrelated information.

B. Insight

Ultimately, the purpose behind studying the use of virtual space is to understand how it facilitates insight. But, there is disagreement on exactly what 'insight' really is. In fact, [7] points out that insight is an overloaded term—referring to either the event of understanding, or as a unit of information, for Information Visualization (InfoVis) and VA practitioners; alternatively, cognitive psychologists understand insight as the "Aha!" moment of understanding. For our work, we operationalize the idea of insight according to [8]. In our view, then, insight is the result of a deep understanding of complex data—and, as described by [1], complexity reduction ("separating the needle from the haystack") plays a central role in generating insights.

III. Method

A. Definitions

Before continuing, it is useful to provide definitions of the calculations underpinning our metric. Here, we will provide a brief description of how we are using ideas from Shannon's model of communication; for a more thorough explanation see [9] or [10]. In addition, we provide pseudocode of our technique, and a short description of metacognition and entity extraction.

a) Information Theory: Information Theory is concerned with the technical problem of transmitting messages through a channel. Messages, in this engineering view of the transmission problem, can be deconstructed (encoded) into a series of bits and sent to the receiver who will then reconstruct (decode) the message. An important thing to understand is that the encoding and decoding of the message is done using the same background reference data—both sender and receiver must have the same knowledge about the underlying data or decoding will fail. Information theory quantifies information in bits, and estimates the complexity of a dataset by how many bits are needed to encode the symbols. For example, if a message is always "X", then the complexity is very low. If instead the message can be any symbol from "A" to "Z", all with equal likelihood, the complexity of the message is much higher. Information Theory (via the Source Coding Theorem) is also the basis for our understanding of how much we can compress a message before we start to lose information.

b) Symbol: A symbol is a representation of some piece of knowledge. These symbols can be defined as anything: words, individual characters, sets of pixels, or notes in a tune. Symbols have some predetermined likelihood of appearing, in the discrete context being examined. The set of symbols, then, is tied to some specific probability distribution based on the scope we choose to look at.

$$p_x = \frac{\text{\# of occurences of symbol x}}{\text{\# of occurences of all symbols}} \tag{1}$$

In our work, the probability of the word occurring is calculated with respect to all the other words in the document, or group of documents. Changes to scope (adding or removing documents to the group) also changes the probabilities, so it is not meaningful to say something like "Document Q adds 15 bits of information to the group" —because we've changed the underlying probability distribution for all the symbols, we've changed the information value of all the symbols. Put plainly, the probability of the word "the" appearing depends on the documents in the set—and if you add a new document to the set, that probability will likely change.

c) Self-Information: The self-information (I_s) of a symbol 'x' is calculated from how often the symbol occurs in a set. I_s always refers to a symbol in a set.

I_s for 'x' is calculated as:

$$I_s = -log_2(p_x) = log_2(\frac{1}{p_x}) \tag{2}$$

and can be understood as the amount of information, in bits, needed to encode a single instance of this symbol in a message. For example, if a message could be one of five possible symbols, and all the symbols appear the same number of times, then $p_x = 1/5$ and the self-information of one symbol is:

$$I_s = log_2(\frac{1}{p_x}) = log_2(\frac{1}{1/5}) = log_2(5) = 2.32 \; bits \tag{3}$$

d) Message: A message is a subset of symbols in a given context. A message may not contain all the words in a set, but an unknown message has a chance of containing any of the symbols. Given that we don't know the contents of a possible message in advance, we need to be able to **estimate** the expected length (in bits) of a message given the number of symbols. We can do this with Shannon entropy.

e) Entropy: Entropy, denoted "H", is a measure of the **average** expected length, in bits, of the symbols that comprise a message; to estimate the size of the message, we will need to multiply by the number of symbols in the message. Entropy always refers to a set of symbols. It can be understood as an indication of the complexity of the vocabulary of a message– "yes" and "no" are simple, but trying to transmit a chapter of "Alice in Wonderland" would require much more complicated vocabulary.

Entropy is calculated as:

$$H = -\sum_{i=1}^{n} p_i * log_2(p_i) = \sum_{i=1}^{n} p_i * log_2(\frac{1}{p_i}) \quad (4)$$

where n is the number of different outcomes possible. To estimate a message with N symbols, simply multiply entropy by N: $H_{message} = H * N$. *In our work, n is the number of unique words in the set of documents in a cluster.*

f) Self-Information of a Message: The self-information of an entire message is, naturally, the sum of the self-information for all the symbols it contains. The self-information of a message is different from the entropy because it is applied to messages where the contents are known. For example, we would calculate the *self-information* for a message that has 12 symbols from a larger set, and we know what these symbols are; this would give the exact number of bits required to code the message. Keep in mind, both entropy and self-information require that we know the set of symbols and their frequency.

