Topic-based Exploration and Embedded Visualizations for Research Idea Generation

Hua Guo, *Member, IEEE,* and David H. Laidlaw, *Fellow, IEEE*

Abstract—This work analyzes sensemaking frameworks and experiments with an iteratively designed visual analysis tool to identify design implications for facilitating research idea generation using visualizations. Our tool, ThoughtFlow, structures and visualizes literature collections using topic models to bridge the information gap between core activities during research ideation. To help users stay focused on a topic while discovering relevant documents, we designed and analyzed usage patterns for two types of embedded visualization that help determine document relevance while minimizing distraction. We analyzed how research ideation outcomes and processes differ when using ThoughtFlow and conventional search engines by augmenting insight-based evaluation with concept-map analysis. Our results suggest that operations afforded by topic models match well with later ideation stages when coherent topics have emerged, but not with early stages when users are still relying heavily on individual keywords to gather background knowledge. We also present qualitative evidence that citation sparklines encourage more exploration of recommended references, and that a preference for paper thumbnails may depend on the consistency between the evidence and the current mental frame.

Index Terms—Empirical study, sensemaking, cognition

1 Introduction

The formation and development of novel research ideas is a complex sensemaking process. We present an empirical study experimenting with ways in which visual analysis systems can support research idea generation. Specifically, we identified two primary sensemaking activities during this process: **framing** (creating and editing the outline of a research proposal) and **elaboration** (gathering related work to support individual arguments), derived design requirements to support the two activities, and distilled design implications from user studies with ThoughtFlow, a visual analysis system we developed to smooth the transition between framing and elaboration using topic modeling.

This paper makes three contributions. First, we demonstrate that grounding the design in sensemaking frameworks helps identify design requirements that target the core sensemaking activities and thus differ from those commonly addressed in text visualization literature. When examined in isolation, neither of the two activities supported by ThoughtFlow is unfamiliar in the visual analytics community. Framing has been studied in the context of analytics environments like Sandbox [1]. For elaboration, a few powerful text-visualization techniques have been developed to facilitate exploration and analysis of scientific literature collections. However, by considering the two activities as the building blocks of research idea generation, we were able to identify new opportunities for bridging the information gap between the two activities. Second, user studies with ThoughtFlow yield empirical validation for an iterative two-phase model of the sensemaking process based on existing sensemaking theories. Finally, after summarizing observed usage patterns of our tool and evaluating analysis outcomes using insight-based evaluation augmented with concept-map analysis, we discuss design implications and lessons from applying the concept-map analysis that could benefit designers of similar sensemaking systems.

The paper is organized as follows. Sec. 3 presents an analysis of initial design requirements derived from user interviews and sensemaking theories. The preliminary tool design and evaluation are described in Secs. 4 and 5. Sec. 6 describes design improvements based on the results of the first evaluation, including a sparkline – word-sized graphics – that conveys the relevance of recommended related work while minimizing space usage, as well as a paper thumbnail of the user's proposal to keep the user oriented while exploring literature collections. Sec. 7 reports an evaluation of the improved design. We conclude by discussing design implications and open challenges in Sec. 8.

• *H. Guo and D. Laidlaw are with the Department of Computer Science, Brown University, Providence, RI, 02912.*
E-mail: huag,dhl@cs.brown.edu

2 Related Work

2.1 Visual Analysis of Scientific Literature Collections

Tools for visual analysis of literature collections usually visualize citation patterns, literature metadata, or content summaries. Our tool also presents this information, but tailors the visualizations to serve both the writing process and the related work search process.

Various visual representations have been developed to convey different aspects of citation patterns. CiteWiz [2] uses a "Newton's shoulders diagram" to show the chronology of publications and researchers, and node-link diagrams to show keyword co-occurrences and co-authorships. PaperCube [3] experimented with an Icicle Tree view to show multilevel citation relationships. Other work has developed techniques to visualize document metadata and contents. PivotPaths [4] can facilitate effective exploration of a literature collection by letting the user choose any of multiple

linked facets of a document collection (e.g. authors, keywords) as visual pivots. CiteRivers [5] visualizes topic structures, author prolificness, and citation patterns of scientific publications and helps the user connect the dots through linked views. SurVis [6] provides selectors on multiple types of document metadata such as dates and keywords to enable versatile queries over a collection.

