The Effect of Semantic Interaction on Foraging in Text Analysis

John Wenskovitch* Virginia Tech Computer Science
Lauren Bradel† Department of Defense
Michelle Dowling‡ Virginia Tech Computer Science
Leanna House§ Virginia Tech Statistics
Chris North¶ Virginia Tech Computer Science

ABSTRACT

Completing text analysis tasks is a continuous sensemaking loop of foraging for information and incrementally synthesizing it into hypotheses. Past research has shown the advantages of using spatial workspaces as a means for synthesizing information through externalizing hypotheses and creating spatial schemas. However, spatializing the entirety of datasets becomes prohibitive as the number of documents available to the analysts grows, particularly when only a small subset are relevant to the task at hand. StarSPIRE is a visual analytics tool designed to explore collections of documents, leveraging users' semantic interactions to steer (1) a synthesis model that aids in document layout, and (2) a foraging model to automatically retrieve new relevant information. In contrast to traditional keyword search foraging (KSF), "semantic interaction foraging" (SIF) occurs as a result of the user's synthesis actions. To quantify the value of semantic interaction foraging, we use StarSPIRE to evaluate its utility for an intelligence analysis sensemaking task. Semantic interaction foraging accounted for 26% of useful documents found, and it also resulted in increased synthesis interactions and improved sensemaking task performance by users in comparison to only using keyword search.

Index Terms: Human-centered computing—Visualization—Empirical studies in visualization; Human-centered computing—Visualization—Visual analytics

1 INTRODUCTION

Prior research has highlighted the utility of spatializations to support the sensemaking process for text analysis [4–6, 11, 16, 21, 23, 30, 34, 48, 52, 53]. By providing a continuous physical workspace, analysts can externalize their hypotheses and organize data into meaningful schemas. However, manually arranging documents is a tedious and time-consuming task. Analysts must read each document and assess its relevance before deciding where the text belongs in an incrementally evolving spatialization. This task is exacerbated in realistic sensemaking scenarios because datasets are rarely small enough to display in full, even on a large, high-resolution display. Additionally, only a small subset of available documents is typically relevant to the analyst's sensemaking task. Analysts must then apply a combination of searching for documents and organizing them spatially. More specifically, analysts are tasked with two primary challenges: foraging for relevant information, and synthesizing the information into a coherent structure and narrative [10, 35].

These foraging and synthesizing tasks are combined in the visual analytics tool StarSPIRE [12], which uses a spatial metaphor to serve as a means of communicating with underlying document relevance and spatial layout models. As the analyst synthesizes information, StarSPIRE encodes their interactions in the workspace to update an underlying user model that captures the analyst's interest foci quantitatively. These are **semantic interactions** in the sense that they directly reflect the analyst's analytical thought process about the meaning of the data (such as organizing documents, highlighting and annotating text, etc.), rather than about manipulating model parameters (e.g., sliders on keyword weights). The user model is then used to support the foraging and synthesis processes.

To support the foraging process, the updated user model is used to determine document relevance and to curate the working set of documents displayed in the workspace. Therefore, in addition to allowing for traditional ***keyword search foraging (KSF)*** for documents (i.e., a user types in keywords and retrieves relevant documents), the updated user model initiates ***semantic interaction foraging (SIF)*** *to automatically forage for documents that may be relevant to the analyst*. SIF displays new documents that the model infers may be of interest to the analyst based on their prior synthesis actions. To support the synthesis process, the updated user model is also used to adjust the spatial layout, allowing the analyst to organize and visualize the working set using a "proximity $\approx$ similarity" metaphor [20].

These two processes work together in a contextual manner. Synthesis actions by the analyst within the spatial workspace (*contextual input*) serve to initiate SIF algorithms, and the resulting newly-foraged documents are automatically positioned within the space (*contextual output*) by the synthesis layout algorithm.

This capability for SIF raises several research questions. Does SIF retrieve useful relevant information? Does it retrieve information that might not be found using KSF alone? How does it affect analysts' interactions, sensemaking process, and analytic performance? To evaluate the utility of semantic interaction foraging for sensemaking tasks, in particular the translation of semantic interactions into SIF, we conducted a comparative user study using a text dataset with a known ground truth from the VAST 2007 Challenge [36]. For foraging, the control condition offered only KSF. The experimental condition also offered SIF in addition to KSF.

We found in this study that KSF and SIF are complementary foraging techniques, each with benefits and limitations regarding the set of documents that each are best at retrieving. We found that the introduction of SIF into StarSPIRE led to a boost in participant comprehension of the scenario in the study dataset, led to an increase in the number of user interactions with the workspace, and led to the discovery of some relevant documents that were rarely located by KSF alone. SIF shows clear effects on which documents participants retrieved, how these documents were retrieved, how the participants interacted with these documents, and the overall information synthesis of the participants.

The contributions of this paper are:

- The design and results of a study to determine the effects of SIF on the sensemaking process using StarSPIRE.
- An analysis of the study results to understand how SIF can benefit the exploration of large document collections.
- Reflections on using KSF and SIF in visual analytics systems.

2 RELATED WORK

2.1 Semantic Interaction

Previous work has demonstrated the success of semantic interaction for manipulating underlying models (e.g., force-directed, multidimensional scaling) to shield users from the complexity of these

*e-mail: jw87@vt.edu
†e-mail: lcbrade@nsa.gov
‡e-mail: dowlingm@vt.edu
§e-mail: lhouse@vt.edu
¶e-mail: north@cs.vt.edu

algorithms [22]. By manipulating the data instead of altering model parameters explicitly, users are able to maintain focus on their analyses, thus staying in the "cognitive zone" [17, 26]. Similar techniques have also been proposed in the user modeling community [2, 3].

Inspired by PNNL's IN-SPIRE [37, 51], systems such as ForceSPIRE [20] and StarSPIRE [12] allow users to directly manipulate data points, which are then translated to parametric model feedback. Dis-Function [14] and Andromeda [41] follow a similar approach with quantitative data. These systems are limited by the size of the datasets that can be analyzed. As the number of data points and/or the data dimensionality increases, the execution time of the spatial layout models increases to the point where a quick interaction-feedback loop is no longer supported.

2.2 Visualizing and Interacting with Text

To visualize large text corpora, Typograph [19] uses varying levels of data abstraction by utilizing extracted topics, keywords, and document snippets. Users can drill down to see the documents at different levels of detail. The multi-model semantic interaction technique in StarSPIRE, in comparison, addresses the scalability challenge by continually updating a small working set of documents. Documents in StarSPIRE are either not present, iconified, or open. We previously presented a visualization pipeline that outlines how interactions are captured, interpreted, and leveraged to compose a working set of documents to visualize [12]. The multi-model visualization pipeline demonstrates how models can be interchanged to best suit the analyst's needs [12]. This pipeline was previously demonstrated using a display layout and a document relevance model, but could easily be extended to include clustering [47], large-scale information retrieval [25], or data streaming and sampling algorithms. For example, Vizster combines a clustering algorithm and a graph layout algorithm to visualize social networks [29].

Work by Ruotsalo et al. has demonstrated the use of direct manipulation to influence information retrieval algorithms [39]. User interactions within a radial topic spatialization were used to infer possible user intent and thereby tune search results, working on the principle that searches evolve incrementally [44]. This is similar to the incremental formalism seen in sensemaking and spatial organization [43]. They found that these interactions did not replace the need for conducting traditional keyword searches, but that the users in the condition that allowed for the use of the spatial interface performed better than those who did not have this technique available. These results closely mimic the results of our user study – inferring user interests through interactions in a spatialization does not replace KSF, yet it augments the underlying models, allowing users to identify more pertinent pieces of information.

2.3 Foraging for Text

Other systems provide mechanisms for visualizing search results beyond the typical ranked list (e.g., term distribution charts [28], self-organizing semantic maps [31]), but these methods have not received widespread adoption and do not provide the nuanced spatial interactions that Intent Radar does [39]. While ranked lists are well-suited to narrow and specific searches, they may not be as well-suited for complex sensemaking tasks. For example, conducting a literature review requires exploring multiple facets of a topic. A simple ranked list of results does not yield insight into documents that are mixtures of different topics. Thus, recommendation systems typically separate foraging and synthesis, presenting results in a separate list. However, StarSPIRE integrates recommendation systems into the sensemaking process by placing recommendations in context with the user's current analytical workspace.