Self-information of a message is calculated as:

$$I_m = -\sum_{i=1}^{n} log_2(p_i) = \sum_{i=1}^{n} log_2(\frac{1}{p_i}) \quad (5)$$

where n is the number of words in the message.

As a side note: in these equations, there is the expectation that the probability of the events occurring are independent. While the frequency of words in English are not completely independent, we are treating them as such in our work. Future work could be done to understand how this impacts our metric.

g) Pseudocode for Metric: Our code for analyzing user-generated clusters is fairly straightforward, but still most easily explained with some pseudocode. The randomized score used for comparison is determined by counting the number of documents in a user-cluster, placing an equal number of randomly chosen documents into an empty cluster, calculating the entropy and self-info metrics (with some MetricValue function, 'MV') of this temporary cluster, and taking the average of each metric over a large number of N runs. The net change for users is the difference between the average of the random clusters and the user-clusters.

This gives us the average entropy or self-information (depending on the metric value function 'MV' that we choose) of a cluster, when the document set is spread randomly between groups. Though it loops through many times, this only needs to be done once for every document set—the average should not change significantly if we choose a large enough number for N. With this, we can calculate the net change in a cluster or set of clusters by simply subtracting the random average from the MV of the cluster/set. Thus, improvement in our metric will be shown as a net negative value for a cluster or cluster set:

Algorithm 1 Random average metric value of a user-cluster

Input: UserCluster, DocumentSet

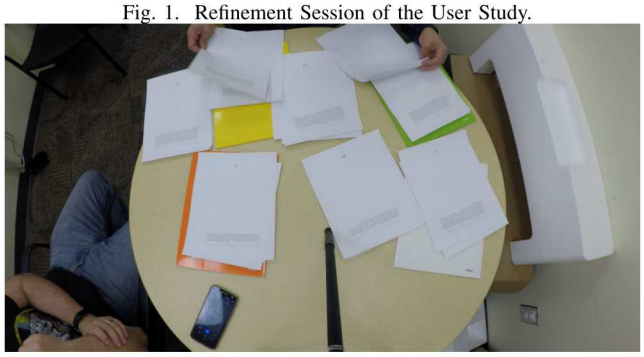
```
MetricSum = 0
for i = 0; i < N; i++ do
    create TempCluster
    for each document in UserCluster do
        select random document from DocumentSet
        copy document to TempCluster
    end for
    MetricSum+ = MV(TempCluster)
end for
Return MetricSum/TotalClusters
```

$$\textit{improvement} = MV(\textit{user-created}) - \textit{randomized} \quad (6)$$

h) Metacognition: In psychology, there have been many attempts to understand the processes that lead to insight. Metcalfe and Weibe [11] introduce the 'feeling of warmth' and 'feeling of knowing' scores, and show that we are largely unable to predict the "spontaneous" insights—but there was some predictive power for more procedural proplem-solving. Durso et al. [12] build on this work, and suggest that while the flash of insight often occurs without conscious warning, there is often a long trail of conceptual changes that take place. While there is considerable variance in how successful we are at predicting our insights, analysis done by Flemming and Lau [13] indicates that measuring a simple high/low confidence condition in repeated trials provides a reliable way of characterizing individuals' metacognitive efficiency (how good they are at thinking about their thinking).

i) Stop Words and Entity Extraction: An important variable that can be manipulated during this phase of our investigation is our 'symbol set', or what words we want to include when calculating entropy and self-information. Many visual analytics projects identify the 'named entities' found in documents—things like people, places, dates, and so on—that form the nodes in the narrative schemes people pull together during their sensemaking process. With the basis of our research being built on some notion of "transitive terms' [6], we want to understand how these entities may relate to our metrics. To automatically identify and extract the entities we used the Named Entity Recognizer, part of the Stanford Natural Language Processing software [14]. In our case, we use the default model that comes with the Stanford NER software, which is geared towards English-language person, organization, and location recognition. As a sort of intermediate level between the full text and the extracted entities, we created a third version of all the documents that has all of the 'stop words' removed—for example, 'an', 'while', 'because', and so on. While this may seem like an unnecessary step

Fig. 1. Refinement Session of the User Study.



when already looking at entity extraction, it has the advantage of retaining relevant words like 'explosive' that might get removed by the default Stanford NER, while still removing some of the noise that might be present in the full text of documents.