Our tool integrates all three types of information to support both exploration of literature collections and evaluation of paper relevance. We use topic clusters as the centerpiece for the related work search process, since previous work suggests that citation relationships often have weak correlations with document similarity [7], [8]. Citation patterns are presented alongside other literature metadata and topic information to facilitate judging document relevance.

2.2 Topic Model Visualizations

The visualization and use of topic models in ThoughtFlow differ from existing topic-model visualization techniques in terms of design requirements and evaluation goals.

Many of the visualization techniques developed for topic modeling results are driven by the need to explore high-level properties of and relationships among topics instead of facilitating related work discovery via topic-based exploration. TextFlow [9], TIARA [10], ParallelTopics [11], and Topic Streams [12] use river metaphors to show topic evolution over time, such as changes in term distribution or topic occurrence, to facilitate tasks such as finding main concepts [10] or identifying rising topics [11]. Liu et al. [13] developed a sedimentation-based visualization to show more complex hierarchical topic evolution. Topicpanorama [14] used a novel layout combining a density-based graph and stacked trees to support examination of topic correlations and comparison of common or distinctive topics across different document sources. Termite [15] uses a matrix representation along with novel term- and topic-sorting techniques to assist users in assessing topic model quality. While providing holistic topic overviews, these techniques and systems do not always offer visual representations or interactions that facilitate discovery of relevant topics and documents given specific user interests; this is a key design requirement for the present work.

Some other topic model visualization techniques are designed with targeted topic and document discovery in mind. TopicNets [16] uses a heterogeneous graph to visualize the relationships among documents, topics, and metadata such as author institutions. UTOPIAN [17] represents document clusters as scatter plots and uses nonnegative matrix factorization to update topic clusters through user interaction. Both systems afford discovery of relevant topics and documents by explicitly visualizing document similarity and topic cluster membership. Our tool design differs from these in that we also support term-based topic and document discovery. More importantly, we also study how users approach topic-based exploration during research idea generation and analyze the benefits and limitations of guiding literature navigation through terms and topics.

2.3 Visual Analysis Tools for Sensemaking

This work aims to deepen our understanding of how visual analysis tools can support the sensemaking process by analyzing the usage of a tool that supports a specific type of sensemaking. Below we discuss the similarities and differences between this work and other sensemaking tools.

Sandbox [1] is a sensemaking tool developed for general analytics tasks. It implements various features, such as annotation and linking between items, drag-and-drop interactions, and analysis templates, to facilitate hypothesis generation and evidence collection. We also consider research idea generation as a process that involves hypotheses and evidence (related work). However, rather than support the flexible spatial arrangement of hypotheses and evidence, we choose to let users express and structure their thoughts using a conventional text-editing interface and then identify ways to augment this interface.

Some visual analytics tools for sensemaking emphasize tracking and using the reasoning process to support analysis. Aruvi [18] demonstrates that explicitly recording and presenting the user's exploration process can help analysts reflect on and optimize the workflow. A study with HARVEST [19] suggests that revisiting action trails helps the user stay oriented during the analysis process. We also considered the sensemaking process when designing ThoughtFlow, but instead of provenance, we focus on designing the interface to match the key phases of the sensemaking process and identify opportunities to improve this process.

This work draws upon and contributes empirical validations of guidelines from Green et al.'s Human Cognition Model [20]. For example, we aimed for insulation of reasoning flow by having separated interfaces for the two research ideation activities, giving the user minimal necessary information during each activity.

3 Design with Sensemaking Theories

Below we discuss how we defined the initial scope of the tool based on a simplified two-phase sensemaking model.

3.1 Two-phase Sensemaking Framework

We started by identifying the major phases of the research idea generation process. We consider this process as a type of sensemaking activity: to assess the novelty and feasibility of a research idea, the researcher usually needs to gather and analyze relevant facts from previous work and update her mental representation of the research area accordingly. This characterization fits especially well when a researcher first steps into a less familiar research area.