2.4 Recommendation Systems

Recommendation systems work by assigning a predicted "rating" or "preference" score to individual items based on the relevance of that item to an analyst [38]. StarSPIRE falls under the "content-based filtering" approach to recommendation systems, in which these preference scores are determined by profiles of both the item in question and the user exploring the collection of all items [15].

The foraging engine of StarSPIRE is also closely related to query-by-example systems, which utilize a set of user-defined query objects. Query-by-example systems can be found in the literature across many types of data, including unstructured text documents [8], multimedia [27, 40], and musical selections [24].

Our intent with this study was not to create a new algorithm for a recommendation system; rather, we sought to evaluate the use of semantic interaction techniques in support of document recommendations. While the StarSPIRE foraging backend is relatively simple, the weights applied to each category of semantic interaction allow for ease of experimentation during the development of the system and can be tuned to each scenario. In the future, these weights could be learned either automatically or based on a large-scale study with additional datasets. We assert that many recommendation systems could be used as a foraging backend to StarSPIRE, which should give even better performance than the heuristic system described in Section 3 and Table 1.

3 StarSPIRE Design

StarSPIRE is a visual analytics system prototype developed to demonstrate semantic interaction with SIF. Many of the implementation details for StarSPIRE can be found in [12], though we briefly summarize the components relevant to the study here. In particular, StarSPIRE contains the following concepts:

1. A *working set* of documents, extracted from a universal set by an *information retrieval model* and *relevance threshold*, representative of the foraging process. This model computes the relevance of a document as a combination of the extracted entities within each document and the term weights in the user interest model. This relevance calculation combined with a threshold serves as a filter for which documents are displayed in the workspace.
2. A *spatialization* of the working set of documents, organized by a *spatial display layout model*, representative of the synthesis process. This model computes a weighted, force-directed layout of the documents, with a document similarity function of co-occurring terms weighted by the term weights in the user interest model. The model places similar documents nearer each other in the layout.
3. A high-dimensional *user interest model*, learned from the user's semantic interactions on the working set and spatialization. The model consists of weights on terms to represent the user's interest level. The user model is input to the retrieval and layout model algorithms.
4. *SIF* occurs as a result of semantic interactions that update the user interest model, which is then input into the retrieval model, thereby updating the current working set that is displayed on screen by the layout model. In contrast, KSF bypasses the interest model and directly manipulates the working set.

StarSPIRE (Fig. 1) provides users with a spatial workspace to view and incrementally arrange documents in a large display space (similar to the Analyst's Workspace [6]). Documents are visualized using a node-link diagram, and are shown as iconified nodes or as open text windows. To avoid a cluttered workspace, edges linking documents (based on term co-occurrence) are only shown radiating from the currently selected node. We designed a set of semantic interactions (some of which are listed in Table 1) by observing real-world analysts who offered usability feedback in informal and formal test settings to tune the parameters. This system is built upon the foundation of ForceSPIRE [20], which implemented the

Table 1: StarSPIRE's available semantic interactions and their associated parametric impact on the user interest model. Effects on the term weights ranged from 15% to 40% depending on interaction.

Semantic Interaction	*Effect on User Interest Model*
Open document	Increase weight of terms in the document, and automatically pin.
Minimize document	Reduce weight of terms in the document.
Remove document	Reduce weight of terms in the document; remove document from working set.
Overlap documents (cluster)	Increase weight of terms co-occurring in the overlapped documents.
Highlight text in document	Increase weight of highlighted terms, add terms to model (if not already present).
Annotate document (notes)	Increase weight of terms in the annotation, add terms to model (if not already present).
Search (KSF)	Increase weight of search terms, add terms to model (if not already present), adjust relevance threshold.
Move or un/pin document	Adjust layout model constraints (layout model only; no effect on user interest model).

synthesis portion of the process and provided the weighted, force-directed spatial layout. StarSPIRE adds the foraging portion of the process, enabling data retrieval beyond what is already displayed in the workspace. StarSPIRE also enables a richer set of visual encodings to reflect term weights and document relevance.

3.1 Visual Encodings

Nodes are encoded with node size and saturation to reflect document relevance based on the underlying user interest model (Fig. 1). These encodings are updated during semantic interactions to reflect incremental and constantly evolving user sensemaking. Edges are labeled with the top-weighted terms that co-occur in both documents, and line thickness encodes the total weight of co-occurring terms to reveal how much the documents have in common.

Terms are extracted from documents using LingPipe [9] and are underlined in the documents. Based on the user interest model, StarSPIRE automatically highlights text using a yellow gradient saturation scale to indicate important terms. This allows for quick skimming of documents to determine if they are worth further investigation. User-created highlights are shown in a distinct green color to differentiate from system-generated highlights. Highlighting turns plain text files into visual glyphs that make them easier to locate again on a large, high-resolution display [11, 46].

StarSPIRE also provides visual cues to help users navigate the workspace. Node outline color is used to indicate read or unread status, and node hue is mapped to specific keyword searches (KSF) the user has executed. Each node is labeled with the document title, which can aid in choosing what documents to read as well as locating previously read documents.

3.2 Semantic Interactions

The semantic interactions and their effect on the user model is described in Table 1. These semantic interactions influence the parameters of the user model, either increasing or decreasing the weights of the associated terms. Additionally, terms can be added or removed from the model through these interactions. In order to allow users to change the course of their analysis without being limited by initial paths of investigation, term weights slowly decay over time to slightly emphasize more recent interactions.

The semantic interactions provide feedback to the user interest model and thereby steer the underlying foraging and synthesis models. After each interaction, the system determines which documents continue to meet the relevance threshold based on the updated user interest model. The relevance threshold can also vary depending on the interaction. For example, removing a document raises the relevance threshold temporarily, allowing more irrelevant documents to be pruned from the workspace. Conversely, explicitly executing a search lowers the relevance threshold temporarily to allow more documents to be added to the workspace. Moving nodes and pinning them to fixed locations in the spatialization are the only interactions that operate solely on the layout without updating the user's interest model. Overall, the system was designed to reflect the incremental nature of the human sensemaking process [35], such that semantic interactions have an incremental effect on retrieval and layout.

3.3 Keyword Search and Semantic Interaction Foraging

StarSPIRE allows for two types of foraging: keyword search and semantic interaction. The system explicitly searches for matching documents when the user executes a keyword search. Executing KSF in this manner serves a dual purpose. First, nodes are color coded according to search hits, which can be used to identify relevant documents already on the screen. Second, documents are foraged from the database that are not currently displayed in the workspace.

StarSPIRE uses SIF when users highlight text, write annotations on a document, or overlap documents. SIF first determines which term weights increased in the model as a result of the interaction.

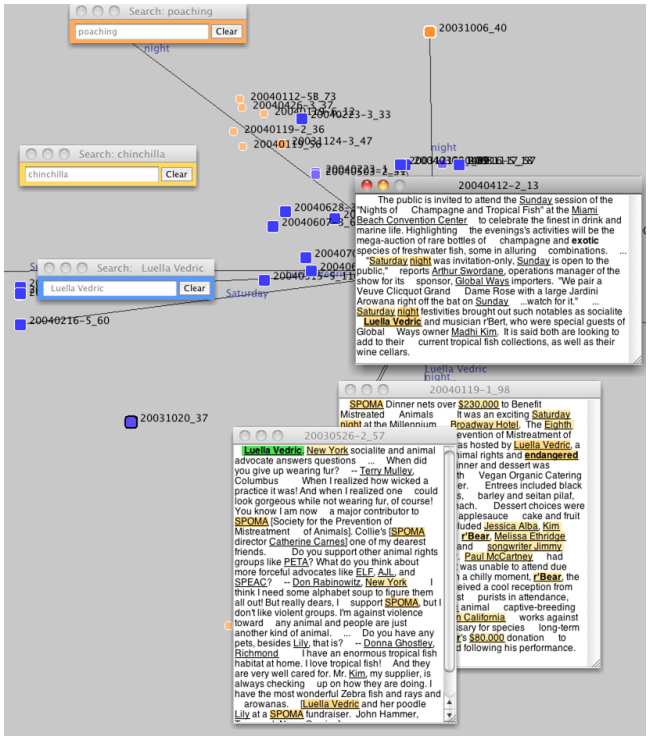

Figure 1: StarSPIRE visual encodings showing document relevance (node size and saturation) and term importance (saturation of automatic yellow highlighting of text).