B. User Study

To investigate insight, we have conducted a user study. Leaning on precedent ([15] [16]), our investigation of insight used a Think-Aloud approach, and also included a metacognitive measure of participants' confidence in each move of a document. The document set is a modified version of an intelligence analyst training dataset ("Sign of the Crescent"), with distractors and junk documents removed. It also has a known correct grouping of documents, making later comparisons to confidence and think-aloud comments easier.

Our study was conducted in the spring of 2018, with a total of 12 participants. Ten were pulled from undergraduate computer science courses, and the other two were graduate students who were passingly familiar with the dataset used in the study. On average, each session lasted approximately 40 minutes, with 25 spent actually manipulating the groups and the rest spent on introducing the task or explaining the details of the uncovered plot. The sessions were filmed in 4K video, from the introduction of the documents to the participants, until the end of the second round of refinements in grouping—where participants explained, in as much detail as they could, their understanding of the terrorist plot described in the dataset. The setting, shown in Figure 1, consisted of a small desk with 4 differently-colored folders on top; the colored folders were chosen both to provide a visual cue to participants while they organized the documents, and to be visibly distinct while reviewing the recorded session at a later time.

Participants were seated with the shuffled stack of 20 modified documents that comprise the dataset, and the 4 unlabeled folders that served as clusters. Each document is a short intelligence report, listing the date and organization it was gathered from (CIA, FBI, etc), and has a large number at the bottom for tracking group membership from the recording. The documents were randomly ordered for each participant, and were not readable until the initial session has begun. Participants were asked to sort the documents into groups based on their content, with the goal of explaining the "who, what, when, where, and how" questions surrounding the terrorist plots. In our prompt we encouraged them to make use of all four folders in developing their groups, and asked that with each move of a document between groups they give some indication of their confidence that the current document belonged with the existing documents in the group. While working on this task, participants were asked to both verbalize their thought process (Think Aloud), and provide verbal feedback on their level of confidence (high, or low) on the cluster membership after each move. To reduce the subjectivity of "levels of confidence" to something that could be roughly analyzed, participants were directed to use "high" or "low" to indicate their confidence; in cases where users habitually slipped into a percentage description, we considered variations such as "90% confident", or "pretty confident" to also indicate high confidence.

The cluster memberships were tracked over time by manual review of the recording, and at the end of the session participants were asked to summarize the clusters and explain why they chose the groupings they did. The sensemaking and grouping process was broken into two time sessions: initially, users were given 15 minutes to read and quickly cluster the documents based on incomplete understanding of the entire dataset. A second, 10 minute period was given to allow them to refine their groupings based on their new understanding of the document contents. In addition to the high/low confidence ratings, these two time periods provide the opportunity to compare our information-theoretic measures to the participants' understanding over time—rather than relying on the high/low scale, we can contrast their first-session answers (score) about the plot to what they understood at the end of the second session.

Both the initial and refinement sessions were scored, using a scheme that requires increasing understanding of the plot details to earn more points. When used flexibly, the scoring system allowed us to award points for plot details mentioned both during the sessions, and during the summaries that followed. As mentioned, this scheme increased the difficulty of each point earned; trivial points were earned for correctly grouping reports on a single plot together, where subsequent points required explanation of general goals within each group—and for the maximum of 20 points, very specific details must have been understood and verbalized. For example, 1 point would be given for grouping all of the reports related to Amsterdam, another point for understanding that there was some plot to bomb something related to shipping, and another point each for recognizing that they were going to use a dirty bomb, and were targeting Boston.

IV. Results

We calculate entropy and self-information of each user-created cluster with equations 4 and 5, respectively. Our expectation was that the information-theoretic measures will decrease (show improvement) over time with respect to the average of the randomly shuffled documents, as described in

Fig. 2. Users' Score after 15 and 25 Minutes.