To derive a framework that would adequately characterize the research idea generation process, we compared four sensemaking frameworks: Russell et al.'s model focusing on the cost structure of sensemaking [21], Pirolli and Card's sensemaking model [22], Klein et al.'s data/frame theory of sensemaking [23], and Zhang and Dagobert's extended sensemaking model [24]. While these frameworks differ in their transitions between basic sensemaking activities and even in the activities themselves, we noted that they all include two phases, called here *framing* and *elaboration*. During framing, the analyst constructs and revises the overall schema (similar to Russell's "representational shift loop;" Pirolli and Card's "sensemaking loop," Klein et al.'s "reframing cycle"); during elaboration, the analyst instantiates

and revises a localized part of the schema. Also, most sensemaking frameworks suggest that the analyst usually loops through the two phases multiple times before completing the sensemaking task.

Given this two-phase framework, we hypothesized that, in writing research proposals, the framing phase maps to 1) creating the proposal outline, often based on the researcher's initial knowledge; and 2) revising and augmenting the overall structure of the proposal, often when more related work is analyzed and triggers new research ideas. The elaboration phase maps onto the process of gathering related work to support or refute a specific argument, as well as any revision of the argument resulting from the information-gathering process.

3.2 Enriching the Model through User Interviews

We then conducted semi-structured interviews to enrich the framework by empirically identifying activities and challenges during proposal writing. We interviewed five researchers – one faculty member, one post-doc, and three graduate students – from either the computer science or cognitive science department at a research university. The interview questions and discussions centered around three aspects of writing research proposals: 1) the general process, 2) tools used, and 3) frustrations or gaps experienced.

We performed open coding of the interview data to extract concepts related to activities and challenges in proposal writing. To capture sensemaking activities that are more fine-grained than *framing* and *elaboration*, we categorized the concepts following the taxonomy of sensemaking activities suggested by the data/frame theory. These activities were then classified as either framing or elaboration. The results show that the two-phase framework could account for the overall flow of the proposal writing process experienced by all the participants.

We identified four categories of sensemaking activities that map onto concrete proposal writing activities. Using terminology from the data/frame theory, framing involves *Seeking a new frame or reframing*, which corresponds to the creation and restructuring of a proposal's flow and arguments, and *Questioning or preserving the frame*, which corresponds to the proposal review and revision based on accumulated evidence. The elaboration phase consists of *Seeking and filtering data* and *Discovering new data or relationships*. These two activities are similar to the *Retrieve* and *Explore* tasks in the visualization interaction taxonomy proposed by Yi et al. [25]. Their differences can be characterized as targeted search versus undirected exploration of the information space: the former usually happens when the researcher has particular keywords, authors, or titles in mind, and the latter when the researcher notices an unexpected thread of related work. The mapping among sensemaking phases, sensemaking activities, and proposal writing activities is shown in Table 1.

We also identified the primary challenges and frustrations experienced by the researchers during each activity:

- **Context switch**: *Switching back and forth between literature search and the writing process or multiple threads of related work costs additional mental energy*. Three participants mentioned context switch as a common issue that can interrupt the flow. Two types of context switch are reported: 1) the switch between articulating a proposal's main ideas and validating whether an argument is well supported by previous work, and 2) the switch among multiple threads of related work.
- **Evidence imbalance**: *Evidence inconsistent with a proposal's hypotheses and arguments is less likely to be discovered and considered during the proposal writing process.* When asked whether they felt they could take both supporting and disconfirming evidence into consideration carefully when articulating the significance and feasibility of an aim, two participants commented that they had, at least once, received feedback on their proposals that pointed out evidence inconsistent with their hypotheses or arguments but missed during literature search. This is consistent with findings suggesting researchers could experience confirmation bias during scientific reasoning [26] and thus be less likely to actively watch for evidence inconsistent with a hypothesis.
- **Disciplinary barriers**: *Users are less likely to discover publications relevant to a proposal but from a less familiar research area.* Some research topics are studied by researchers from multiple areas, and titles and abstracts of papers from different disciplines sometimes emphasize different elements or even use somewhat different terms to discuss the same topics. Hence literature reviews using keyword-based search often miss publications from unfamiliar areas. This issue was raised by two participants when asked about their confidence in the comprehensiveness of their literature review while writing proposals.
- **Citation-based clusters**: *Citation relationships, while used extensively by the participants, do not guarantee relevance or comprehensive coverage of related work.* All participants reported that they relied heavily on citation relationships to identify related work. However, it has been shown that citation relationships often have low correlation with subject similarity [7], and reliance on citation relationships in literature search may result in missing related work outside a familiar citation network.