Next, StarSPIRE uses these terms to search the repository of all documents in the database that are not currently displayed in the workspace. This forms a set of documents that are candidates for addition to the workspace. These documents are then ranked in terms of relevance by matching them to the user's interest model. The top n documents that surpass the relevance threshold are then added into the workspace where they are laid out according to the current display layout model, placing the search results in context of the user's current work. This eliminates the user's need to swap views to execute a query, review results, and add information to the synthesis space. In this manner, synthesis-related actions are leveraged to forage for information, while foraging actions aid in synthesizing information by updating the visual encodings and spatial layout.

For example, when a user overlaps two documents that they think are related, StarSPIRE increases the weight on the terms shared between those two documents in the user interest model, inferring their importance to the user. StarSPIRE then forages for additional documents containing those terms, ranks the documents on relevance to the user interest model, and adds the most relevant to the working set and inserts them into the layout model shown on the screen.

KSF is the traditional method of obtaining potentially relevant documents. Adding SIF functionality enables the system to passively search for information as the analyst is synthesizing documents into their workspace. Because it is based on the user interest model, SIF utilizes many more search terms than are typically contained in KSF queries. This allows for richer matching to find new documents that closely fit the user's perception of what is important, and can help to overcome the difficulties users have in choosing good search terms.

4 Study Design

The goal of this study is to quantify the impact, if any, of introducing contextualized SIF into the sensemaking process. Specifically, how does StarSPIRE with SIF compare to StarSPIRE without SIF? To accomplish this, we conducted a comparative user study with SIF+KSF (referred to as the "SIF+KSF" group) as the test condition and only KSF (referred to as the "KSF" group) as the control condition.

4.1 Task Description

To ensure that users would not be able to simply read all documents in the dataset, and thus would have to forage for a small subset of relevant documents, we chose the large VAST 2007 Challenge dataset named Blue Iguanodon[1] [36]. This dataset presents a law enforcement/counterterrorism scenario composed of multiple latent subplots within the overarching scenario of illegal exotic animal sales. Participants were asked to explore these documents to investigate the scenario. The documents themselves include news articles, blog posts, photographs, hand-drawn comics, and spreadsheets. All of the data, except for the spreadsheets, was used in this study. Because StarSPIRE does not currently contain support for images, all images and comics were transcribed to describe their contents. This resulted in 1486 documents. These documents were processed using LingPipe [9] for entity extraction. After eliminating all entities that only appeared a single time, 1440 entities remained in the term set.

The original Blue Iguanodon dataset does not contain a clear starting point, but to aid the participants, we slightly modified the task description to indicate a starting document for analysis: an article describing an outbreak of a disease called "monkeypox" and implying that chinchillas may be carriers of this disease. Their goal was to identify the cause of this outbreak. The task is suitable for students as well as professionals, requiring no specialized analytical experience or domain knowledge. Also, there is a ground truth for the task: the VAST 2007 Challenge has an associated scoring guide, which enabled us to quantitatively evaluate the quality of analysis.

[1] In addition to the contest summary paper cited above, more information about the 2007 VAST Challenge and the Blue Iguanodon dataset can be found at `http://www.cs.umd.edu/hcil/VASTcontest07/`.

Participants used StarSPIRE on six 30" LCD panels, tiled in a 2x3 grid, on a 24-megapixel display system. This apparatus was chosen to give users ample space to perform spatial synthesis, and avoid the need to close documents purely for lack of space. Large high-resolution displays have been shown to have many benefits for cognitively intensive sensemaking tasks [5, 7, 18].

Participants were given identical training on StarSPIRE with a smaller dataset of 111 short text documents. After a demonstration of the tool's functionality, participants were instructed to solve an analytical task in order to grow comfortable using StarSPIRE's interface. Participants were then given 75 minutes to complete the sensemaking task, requiring participants to explore the 1486 document set to identify the hidden plots regarding illicit activity. The task required participants to sort out and synthesize relevant information from many documents into a coherent hypothesized narrative. All participants used the full allotted time. Although it was unlikely to detect all of the interconnected subplots in this short time frame, a reasonable and uniform time for analysis helped to prevent fatigue and ensure quality analysis. To motivate participants, monetary prizes in addition to the initial compensation were granted to the top three performing participants.

After completing the 75-minute analytical session, participants answered survey questions pertaining to the who, what, and where of the plot, and described their overall hypothesis. All participants had access to their final workspace during the survey to be able to reference their annotations and open documents. Next, the participants drew and annotated their spatial organizational schema on paper. Finally, users completed a survey to give feedback on their analytical strategy, difficulties encountered, and how StarSPIRE helped or hindered their analysis. The proctor conducted a brief semi-structured interview for any remaining comments. Also, participants were able to pause and ask questions at any point during the sensemaking session. The entire session, from informed consent to final survey and interview, spanned approximately two hours.

We collected logs of all interactions performed by users as well as snapshots of the underlying model parameter values, took screenshots every minute, and saved their final workspaces so that they could be loaded and examined at a later date.

4.2 Participants

We recruited 18 graduate and undergraduate students from varying academic backgrounds. Participant ages ranged from 18 to 42 ($\mu = 23$, $\sigma = 5.6$). Twelve participants were male and six were female. Twelve were computer science students, five from engineering disciplines, and one from mathematics. Six participants were graduate students, and twelve were undergraduates. Each participant was randomly assigned a condition (KSF or SIF+KSF, described in the next subsection) such that each condition had an equal number of participants.

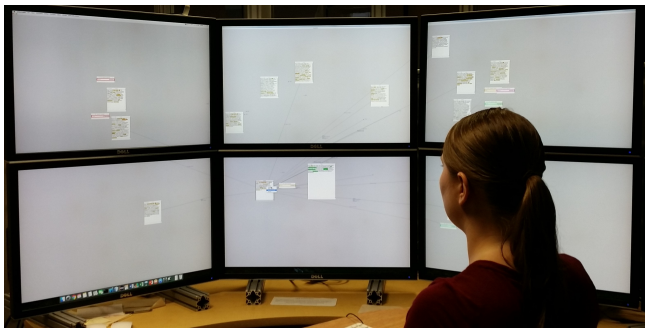

Figure 2: A participant interacting with StarSPIRE.

4.3 Study Conditions

This study consisted of two conditions. The test condition is referred to as the "SIF+KSF" group, in which participants had access to the full StarSPIRE system. Participants assigned to this group could use both semantic interaction foraging and keyword search foraging when exploring the document collection. In other words, StarSPIRE foraged for new documents to recommend to each participant based on their explicit keyword searches, as well as by their interactions in opening, minimizing, removing, overlapping, highlighting, and annotating documents. The semantic interactions provided to these participants are listed in Table 1.

The control condition is referred to as the "KSF" group. Participants assigned to this group could only forage for new documents via explicit keyword searches that they typed into search boxes. Participants still had the ability to perform the semantic interactions listed in Table 1 that updated the user model, but automatic foraging did not occur as a result of those actions. For example, participants could still highlight phrases within the documents to support their own synthesis process and to support the layout and automatic highlighting, but StarSPIRE did not automatically forage for documents related to those phrases or the updated model. The StarSPIRE system was identical in both conditions, except that the SIF functionality was turned off in the KSF condition. Participants were unaware of the different conditions for the study, and no change to the user interface was evident to the KSF participants.

5 Study Results

Using a combination of log files, screenshots, solution sheets, surveys, and interviews, we quantitatively and qualitatively evaluate how SIF impacted the sensemaking process. Specifically, we examine (1) how well users performed, (2) how well they foraged for relevant documents, (3) which relevant documents they discovered and how they found them, (4) what interactions they performed, and (5) what strategies they applied.

Each of the following subsections begins with a summary of the research question addressed, followed by the study results and a discussion of their significance. We report both significant and non-significant results, showing both conclusions drawn from this study as well as directions for further investigation.

5.1 SIF+KSF Participants Averaged Higher Scores

In this subsection, we investigate how the introduction of SIF affected the participant scores resulting from their exploration of the Blue Iguanodon document collection. We found that SIF+KSF group members exhibited significantly higher average scores.

5.1.1 Results

Using the published scoring rules from the VAST 2007 Challenge [36], we computed a performance score for each participant. Participant scores ranged from 1 to 17. The maximum possible score was 58, although we did not expect participants to approach this value given the time constraints of this study. No participants identified any subplots outside of the plot indicated in the starting document. The highest possible score considering only the initial plot was 27. The scores were higher in the SIF+KSF group than in the KSF group (SIF+KSF: $\mu = 8.0$, $\sigma = 5.4$, $min = 3$, $max = 17$; KSF: $\mu = 4.2$, $\sigma = 3.3$, $min = 1$, $max = 10$). The individual scores with their means are shown in Fig. 3.