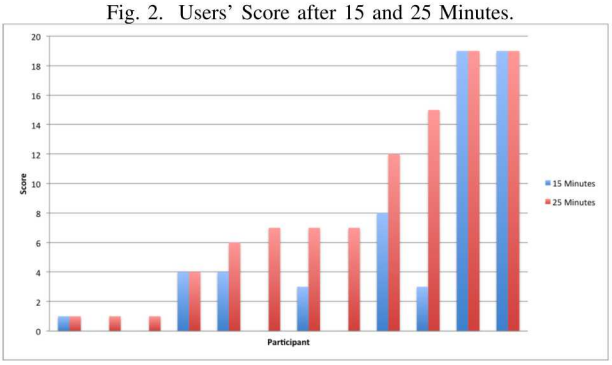
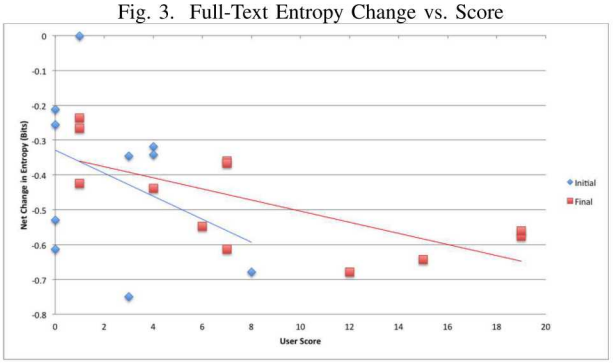


Fig. 3. Full-Text Entropy Change vs. Score



section III-A. We also expected that the think-aloud feedback and confidence scores would indicate increased understanding of the document set as the users spend more time with understanding the plot. However, initial analysis shows that the metacognitive measure (high/low confidence scores) does not correlate with the points the participants earn by stating their understanding out loud. Additionally, reviewing the sessions reveals that the metacognitive confidence scores also fail to correspond with meaningful groupings of documents.

The goal of the user study is to determine if the complexity reductions that we can detect in the workspace correspond with increases in users' insight, but with the failure of our fine-grained measure of understanding, we will instead rely on the coarse time periods of the initial and refinement sessions for analysis. These coarse time periods are evaluated with a user score, which we will compare to our information theoretic metrics rather than the high:low confidence ratio originally planned; this change will be discussed in more detail in section V. These results *do* show improvement over time: after the initial session, the average score is 2.3 points (out of 20 possible). After the refinement session the average score is 8.25 points. Figure 2 shows the scores for each participant, between the two sessions. Two participants completed their analysis of the documents in only one round, so their sessions are considered part of the final session group.

As described in section III, there are three ways to treat the document set—we can use the full text, or remove the "stop words", or use only the extracted entities. Here, we use all three approaches on both the entropy and self-information scores, and plot the net change in our metric to the score earned with that configuration of groups.

Our metric values show a strong link between score and entropy (Figures 3, 4, and 5). The self-information metric also demonstrates the connection we were looking for (Figures 6, 7, and 8), and the relationship seems to grow stronger with the better-targeted sets of symbols (no stop words, or just entities). While the small sample and qualitative scoring system doesn't lend to a meaningful statistical analysis, we do consider these results to support our hypothesis that decreased information-theoretic metrics correspond to improved insight.

We also found that our method resulted in mutual improvements between the entropy and self-information metrics. Com-