The first two challenges are associated with both framing and elaboration. However, we classified them as framing challenges because they arise due to the need to create and maintain unbiased explanatory frames. The last two challenges are associated primarily with elaboration activities.

With the above challenges and the two-phase framework, we derived two high-level design requirements:

- **R1: Support both framing and elaboration**. We hypothesized that by providing an integrated environment that supports both phases, we could identify information gaps during these two phases that increase context-switch cost and then refine the interface design to reduce that cost.
- **R2: Provide multiple entry points for identifying relevant work**. We hypothesized that the last three challenges could be partially addressed by enabling exploration of paper clusters that reflect semantic similarities. Therefore, we aimed to support exploration based on topic modeling alongside keyword-based search.

4 Design Iteration 1

We first implemented two views, one for editing the proposal and the other for exploring a literature collection.

Phase	Sensemaking activity	Proposal writing activity	Challenges
Framing	Seeking a new frame / Reframing	Writing and restructuring the proposal outline	Context switching
	Questioning / preserving a frame	Explain / incorporate new evidence	Evidence imbalance
Elaboration	Seeking and filtering data	Search for papers by keywords, authors, or title	Disciplinary barriers
	Discovering new data / relationships	Explore unexpected threads of related work	Citation-based clusters

TABLE 1: Mapping among sensemaking phases, abstract sensemaking activities, and proposal writing activities, as well as challenges identified in interviews with researchers in computer science and cognitive science.

Both use topic-modeling results from a Latent Dirichelet Allocation [27] to help discover related work. This section describes the design of the two views and how we used topic modeling to generate data for the interface.

4.1 Design of the *Write* View

The *Write* view resembles a standard and familiar text editor. When the user enters the view for the first time, she can either create a new document or upload an existing proposal draft. On the right side of the view is the *Reference Panel*, which contains three types of references: recommended, cited, and bookmarked. Recommended references are updated every time the user selects a paragraph by computing the topic distribution for the paragraph. Documents with topic mixtures similar to that of the paragraph are returned as recommendations. Users can bookmark publications when navigating the *Explore* View, which appears in the bookmarked references list. References from both lists can be cited as relevant to a selected paragraph and are then added to the citation list.

4.2 Design of the *Explore* View

The *Explore* View is designed to facilitate semantics-driven navigation of a literature collection. We identified three entry points that users can pick to express their research interests: term, topic, and publication. A term is a single word that can be a part of a topic or a publication's title or abstract. A topic contains a list of weighted terms and a set of documents. A document has a topic distribution and a set of terms in its title and abstract. Below we describe how the *Explore* View supports navigation between each pair of entry points.

The top half of *Explore* View is shown in Fig. 2. On the left, an overview contains the terms and topics identified from topic modeling. Initially, we experimented with two layouts commonly used for visualizing topic models: matrix and node-link diagram. However, the matrix representation led to much wasted space due to the sparsity of the term-topic matrix. The node-link diagram is space-efficient but does not support easy ordering of terms and topics, which is important given users' strong prior research interests in our use case. We eventually settled on a parallel-list representation: a list of terms and a list of topics. In the term list, every term is represented using a bar, the width of which is proportional to the weighted sum of the term's weight across all topics. In the topic list, each topic is represented using a segmented bar, with each segment representing a term contained in the topic. The width of each segment is proportional to the weight of the term in the topic. Each term in the term list is connected with all containing topics, and the user can mouseover or click a term to highlight topics containing that term (**term** $\rightarrow$ **topic**). The user can also click on a segment in a topic to select the corresponding term (**topic** $\rightarrow$ **term**). Both lists are sorted, and only the top 50 terms and topics are initially visible. The term list is sorted by the sum of term weights across all topics, and the topic list is sorted by the weighted sum of the weights of all its terms. The user can scroll through both lists to browse more terms and topics or use a search bar to promote terms to the first page. The term list can be reordered to prioritize terms that frequently co-occur with selected terms. The topic list can also be reordered based on the total weights of selected terms in each topic.