Due to the small sample size ($n = 9$ for each group), we first performed two Shapiro-Wilk tests [42] for normality, to learn whether or not the participant scores in each group were normally distributed. The non-significant outcomes of this test at the $\alpha = 0.05$ level ($W = 0.064$ for the SIF+KSF group, $W = 0.229$ for the KSF group) indicated that the scores were approximately normally distributed.

Following this, we performed a t-test assuming unequal variance, using the alternative hypothesis that the SIF+KSF scores would be higher than the KSF scores. At a significance level of $\alpha = 0.05$, we found that the SIF+KSF group scores were significantly higher than the KSF group scores ($t = 1.8045$, $df = 13$, $p = 0.0471$). This process of non-significant Shapiro-Wilk test preceding an unequal variance t-test was used for all other inferential statistics presented in the following subsections.

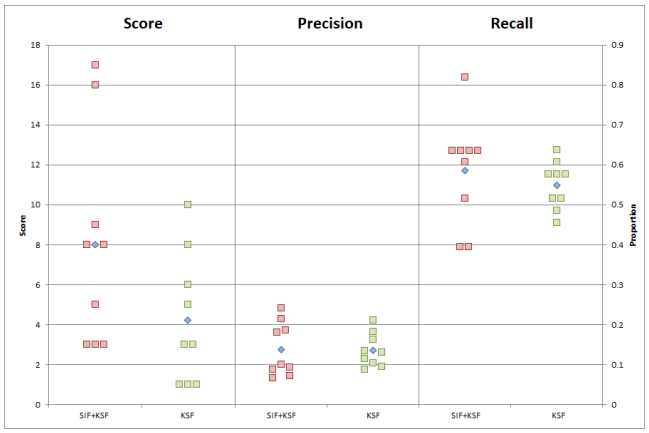

Figure 3: Score (left y-axis scale), precision, and recall (right y-axis scale) of foraging performance by all participants. Mean group scores are shown as blue diamonds. We found a statistically significant difference between conditions in score, but not in precision and recall.

5.1.2 Discussion

The results from this section show that, on average, participants in the SIF+KSF group understood the plot to a greater degree than those in the KSF group. Though the p-value is near the $\alpha = 0.05$ significance threshold, this is due in part to the small sample size of 9 participants in each group. The mean score for SIF+KSF participants was nearly double that of the KSF participants. However, we also note that the inclusion of SIF produced a higher variance in scores than participants who were only afforded KSF. We suspect that this is due in part to the variable number of semantic interactions performed by SIF+KSF participants – both the choice and frequency of semantic interactions used influences the set of documents that are foraged, and thereby influences how well the participant understands the plot. We discuss further explanations for the effect of the inclusion of SIF on documents foraged in the next two subsections.

5.2 No Change to Precision and Recall between Groups

In this subsection, we investigate how the introduction of SIF affected precision and recall scores for foraging performance. We found no significant difference in foraging precision and recall between the SIF+KSF and KSF groups.

5.2.1 Results

In evaluating the foraging performance of participants, we compute precision, recall, and F-measure values for the relevant documents found by each participant. These results are summarized in Table 2. We compute *precision* to be the number of relevant documents found divided by the total number of documents retrieved and *recall* as the number of relevant documents found divided by the number of relevant documents in the known solution. *F-measure* is computed as $2 * precision * recall / (precision + recall)$. In this scenario, there were 33 documents relevant to the known solution. We used the participant log files to identify which documents were retrieved into the workspace in order to calculate precision, recall, and F-measure (shown in Fig. 3).

The SIF+KSF group averaged a precision score of 0.14 ($\sigma =$ 0.07), a recall score of 0.59 ($\sigma = 0.13$), and an F-measure of 0.21

Table 2: Scores, counts of documents retrieved, and precision-recall statistics for each condition.

	SIF+KSF Avg.	KSF Avg.	All Avg.
Score (out of 27)	8.0	4.6	6.1
Unique Relevant Docs Retrieved (out of 33)	19.3	18.1	18.7
Total Unique Docs Retrieved (out of 1486)	178.0	145.2	161.6
Precision	0.14	0.14	0.14
Recall	0.59	0.55	0.57
F-Measure	0.21	0.21	0.21

($\sigma = 0.09$). Similarly, the KSF group averaged a precision score of 0.14 ($\sigma = 0.04$), a recall score of 0.55 ($\sigma = 0.06$), and an F-measure of 0.21 ($\sigma = 0.05$). It is noteworthy that both groups had very similar precision, recall, and F-measure scores. This result is counter to our initial hypothesis, which was that SIF would increase recall but might penalize precision.

We did not observe a significant difference between SIF+KSF and KSF conditions in the total number of unique documents retrieved ($t = 0.8681$, $df = 12$, $p = 0.4024$). Across conditions, the number of unique documents retrieved ranged from 70 to 315 ($\mu = 162$, $\sigma = 80$), which corresponds to 4.7% to 21.2% of the entire dataset retrieved. The SIF+KSF participants retrieved between 70 and 315 unique documents ($\mu = 178$, $\sigma = 102$), and the KSF condition participants retrieved between 90 and 239 documents ($\mu = 145$, $\sigma = 50$). Although these documents were imported into the workspace, not all of them were read. This in and of itself is a promising result. Participants were able to mentally filter out many of the irrelevant documents in their synthesis phase.

5.2.2 Discussion

Because we are evaluating the influence of semantic interactions on foraging, our computations of precision and recall used the number of documents (and relevant documents) retrieved, rather than using a similar measure such as number of documents opened or interacted with. This choice allows us to measure what the system is giving the analysts to read, rather than exploring what the analysts are focusing on. It is certainly possible that altering these computations could affect our non-significant results.

Overall, it is interesting to note that the foraging results for the SIF+KSF group consistently show a standard deviation twice that of the KSF group. This is further evidence that SIF introduces greater variability into the foraging process. The number of documents retrieved from the dataset varied based on user analytical strategy.

5.3 SIF and KSF Serve Complementary Document Foraging Roles

In this subsection, we investigate how the introduction of SIF affects the set of relevant documents retrieved and how they were retrieved. We found that KSF and SIF each have their own advantages towards retrieving certain sets of documents, and that highlighting was the primary semantic interaction used to retrieve documents.

Table 3: Quantity and percentage of relevant documents retrieved using the various interaction methods for the two conditions.

	SIF+KSF	KSF
Relevant Docs Retrieved (includes re-finds)	22.4	20.1
Total from SIF	5.8 (26%)	
SIF from Highlight	5.0 (22%)	
SIF from Annotate	0.2 (1%)	
SIF from Overlap	0.6 (3%)	
KSF from Search	16.7 (74%)	20.1 (100%)

5.3.1 Results

Document discovery results are summarized in Fig. 4. Of the relevant documents in the collection, the "chinsurrection" documents were almost universally found by every KSF participant (one KSF participant missed one of the documents). In contrast, some SIF+KSF participants missed them. These documents are central to the main plot of the investigation, which most of the participants at least partially solved. All of these documents contain the name of the central nefarious character, "Cesar Gil," that most of the users cited in their solutions. All KSF group users explicitly searched on his name, but three of the SIF+KSF users did not, and consequently some of those three missed a subset of these documents.

In contrast, the other relevant documents were found more often by the SIF+KSF participants. In particular, three of these documents were not found by any of KSF users, yet were found by 3/9 of the SIF+KSF users. One of these documents contained supporting evidence for the main plot described above, but did not identify Cesar Gil by name. The other two documents contained information relevant to a second subplot that interconnects with the main plot, about another character named "rBear," although none of the participants succeeded in solving this plot. This character's name was never explicitly searched for by any of the participants, so it is likely this information was retrieved through SIF, perhaps exploiting other keywords in common between the two plots, such as "monkeypox" (a highly weighted term in the final states of many of the participants' user interest models). This indicates that it was valuable to have the SIF mechanism to expand the scope of investigation to this other relevant but less obviously connected information, beyond keywords on which users might not think to explicitly search.

The SIF+KSF group located some relevant documents through their semantic interaction foraging ability, while the KSF group used only the keyword search means. The percentages are shown in Table 3. We examined the interaction logs of the participants in the SIF+KSF condition to determine if they retrieved relevant documents via semantic interactions that executed SIF retrieval. Eight out of nine SIF+KSF users retrieved new relevant documents using SIF. Including re-finds (relevant documents that were located, removed from the working set by the user, and then located again), SIF accounted for 26% of the total number of relevant documents retrieved by the SIF+KSF group. For individual SIF+KSF users, this percentage ranged from 0% to 100%, demonstrating the wide variety of user strategies. This also suggests that users succeeded in finding useful information via SIF, information that might not have been found through explicit KSF. By far, most of the SIF-retrieved relevant documents were retrieved as a result of highlight interactions, indicating the importance of this type of semantic interaction.