Fig. 4. No Stop-Words Entropy Change vs. Score

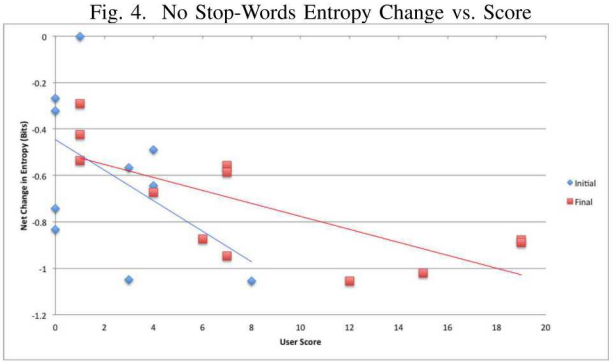


Fig. 5. Entities-Only Entropy Change vs. Score

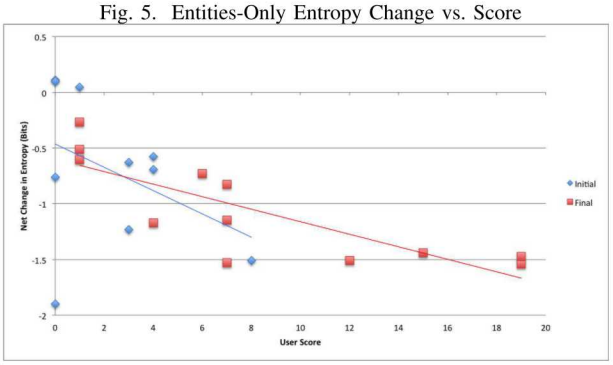


Fig. 6. Full-Text Self-Info Change vs. Score

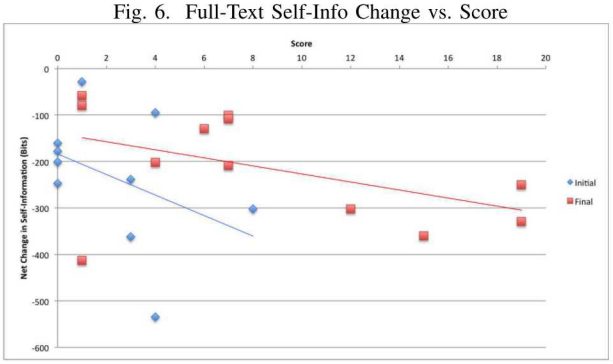


Fig. 7. No Stop-Words Self-Info Change vs. Score

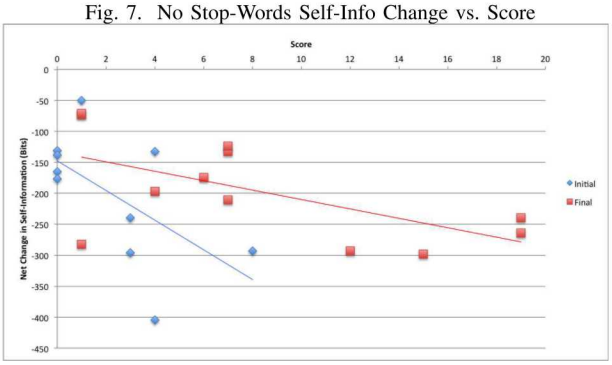
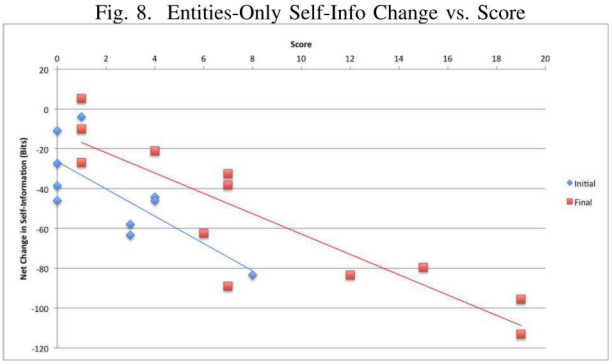


Fig. 8. Entities-Only Self-Info Change vs. Score



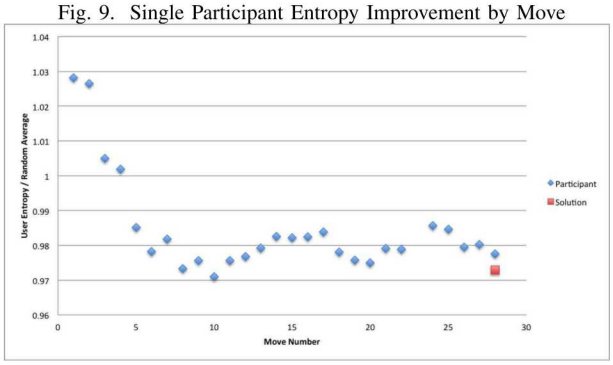
paring the net improvement across clusters, there is between a 2.5% to 6.4% reduction in entropy and self-information in the sorted document groups (shown in Table I). This means, for example, that the entities-only symbol set sorted into clusters that were, on average, only 93.6% of the size of the un-sorted, random clusters. Further, entropy and self-information scores improve by approximately the same amount for each symbol set used.