The circle to the left of each segmented bar serves as both a topic selector and an indicator of the number of publications assigned to a topic, with its area proportional to the number of publications. Clicking on the topic selector updates the document panel (not shown in Fig. 2) to contain a list of documents from the selected topic cluster (**topic** $\rightarrow$ **document**). The user can use the document panel to read paper metadata and abstracts and bookmark papers for later review. We initially used rectangles instead of circles, with the length of each rectangle proportional to the number of publications. However, users found the bars unintuitive as topic selectors, and since they did not feel it necessary to accurately estimate the size of each topic cluster, we used circles to invoke the radio-button metaphor.

The interface also accommodates explicit query for publications. The user can directly input keywords to see a list of documents containing those keywords in the title (**term** $\rightarrow$ **document**). Selecting a publication returned by the query reorders the term topic lists so that terms and topics associated with the selected publication are ranked higher (**document** $\rightarrow$ **term** and **document** $\rightarrow$ **topic**).

4.3 Use Cases

Below we walk through a set of example use cases from the user study to illustrate how users interact with the tool.

Explore related work by searching by paper title. The user actively looks for specific related work by searching for papers about "frontal lobe connectivity" in the *Explore* view. She selects a paper from the search results, and the term and topic lists automatically update to be sorted based on associations with the paper. The primary topic assigned to the selected paper is ranked first in the topic list by default, and terms associated with the primary topic are highlighted. The document panel also updates to show papers from the primary topic cluster, automatically scrolled down so that the paper selected by the user is in view.

Explore related work by following a recommended citation. After finishing the draft of a paragraph in the *Write* view, the user views the list of papers recommended by the system