5.3.2 Discussion

From this analysis of foraging behavior, we can see benefits of both SIF and KSF. KSF is useful when specific terms of interest are known; a keyword search for "Cesar Gil" added many of the "chinsurrection" documents into the working set, indicating that KSF is still valuable for foraging, especially for terms that are more obvious targets of investigation. Simultaneously, KSF is limited when those precise search terms are not present in other relevant documents. SIF, in contrast, can locate documents related to the current direction of exploration without the analyst knowing precisely what to search for, but with the limitation that SIF may not locate *all* of the documents that an analyst may be seeking. This limitation can be addressed by more accurate learning and retrieval models in the future.

Interestingly, the SIF+KSF group earned higher analysis scores on average, despite not finding all of the core "chinsurrection" documents. Instead, they earned higher scores by building up a more complete plot with the supplemental documents they found through SIF. The sensemaking process is boosted by SIF locating this broader supplemental information, beyond the obvious core documents.

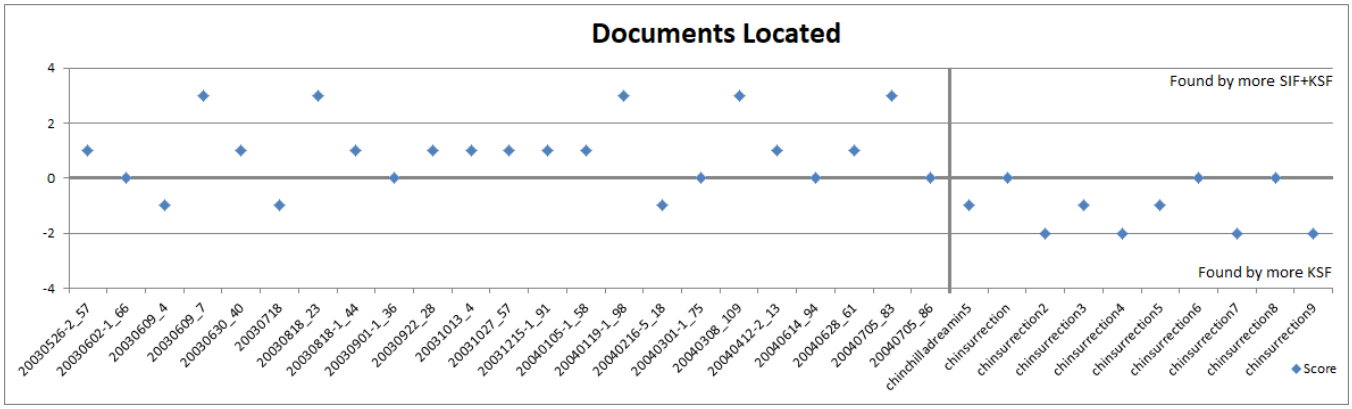

Figure 4: The difference between the number of SIF+KSF group participants and KSF group participants who found each relevant document. Positive scores (above the horizontal axis) mean that more SIF+KSF participants found the document, while negative scores (below the horizontal axis) mean that more KSF participants found the document. More SIF+KSF participants found a majority of the documents, but more KSF participants found the core "chinsurrection" documents.

5.4 SIF+KSF Participants Performed More Synthesis Interactions

In this subsection, we investigate how the introduction of SIF affected the number of semantic interactions performed by participants. We observed significant differences between study conditions in terms of how much information users externalized to the workspace via some synthesis-related actions, which may have contributed to the potential trend of improved performance by the SIF+KSF participants compared to the KSF participants.

5.4.1 Results

In order to track how users synthesized information, we once again analyzed the interaction logs (Fig. 5). We identified the following semantic interactions as being directly related to synthesis through the externalization of the user's thought processes: highlighting, annotating, and document overlapping (clustering). The SIF+KSF condition participants performed significantly more highlights ($t = 2.3227$, $df = 16$, $p = 0.0169$) and significantly more annotations ($t = 2.0809$, $df = 9$, $p = 0.0336$). There was no significant difference between the number of times that users clustered documents by overlapping them (SIF+KSF $\mu = 15.6$; KSF $\mu = 11.2$), nor was there a significant difference in the number of keyword searches performed by each group (SIF+KSF $\mu = 19.2$; KSF $\mu = 18.0$). Users varied in their preferences for performing various interactions (e.g., some preferred annotating over highlighting, others preferred overlapping documents).

Figure 5: The panels from left to right show the total number of highlight, annotation, document overlap, and search interactions performed by each participant. Means are shown as blue diamonds. The highlight and annotation conditions are significantly different, with the SIF+KSF group performing more actions than the KSF group in both. The overlap and search conditions are not significantly different.

5.4.2 Discussion

We can infer from these results that the SIF+KSF users externalized more of their understanding of the dataset and hypotheses about what information was relevant. Overall, these participants provided more feedback to the user model regarding their interests. This feedback was not only used to retrieve documents, but also to augment the spatialization in terms of document positioning, visual encodings, and automatic text highlighting. This process serves to continually give analysts visual feedback on what documents it believes will be most relevant or interesting for the analyst to read. Therefore, the system is more likely to indicate good documents on the display for the user to open and read next based on their interests. Both study conditions were provided with this relevance feedback based on their underlying interest model, although the SIF+KSF participants benefited from this feature more than KSF participants.

Furthermore, user's highlighting and annotating documents aids in auto-highlighting of the text in open documents, making them easier to skim. It also helps transform open documents into distinguishable visual glyphs that aids in re-finding information, making analysts more efficient in navigating the workspace and referencing the workspace for filling out their final solution reports [11, 46].

The significant difference between study conditions may have been a result of a positive-reinforcing feedback loop. As users made highlights in documents or wrote notes, the system retrieved and identified documents that it believed the users would be interested in. This may have encouraged the users in the SIF+KSF condition to continue performing these actions. Thus, synthesis-related actions foraged for information, both on and off the screen, which led to more data being interpreted and formulated into hypotheses. It is interesting though that this did not seem to significantly reduce their use of search. This might suggest a possible design opportunity for more clear visual connection between KSF and SIF.

5.5 Participants Exhibited a Variety of Strategies

In this subsection, we investigate the structure and layout of the final workspaces for both groups of participants. Overall, participant strategies for use of the workspace mirrored previous results about sensemaking with large display spaces [5, 7]. Users organized a variety of spatial representations of the document collection as part of their distributed cognitive process.

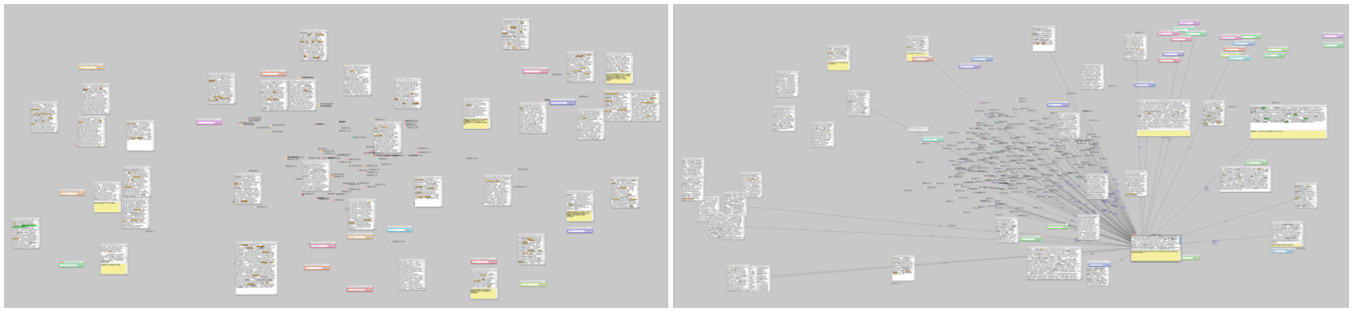

Figure 6: Final workspace of **(left)** user KSF #9 showing the spatial organization of documents, annotations, and search boxes that label the space, and **(right)** user SIF+KSF #4 showing a large central pool of unopened documents, with opened documents arranged on the periphery.