Finally, because we control for the number of documents in each user-created cluster, we can use this method to plot the net change in a user's groups, move-by-move. Starting at the first instance of a 4-cluster workspace (for consistency of comparison), we calculate the size of the clusters and show the net reduction as a percentage of the randomized average. Our results, demonstrated in Figure 9 for a single participant, indicate that as participants spend time making sense of the documents, the entropy of their clusters decreased further with respect to the randomized average. The large decreases at the beginning of the process likely stem from the inclusion of more documents to the workspace—as the users read and place reports, there is both larger "random average" entropy

TABLE I
% VS. RANDOM, ENTROPY AND SELF-INFORMATION

Symbol Set	Entropy	Self-Info
Full Text	97.6	97.3
No Stop Words	96.4	96.1
Entities Only	93.6	94.9

Fig. 9. Single Participant Entropy Improvement by Move



AND more similarity in symbols that could drive entropy reduction. Review of the session video shows that, although the participant did not verbally indicate strong connections, there are some non-verbal cues that suggest moments of insight (excited finger tapping, rapid scanning through other documents) that happen in conjunction with the entropy-reducing moves. This suggest that adding documents to the workspace drives up the entropy, while the consecutive moves that result from increased understanding drives the entropy down. These moves also often move documents together in a way that matches the known solution, so it is tempting to call these "good" moves. This reinforces the results of the large-time-step, initial-versus-refinement session analysis described by Figures 3 through 8.

V. DISCUSSION

Our results demonstrate that the entropy of a user's clusters decreases as they group the documents, and that improved sensemaking scores correspond to lower entropy. This supports our hypothesis that the measures are linked to understanding, but it comes with some qualifications. First, we focus on a text-based sensemaking task, in a workspace with some inherent affordance of spatial grouping; applications to other domains need further investigation. More critically, our initial goal was to link *insight*, not simply understanding. There is doubtless a great amount of overlap between these two ideas, but to actually pinpoint instances of insight is more difficult with the failure of our metacognitive measure—we must use the think-aloud results, and hope that participants verbalized their thought processes well enough that we can identify those 'Aha!' moments. Reviewing the sessions, it seems that the weakness of our metacognitive measure was in the open-ended nature of our task; users were creating a semantic structure for the documents, then assigning confidence based on how well each document fit into *their own* structure, rather than how well the grouping helps understand the plot. This means that, for example, someone may put 15 of the 20 documents into one pile with high confidence, because they are all FBI reports.

The fact that unusually low net entropy scores can be achieved by very low-scoring participants (as seen in Figure 3, for example) suggests that our approach must be interpreted with respect to the semantic meaning created during the

process; an automated system may manage to minimize the net entropy, but without a human in the loop there may be no meaning to be gleaned from the groups. This also touches on the issue of useful truths versus useless truths: a person might group documents together in a way that is technically correct, according to a scheme that does nothing to help them make sense of the information. An example would be grouping documents by month—it's correct, but possibly not useful.

There is still much more that can be investigated or implemented. Our intent, when starting this exploration, was to uncover a method for evaluating visual analytics tools that could be compared across systems; our research hints at a solution, but much more implementation and testing is required before we could claim progress in this area—most notably, a user study with a much larger set of documents. A potential application is using cluster compressiblity (entropy) indicators as a feedback tool for analysts in real time, perhaps encoded as a visual indication of how well a document might relate to the existing contents of a group—added to systems like those by Wenskovitch and North [17] or Endert et al. [6]. For example, in an interactive VA system like StarSPIRE [18] or Jigsaw [19], some color coding might be shown while interacting with objects in the spatial/graph layouts; moving an object near others could provide an indication of the relative improvement in complexity achieved by grouping the objects together. Alternatively, understanding how regions of a workspace with spatial layouts have their entropy or self-information change as a person works might be used to provide feedback to the machine—perhaps suggesting what documents a user does, or doesn't, find useful. Or, perhaps these values can be adapted for use as a weighting metric, to directly manipulate the layout of the workspace.

Finally, it is worth emphasizing that this work focused on a fairly narrow range of the visual analytics problem space. Text-based intelligence analysis is far from the only task that VA is called on to support, and we feel that there are potential applications of information theory beyond this task. As we know, almost anything that can be represented in a digital format can be compressed and transmitted—to apply this metric to a domain other than text-based analytics, we will need to give some thought to what the meaningful atoms are in a particular representation. Where we use words as symbols in this exploration, perhaps genes would work as symbols in another application; in yet another situation, maybe various color palettes would be appropriate as symbols. Where we use explicit groupings in folders to define what constitutes a cluster, maybe the pattern of access would be a more appropriate delineation in another application. It seems like a topic full of potential for further exploration.