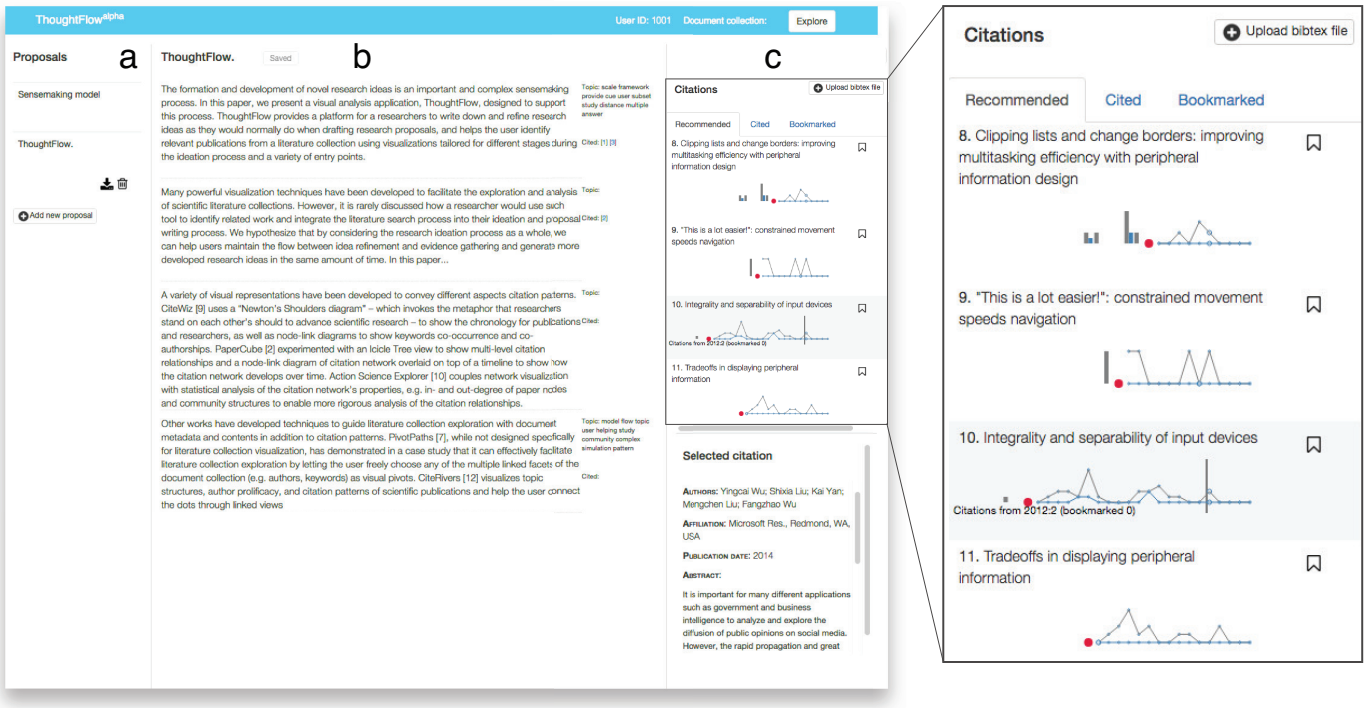


Fig. 1: The *Write* view. Part a is a panel for managing the proposals. Part b is the proposal editing area. Part c is the *Reference Panel* that displays recommended, cited, and bookmarked references. Each citation entry contains a citation sparkline, which is introduced during the second design iteration. In each sparkline, the red dot represents the current citation. The bars to the left of the dot represent reference counts and the curves to the right of the dot represent citation counts. The horizontal positions of the marks represent years. Bars and curves in gray represent total counts of references / citations, while those in blue represent counts of references / citations bookmarked by the user. This helps the user judge which recommended citations are related to publications that she is aware of and which are outside the radar. For instance, citation 10 has been referenced by some publications in the user's bookmark, while citation 11 may lie outside the body of work the user is familiar with.

and chooses one to explore. This action takes her to the *Explore* view. The term list, topic list, and document panel have updated based on the selected recommendation. The user finds and bookmarks a relevant paper from the same topic cluster as the recommended paper.

Explore related work by searching for a specific term. The user wants to focus on studies on anxiety with specific populations. She searches for the term "anxiety", and the topic list is sorted so that those containing the term "anxiety" are at the top of the list. The term list is also sorted so that the terms co-occurring with "anxiety" in topic clusters are ranked higher. The user browses the term list and notices the term "children". She mouses over the term and finds a topic cluster about anxiety in children and development.

4.4 System

This section describes how the system performs topic modeling over literature collections to generate topic clusters and enable citation recommendation.

4.4.1 Creating literature collections

Researchers usually have one or more specific research interests before generating research ideas, and ThoughtFlow lets the user work with literature collections fitting their research interests. For our user studies, we created a literature collection containing 12,689 SIGCHI publications extracted from a publication dataset curated by ArnetMiner [28]. In addition, ThoughtFlow provides two ways to create literature collections by querying PubMed: 1) programmatically submitting user-provided search queries to PubMed; 2) taking a user-uploaded bibtex file and querying PubMed to download metadata for each entry in the bibtex file together with their references and citations. We use an open-source PubMed library, metapub [29] for all the PubMed queries, and all publications are stored in a MySQL database using Django. Most user-study participants chose to work with customized literature collections instead of the SIGCHI dataset, and we used eight different literature collections in total in the user study, with sizes ranging from 337 to 64,714 publications.

4.4.2 Training and using topic models

Once the document collection is generated, ThoughtFlow performs topic modeling on the collection to identify topics and assign topic distributions to each paper. We use an off-the-shelf implementation of Latent Dirichelet Allocation [27] from *gensim*, a Python Library. A topic model is trained on each literature collection and cached for repeated use. The number of topics is set to be equal to the total number of publications in the collection divided by 50.