5.5.1 Results

The final screenshots of user workspaces shared a common artifact, likely caused by a low relevance threshold that kept a high number of documents on the display. Most participants' final workspaces contained a central pool of unopened document nodes with documents arranged around the periphery of the display. The nodes in the central pool represented weakly relevant information. Nodes that were highly relevant to a specific cluster of documents were positioned near the cluster.

Users were asked to sketch spatial representations of their final workspaces – how they perceived the space. Users adopted different methods for labeling the space, even within their own drawings, which would make automatic cluster detection and classification difficult [21]. For example, users created specific tags for areas of the display that directly matched extracted entities (e.g., monkeypox, Cesar Gil), but they also tagged areas of the space with cognitively meaningful labels (e.g., who, what, where). This behavior has been previously observed, where users label their spatial workspaces in fuzzy and complicated manners that would be difficult to match by another person or algorithm [13].

The number of open documents on the final workspaces varied greatly, from 2 to 35 ($\mu = 15.56$, $\sigma = 10.51$). There was also substantial variance within each condition. The SIF+KSF condition participants kept a range of 2 to 34 documents open on their final workspace ($\mu = 14.78$, $\sigma = 11.69$). The KSF condition participants ranged in keeping 3 to 35 documents open ($\mu = 16.33$, $\sigma = 9.84$). The lack of significant difference implies that any trend in performance between the conditions cannot be explained by the number of open documents alone.

Interestingly, participants who had very few documents open on their final workspace still drew spatial schemas indicating where they had opened and then minimized or closed documents. For example, participant SIF+KSF #1 opened documents 57 times, but only had two documents open on his final workspace, neither of which were relevant to the solution. In fact, participant SIF+KSF #1 did not have any relevant documents on his final workspace, open or closed. He preferred a neat and clutter-free workspace and deleted documents after he had read and processed the information. For reference, this participant retrieved 20 relevant documents and had the second highest score. The highest scoring participant overall, SIF+KSF #8, assumed quite the opposite strategy. She retrieved a total of 310 documents, 28 of which were relevant. She also opened documents 57 times, yet she kept 16 open on her final workspace, 11 of which were relevant to the overall solution.

In the KSF condition, participants KSF #1 and KSF #2 had the highest scores in their group. They also adopted differing strategies in terms of keeping documents on the display. KSF #2 retrieved 117 documents, 21 of which were relevant. She had 15 documents open on her final workspace, 10 of which were relevant. KSF #1 retrieved 118 documents, 15 of which were relevant. He had 6 documents open on his final workspace, but none were relevant. However, three documents were opened and then minimized, indicating that they were read.

5.5.2 Discussion

We found no correlation between any of these organizational strategy metrics and user performance. This can be attributed to individual differences in analytical strategies, such as user ability, the desire to keep a neat workspace (or not minding having the display filled with open and closed documents), or needing to focus on one or two documents at a time so as not to get distracted. These preferences were explained during the post-study surveys and semi-structured interviews. We see that StarSPIRE supports a variety of analytical strategies and user preferences without a particular strategy having an adverse impact on sensemaking quality and performance. Our results also replicate previous work that demonstrates how users remember spatial locations of items on a large, physical workspace, both during data analysis and after the fact when the display is empty [5, 32].

6 Discussion

We begin this section by summarizing the lessons learned from this study, and discussing ways by which these lessons can be applied beyond StarSPIRE and into visual analytics in general (Section 6.1). Following this, we discuss two issues that surfaced through our observations of participants and analysis of the study data. The first was a sometimes overwhelming number of documents staying on the display, which suggests a need to modify the relevance threshold (Section 6.2). The second is the problem of cognitive tunneling, which we noticed when no participants identified additional subplots in the data (Section 6.3). We discuss these issues and suggest methods for alleviating the problems in future work. We also discuss a feature that proved to be surprisingly important to the users, the automatic text highlighting (Section 6.4). In addition, we discuss the potential for tuning semantic interactions to individual users (Section 6.5), and the limitations of our study (Section 6.6).

6.1 General Principles

Throughout the experimental results detailed in the previous section, we saw that incorporating a "passive" foraging mechanism like SIF that retrieves documents based on a learned user model can recommend documents to analysts that they may not have found via traditional keyword search means. The result of these document recommendations is that analysts gain a better understanding of the underlying plot in the document collection (evidenced by their higher scores), and interact with the workspace more, leading to a feedback cycle that continues to improve document recommendations with each additional interaction. At the same time, the quality of the documents recommended (measured by the foraging precision and recall scores) is not reduced.

KSF and SIF represent independent, complementary mechanisms for information retrieval, each with strengths and weaknesses. As such, SIF should not be used by system designers as a replacement for keyword search. Indeed, the inclusion of a search box can greatly benefit the usability of a visualization system as datasets increase in size [1]. Rather, our findings suggest that implicit or passive search mechanisms can be included in visualization systems to draw the user's attention to related objects that may not necessarily include identical search terms.

Similarly, our findings suggest that interactions alone are sufficient to drive these search mechanisms, building a user model by interpreting the interest of a user based on how they interact with other documents. We noted previously (Section 2.4) that our learning rules to generate a user model are relatively simple, and that more thorough recommendation systems that follow an interaction-driven approach could certainly outperform our findings. However, our results show that even this simple approach can cause substantial improvement. We additionally assert that existing visualization systems could make use of our simple approach of mapping interactions to weight modifications, ultimately to the benefit of a user.

For example, Andromeda [41] allows users to manipulate projections to learn a set of attribute weights that will best approximate the user-provided projection. To use this mechanism, a user uses drag-and-drop interactions to manipulate the current projection, dragging observations to various positions in the workspace to communicate desired similarity/dissimilarity relationships. The new weights are not learned until the user has finished their repositioning interactions and click an "Update Layout" button. With a StarSPIRE-like approach, the system can begin to learn the user's desired similarity/dissimilarity goals while the user is still performing these interactions. As the user moves more observations, Andromeda could begin to recommend additional observations to reposition, and could even recommend the positions to place the additional observations. Such suggestions could lead to better reprojection results as the feedback from user to system is increased.

Similarly, Intent Radar [39] allows users to manipulate a projection of keywords centered about a "radar" display. Users can reposition keywords closer to the center of the radar to indicate that a particular keyword is more important to their interests. Using the StarSPIRE approach, the system could learn from this sequence of interactions, perhaps discovering documents that contain these keywords and recommending other keywords within those documents.

6.2 Relevance Threshold

In this study, it appears that the relevance threshold may have been set too low, causing too much irrelevant information to remain on the display. This was observed during the pilot study, but we chose to maintain this relevance threshold level so that the system would not over-prune the workspace, which can be more problematic. As a result, many users ended up with a central "pool" of data and arranged their open documents on the perimeter.

Participants retrieved a widely varying number of documents ($\mu = 161.61$, $\sigma = 79.52$). The number of documents removed also varied greatly ($\mu = 31.50$, $\sigma = 20.34$). This can be attributed to differences in analytical strategies by the participants. During the post-study interviews, it was revealed that some users (e.g., KSF #1) preferred to keep a clutter-free workspace and keep as few documents open as possible. Others (e.g., SIF+KSF #8) did not feel overwhelmed by the excess information and preferred having a great deal of information to pull from. These two participants earned the highest scores in their groups. The data gathered in this study show that the variation in clutter represented by the participants' layouts did not correlate with their performance; however, it is reasonable to assume that an excessive amount of clutter would impact task performance.

In order to support these varied styles, it may be prudent to alter the document relevance threshold to adapt to each user instead of having static values based on interactions. The model could incrementally learn from the interactions users perform and update as needed. For example, if a user has a tendency to delete documents from the workspace, the threshold for keeping documents should be raised so that more are automatically pruned from the display. An alternative to this strategy may be to begin with a more strict threshold to only show closely-related documents. Then, once a foraging saturation is reached, the threshold could be lowered incrementally to bring in new documents. Likewise, if the system is able to detect a large number of documents that are just under the current relevance threshold, say related to a new subplot that has just been encountered, the relevance threshold could be lowered to bring in all of those new documents.

6.3 Cognitive Tunneling

While the Blue Iguanodon dataset contained multiple subplots, no users branched out to identify any other plot aside from the main plot mentioned in the starting document. Some participants pursued alternative hypotheses for this subplot, but none correctly identified adjacent subplots. Interestingly, a few participants read documents containing information on different subplots, and one even executed searches for relevant entities involved in a second subplot. However, they did not include this information in their solution. Our instructions to the study participants did not indicate that there was only a single plot within the dataset; we merely provided them with the starting document and allowed them to begin exploring. Participants may have implicitly assumed that they should focus on the specific plot hinted at in the starting document, ignoring other interesting threads that they uncovered in the data. This indicates that many of the participants in both study conditions fell victim to a phenomenon similar to cognitive tunneling [33] or satisfaction of search [45], in which an analyst narrows their attention to target an initial discovery, ignoring other possibilities.