VI. Conclusion

Measuring complexity means analyzing the documents that an analyst is working to understand, and assigning value to the contents in such a way that grouping related information together produces a detectable change. Shannon's information-theoretic notions of entropy and self-information provide just such measures of complexity. We've shown that these measures *do* change as users sort and semantically structure their workspace, and we've shown a link between the measures and the users' performance on a sensemaking task.

Complications with our approach to tracking insight means we were not able to get the fine-grained look at how a user's "aha!" moments (described in [7]) correspond with our information-theoretic measures. However, if we consider insight to be the more gradual, knowledge-building process described by North [8] we show that increases in insight *do* correspond with improvements in our measures. Taken collectively, our work suggests that we can potentially quantify insight generation via Shannon entropy and self-information measures.

Put concisely, we demonstrate that the entropy and self-information values of clusters in a workspace improve as documents are put into semantically meaningful groups by users.

Acknowledgments

S.H. thanks Laura McNamara for guidance and feedback throughout this work, Scott McCrickard and Steve Harrison for the constant support they provided as part of my advisory committee, and many colleagues at Sandia Labs and Virginia Tech for their patience and advice.

References

[1] L. A. McNamara, T. L. Bauer, M. Haass, and L. Matzen, "Information theoretic measures for visual analytics: The silver ticket?" in *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, ser. BELIV '16. New York, NY, USA: ACM, 2016, pp. 53–61. [Online]. Available: http://doi.acm.org/10.1145/2993901.2993920

[2] J. Scholtz, "Beyond usability: Evaluation aspects of visual analytic environments," in *Visual Analytics Science and Technology, 2006 IEEE Symposium On*. IEEE, 2006, pp. 145–150.

[3] Y. a. Kang and J. Stasko, "Examining the use of a visual analytics system for sensemaking tasks: Case studies with domain experts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2869–2878, Dec 2012.

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

[5] C. Andrews, A. Endert, and C. North, "Space to think: large high-resolution displays for sensemaking," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 55–64.

[6] A. Endert, S. Fox, D. Maiti, S. Leman, and C. North, "The semantics of clustering: Analysis of user-generated spatializations of text documents," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ser. AVI '12. New York, NY, USA: ACM, 2012, pp. 555–562. [Online]. Available: http://doi.acm.org/10.1145/2254556.2254660

[7] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky, "Defining insight for visual analytics," *IEEE Computer Graphics and Applications*, vol. 29, no. 2, pp. 14–17, March 2009.

[8] C. North, "Toward measuring visualization insight," *IEEE Computer Graphics and Applications*, vol. 26, no. 3, pp. 6–9, May 2006.

[9] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, Jan. 2001. [Online]. Available: http://doi.acm.org/10.1145/584091.584093

[10] T. L. Bauer, "Information and meaning: Revisiting shannon's theory of communication and extending it to address today's technical problems," Sandia National Laboratories, Tech. Rep., 2009.

[11] J. Metcalfe and D. Wiebe, "Intuition in insight and noninsight problem solving," *Memory & cognition*, vol. 15, no. 3, pp. 238–246, 1987.

[12] F. T. Durso, C. B. Rea, and T. Dayton, "Graph-theoretic confirmation of restructuring during insight," *Psychological Science*, vol. 5, no. 2, pp. 94–98, 1994.

[13] S. M. Fleming and H. C. Lau, "How to measure metacognition," *Frontiers in human neuroscience*, vol. 8, 2014.

[14] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[15] K. A. Young, "Direct from the source: the value of 'think aloud' data in understanding learning," *The Journal of Educational Enquiry*, 2005.

[16] J. Nielsen, T. Clemmensen, and C. Yssing, "Getting access to what goes on in people's heads?: Reflections on the think-aloud technique," in *Proceedings of the Second Nordic Conference on Human-computer Interaction*, ser. NordiCHI '02. New York, NY, USA: ACM, 2002, pp. 101–110. [Online]. Available: http://doi.acm.org/10.1145/572020.572033

[17] J. Wenskovitch and C. North, "Observation-level interaction with clustering and dimension reduction algorithms," in *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics*, ser. HILDA'17. New York, NY, USA: ACM, 2017, pp. 14:1–14:6. [Online]. Available: http://doi.acm.org/10.1145/3077257.3077259

[18] L. Bradel, C. North, and L. House, "Multi-model semantic interaction for text analytics," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE, 2014, pp. 163–172.

[19] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: supporting investigative analysis through interactive visualization," *Information visualization*, vol. 7, no. 2, pp. 118–132, 2008.