One explanation for this issue is due to how information was retrieved and synthesized by participants. Documents added to the workspace were those containing terms that most closely matched the user's model for both study conditions. This could have led to confirmation bias and a tendency to ignore alternatives in the plot. That said, confirmation bias is a feature of participants, not retrieval systems. For example, an analyst investigating the question "Do chinchillas have monkeypox?" could initiate a search that returns information that both confirms and refutes the question. The decision to pursue one conclusion or the other occurs during the synthesis process that follows the search. Indeed, results from Section 5.3 show that introducing SIF may work to alleviate the effects of confirmation bias, because SIF returns related documents that may not be found through traditional keyword search. Our results show that SIF presented participants with a broader set of documents, many with more subtle ties to the currently investigated hypothesis.

This cognitive tunneling effect can also be attributed to each user being provided with an explicit starting point in the analysis. While this was intended to focus the investigation of the study participants during a time-limited task and reducing the variation between users, it also has the effect of limiting open-ended investigation within the document set. The automatic and dynamic highlighting may further influence this effect by "steering" the participants towards searching for highlighted terms and missing potentially useful documents that are not significantly highlighted. Studies from the visual search community have noted that prevalence [50] and detectability [49] play significant roles in target location.

One way to alleviate this problem is to introduce novel documents to the workspace in addition to highly relevant documents. Ruotsalo et al. achieved this by sampling from a distribution of documents according to their relevance [39]. This allowed for closely related documents to be shown as results, but also occasionally to show novel documents. Again echoing the idea of detectability for visual

search, it is advantageous to visually indicate novel documents within the spatialization to draw user focus to them. We noted that participants tended to open large, bright documents, even if they were in a cluster of many documents. How to integrate the notion of veering away from the user's model to highlight novel information within the current multi-model semantic interaction pipeline remains an open research challenge.

Another possibility is to provide large-scale overview spatializations of the full document collection. ForceSPIRE accomplished this by simply displaying the entire document collection at start-up, but therefore only worked for small document collections [20]. StarSPIRE abandoned that approach in order to handle larger document collections, instead focusing on retrieval. This leads to opportunities to integrate other types of overviews of large document collections, such as sampling, clustering, and topic modeling [17].

6.4 Automatic Text Highlighting

According to user comments, the user-tuned automatic text highlighting (shown in the interface in Fig. 1) proved to be one of most valuable features in StarSPIRE. This feature gave the users a subtle yet salient visual representation of the underlying interest model. The highlights had the potential to change with each interaction, thus continuously representing the current underlying model's state, reflecting which terms the user interest model had placed the most emphasis on. This visual feedback about the state of the model was conveniently presented, in context, directly within the documents and served as a form of "explainable AI".

However, automatic highlighting proved to be much more useful than merely giving feedback. Participants leveraged the automatic highlighting to (1) determine which documents in a collection are worth reading, and (2) determine which portions of a document to focus on, particularly in longer documents. Thus, StarSPIRE gave users feedback at multiple levels of data abstraction through a visual metaphor that users found easy to interpret. At the graph level (many documents), the node relevance size-encoding and text highlighting served to guide users toward relevant documents to read. They could quickly identify pertinent documents with a quick glance at the highlighted terms. At the individual document level, text highlighting directed users to portions of the document to read. This was particularly useful to home in on specific paragraphs in long documents, and to identify where multiple reports contain similar information that the user has already read before, allowing them to skip over that content. Further, no users complained about the system recommending improper documents. The incremental dynamics of the working set, layout, and visual encodings did not appear to frustrate them. These features appeared to help direct the user's attention, but further research is necessary to measure any increased analyst efficiency produced by auto-highlighting.

6.5 Tuning Semantic Interactions

Participants employed vastly different strategies in conducting their analysis. For example, foraging performance was similar across the participants, but how they went about foraging varied greatly. It is not unreasonable to assume that users had different preferences in terms of what interactions they performed. Currently, interactions are interpreted the same way for all users. We may be able to tune the impact of the interactions to better approximate the preferences of each individual user, perhaps using machine learning methods to tune the parameters of the semantic interactions.

For example, if a user repeatedly closes documents that were retrieved as a result of highlighting sentences in a document, the system could reduce the impact that highlighting has on extracted entities. In this manner, we can attempt to avoid over-assuming the intent of users. Instead, we can begin with a baseline interpretation of interactions and incrementally tune these interpretations through a meta-level semantic interaction learning process.

6.6 Limitations

Three noteworthy limitations to this study include the small number of participants, the single document set used, and the short task time duration. First, it is possible that we could have discovered stronger significance levels and more significant results with additional participants, reducing some of the variance in the participant scores and behaviors. Additionally, using only the Blue Iguanodon set of documents in this study limits the generalizability of our findings. This is due to the large amount of noise in the document set (since many documents are irrelevant to the main plot), as well as the short average length of these documents. It is certainly possible that these findings could differ with a collection of longer documents. Finally, while supporting longer-term sensemaking sessions is certainly a goal of this work, the current study focuses on short two hour sessions. It would be interesting to see the value of SIF functionality in multi-day scenarios.

An additional consideration in the study design is that, while the KSF group did not have the advantage of SIF, they did have other advantages associated with semantic interaction and the user interest model, such as automatic highlighting, relevance-based node sizing, etc. Most KSF-only systems would likely not include these features. Thus, the actual difference between the experimental condition and the control condition in this study was perhaps smaller than it would be in a realistic setting. The results found in this study might actually be amplified when comparing SIF to traditional KSF approaches.

7 Conclusions and Future Work

We conducted a comparative user study with StarSPIRE to examine the impact of SIF on the sensemaking process. The study showed that foraging performance was similar between conditions. However, the group afforded with SIF functionality performed significantly more synthesis-related semantic interactions. They externalized more information (a process associated with synthesis) and injected more feedback into the underlying user interest model. The system was then able to forage and identify a broader set of relevant information in the spatial workspace. This led to improved sensemaking task performance, for foraging large textual information and synthesizing a coherent and complete hypothesis narrative, as scored against a known ground-truth solution. Participants in the SIF+KSF condition retrieved 26% of their total relevant documents through SIF on average. Executing traditional keyword searches retrieved the remaining 74% of relevant documents. SIF and KSF proved to be effective complementary retrieval techniques.

Based on this user study, participants were able to solve a portion of the given sensemaking task while retrieving and reading only a small portion of the overall dataset. However, there was not a clear superior analytical strategy, which demonstrates that StarSPIRE supports multiple avenues for sensemaking. We also identified potential improvements for StarSPIRE and its underlying layout and relevance models that inform the design of future semantic interaction systems. The automatic text highlighting and node size relevance encoding, both features enabled by semantic interactions that learn a user interest model, were especially appreciated by users.

We intend to implement the identified changes, including altering the relevance model to include novel documents in addition to the documents that most closely match the current user interest model. After these proposed changes are made, we plan to conduct a longitudinal study to observe long-term usage of StarSPIRE on real-world data. Example tasks include conducting an in-depth literature review and learning about a current event in the news.

Acknowledgments

This research was supported by NSF Grants IIS-1218346 and IIS-1447416. The authors would like to recognize the role of comments from reviewers and discussions with InfoVis Lab @ VT research group members in improving this work.

REFERENCES

[1] J. Abello, F. V. Ham, and N. Krishnan. Ask-graphview: A large scale graph visualization system. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):669–676, Sept 2006. doi: 10.1109/TVCG.2006.120

[2] J.-w. Ahn and P. Brusilovsky. Adaptive visualization for exploratory information retrieval. *Information Processing & Management*, 49(5):1139–1164, 2013. doi: 10.1016/j.ipm.2013.01.007

[3] J.-w. Ahn, P. Brusilovsky, D. He, J. Grady, and Q. Li. Personalized web exploration with task models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pp. 1–10. ACM, New York, NY, USA, 2008. doi: 10.1145/1367497.1367499

[4] J. Alsakran, Y. Chen, Y. Zhao, J. Yang, and D. Luo. Streamit: Dynamic visualization and interactive exploration of text streams. In *2011 IEEE Pacific Visualization Symposium*, pp. 131–138, March 2011. doi: 10.1109/PACIFICVIS.2011.5742382

[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*, CHI '10, pp. 55–64. ACM, New York, NY, USA, 2010. doi: 10.1145/1753326.1753336

[6] C. Andrews and C. North. Analyst's workspace: An embodied sensemaking environment for large, high-resolution displays. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 123–131, Oct 2012. doi: 10.1109/VAST.2012.6400559

[7] C. Andrews and C. North. The impact of physical navigation on spatial organization for sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2207–2216, Dec 2013. doi: 10.1109/TVCG.2013.205

[8] M. Q. W. Baldonado and T. Winograd. Sensemaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI '97, pp. 11–18. ACM, New York, NY, USA, 1997. doi: 10.1145/258549.258563

[9] B. Baldwin and B. Carpenter. Lingpipe. *Available from World Wide Web: http://alias-i.com/lingpipe*, 2003.

[10] E. A. Bier, S. K. Card, and J. W. Bodnar. Entity-based collaboration tools for intelligence analysis. In *2008 IEEE Symposium on Visual Analytics Science and Technology*, pp. 99–106, Oct 2008. doi: 10.1109/VAST.2008.4677362

[11] L. Bradel, A. Endert, K. Koch, C. Andrews, and C. North. Large high resolution displays for co-located collaborative sensemaking: Display usage and territoriality. *International Journal of Human-Computer Studies*, 71(11):1078–1088, 2013. doi: 10.1016/j.ijhcs.2013.07.004

[12] L. Bradel, C. North, L. House, and S. Leman. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 163–172, Oct 2014. doi: 10.1109/VAST.2014.7042492

[13] L. Bradel, J. Z. Self, A. Endert, M. S. Hossain, C. North, and N. Ramakrishnan. How analysts cognitively "connect the dots". In *2013 IEEE International Conference on Intelligence and Security Informatics*, pp. 24–26, June 2013. doi: 10.1109/ISI.2013.6578780

[14] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 83–92, Oct 2012. doi: 10.1109/VAST.2012.6400486

[15] P. Brusilovski, A. Kobsa, and W. Nejdl. *The adaptive web: methods and strategies of web personalization*. Springer Science & Business Media, 2007.

[16] G. Chin, Jr., O. A. Kuchar, and K. E. Wolf. Exploring the analytical processes of intelligence analysts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 11–20. ACM, New York, NY, USA, 2009. doi: 10.1145/1518701.1518704

[17] A. Endert, L. Bradel, and C. North. Beyond control panels: Direct manipulation for visual analytics. *IEEE Computer Graphics and Applications*, 33(4):6–13, July 2013. doi: 10.1109/MCG.2013.53

[18] A. Endert, L. Bradel, J. Zeitz, C. Andrews, and C. North. Designing large high-resolution display workspaces. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pp. 58–65. ACM, New York, NY, USA, 2012. doi: 10.1145/2254556.2254570

[19] A. Endert, R. Burtner, N. Cramer, R. Perko, S. Hampton, and K. Cook. Typograph: Multiscale spatial exploration of text documents. In *2013 IEEE International Conference on Big Data*, pp. 17–24, Oct 2013. doi: 10.1109/BigData.2013.6691709

[20] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, Dec 2012. doi: 10.1109/TVCG.2012.260

[21] 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*, AVI '12, pp. 555–562. ACM, New York, NY, USA, 2012. doi: 10.1145/2254556.2254660

[22] A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. Observation-level interaction with statistical models for visual analytics. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 121–130, Oct 2011. doi: 10.1109/VAST.2011.6102449

[23] P. Fiaux, M. Sun, L. Bradel, C. North, N. Ramakrishnan, and A. Endert. Bixplorer: Visual analytics with biclusters. *Computer*, 46(8):90–94, August 2013. doi: 10.1109/MC.2013.269

[24] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith. Query by humming: Musical information retrieval in an audio database. In *Proceedings of the Third ACM International Conference on Multimedia*, MULTIMEDIA '95, pp. 231–236. ACM, New York, NY, USA, 1995. doi: 10.1145/217279.215273

[25] C. Gormley and Z. Tong. *Elasticsearch: The Definitive Guide*. O'Reilly Media, Inc., 1st ed., 2015.

[26] T. M. Green, W. Ribarsky, and B. Fisher. Building and applying a human cognition model for visual analytics. *Information Visualization*, 8(1):1–13, 2009. doi: 10.1057/ivs.2008.28

[27] A. Gupta and R. Jain. Visual information retrieval. *Commun. ACM*, 40(5):70–79, May 1997. doi: 10.1145/253769.253798

[28] M. A. Hearst. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pp. 59–66. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1995. doi: 10.1145/223904.223912

[29] J. Heer and D. Boyd. Vizster: visualizing online social networks. In *IEEE Symposium on Information Visualization*, pp. 32–39, Oct 2005. doi: 10.1109/INFVIS.2005.1532126

[30] Y. Kang, C. Gorg, and J. Stasko. How can visual analytics assist investigative analysis? design implications from an evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):570–583, May 2011. doi: 10.1109/TVCG.2010.84

[31] X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '91, pp. 262–269. ACM, New York, NY, USA, 1991. doi: 10.1145/122860.122887

[32] K. Logan. *Spatial History: Using Spatial Memory to Recall Information*. PhD thesis, Virginia Tech, 2012.

[33] A. Mack and I. Rock. *Inattentional blindness*, vol. 33. MIT press Cambridge, MA, 1998.

[34] K. A. Olsen, R. R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams. Visualization of a document collection: The vibe system. *Information Processing & Management*, 29(1):69–81, 1993. doi: 10.1016/0306-4573(93)90024-8

[35] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, vol. 5, 2005.

[36] C. Plaisant, G. Grinstein, J. Scholtz, M. Whiting, T. O'Connell, S. Laskowski, L. Chien, A. Tat, W. Wright, C. Grg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko. Evaluating visual analytics at the 2007 vast symposium contest. *IEEE Computer Graphics and Applications*, 28(2):12–21, March 2008. doi: 10.1109/MCG.2008.27

[37] PNNL. In-spire visual document analysis, 2010.

[38] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pp. 1–35. Springer, 2011.

[39] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, and S. Kaski. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pp. 1759–1764. ACM, New York, NY, USA, 2013. doi: 10.1145/2505515.2505644

[40] D. P. Russ Burtner, Shawn Bohn. Interactive visual comparison of multimedia data through type-specific views, 2013. doi: 10.1117/12.2004735

[41] J. Z. Self, R. Vinayagam, J. T. Fry, and C. North. Bridging the gap between user intention and model parameters for data analytics. In *SIGMOD 2016 Workshop on Human-In-the-Loop Data Analytics (HILDA 2016)*, p. 6, June 2016.

[42] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[43] F. M. Shipman and C. C. Marshall. Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work (CSCW)*, 8(4):333–352, 1999. doi: 10.1023/A:1008716330212

[44] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pp. 449–456. ACM, New York, NY, USA, 2005. doi: 10.1145/1076034.1076111

[45] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136, 1980. doi: 10.1016/0010-0285(80)90005-5

[46] K. Vogt, L. Bradel, C. Andrews, C. North, A. Endert, and D. Hutchings. *Co-located Collaborative Sensemaking on a Large High-Resolution Display with Multiple Input Devices*, pp. 589–604. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi: 10.1007/978-3-642-23771-3_44

[47] 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*, HILDA'17, pp. 14:1–14:6. ACM, New York, NY, USA, 2017. doi: 10.1145/3077257.3077259

[48] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of Visualization 1995 Conference*, pp. 51–58, Oct 1995. doi: 10.1109/INFVIS.1995.528686

[49] J. M. Wolfe and T. S. Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):0058, 2017.

[50] J. M. Wolfe and M. J. V. Wert. Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology*, 20(2):121 – 124, 2010. doi: 10.1016/j.cub.2009.11.066

[51] P. C. Wong, E. G. Hetzler, C. Posse, M. A. Whiting, S. Havre, N. Cramer, A. R. Shah, M. Singhal, A. Turner, and J. Thomas. In-spire infovis 2004 contest entry. In *INFOVIS*, vol. 4, pp. 51–52, 2004.

[52] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. The sandbox for analysis: Concepts and methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pp. 801–810. ACM, New York, NY, USA, 2006. doi: 10.1145/1124772.1124890

[53] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. Dust & magnet: Multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256, 2005. doi: 10.1057/palgrave.ivs.9500099