To recommend citations for a user-created paragraph, the system first uses the cached topic model to compute the paragraph's topic distribution. It then retrieves all publications from the literature collection with similar topic mixtures, ranks them based on cosine similarity between their abstracts and the paragraph, and returns the ranked list as the recommendations.

Fig. 2: The top half of the *Explore* View (the bottom half contains a document panel, not shown here). Part a shows two search bars for selecting terms in the term list and finding publications with given terms, respectively. Part b shows lists of terms and topics extracted from the current literature corpus, and the links between the two lists indicate term membership. Part c shows results of publication search results, grouped by primary topics. Part d is the paper thumbnail introduced during the second design iteration , where the leftmost rectangles are citation containers. The yellow rectangle within the top container shows that one paper has been cited in the proposal's first paragraph. The name of the citation is shown on mouseover, and the citation can be selected so that the user can explore publications on the same topic. The area with green texts is a reduced view of the full proposal. The texts become more transparent the longer they remain unchanged. The middle section is a set of paragraph selectors for use in jumping to the selected paragraph.

5 User Study 1

After the preliminary design had been implemented, we conducted a user study to understand usage patterns of the tool and identify further design improvements. We were interested in answering the following questions:

- Is the design of the topic modeling visualization intuitive, i.e. can users correctly interpret the connections among documents, terms, and topics?
- How do users locate interesting document clusters? What entry points do they use?
- Do terms and topics provide sufficient cues to judge a paper's relevance?

5.1 Study Design

Four graduate students from the computer science department participated in the study. Three of them were interviewed to solicit design requirements (the other two who participated in the initial interview were now unavailable because of relocation). Before the study, participants were asked to prepare and bring with them 1-3 paragraphs from a research paper or proposal that they were working on. The paragraphs were to describe nascent ideas for which they would like to find more related work for to demonstrate the ideas' novelty, feasibility, and significance. During the study, each participant was briefly introduced to how to use the interface and was then asked to use the tool for around 30 minutes to develop those paragraphs further. We used a think-aloud protocol and captured screen recordings during the user study.

5.2 Analysis and Observations

We performed qualitative analysis of user-study videos and user-interaction logs. We visualized the interaction logs to identify interesting behavioral patterns and analyzed corresponding user comments from the study videos. Overall, participants found the term-topic visualization helpful and intuitive, and they were able to find new relevant publications by exploring topic clusters. However, we also observed the following suboptimal analysis patterns.

Lack of citation information discourages the use of citation recommendation. Participants felt that the citation recommendation could be useful, but they rarely used it during the study sessions. Feedback from the post-study interview suggests that sometimes participants could not judge a recommendation's relevance or importance just by glancing at the titles, and decided it would be more straightforward just to search for information directly.

Users frequently switched between *Explore* and *Write* views to reorient themselves while searching for related

work. This pattern is exemplified by the interaction log shown in Fig. 3. We observed four triggers for switching: 1) to find supporting evidence for contents just created; 2) to elaborate on evidence collected; 3) to cite evidence collected; 4) to reread contents created and decide what evidence to search for next. The first two types of triggers were expected and necessary. The last two, however, could be avoided, thus reducing context switching.

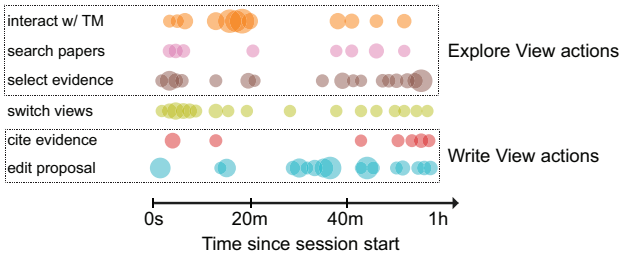


Fig. 3: Interaction log of a session showing frequent view switch during literature search. Each circle represents one or more actions of the same type. The top three rows of actions happened in the Explore view and the bottom two in the Write view. The first 15 minutes contained many view-switch actions that were not followed by proposal-editing actions, since the user was simply switching to cite evidence or to reorient and decide what to search for next.

6 Design Iteration 2

Given the two observations from the first user study, we added two visual designs to ThoughFlow to address the following two design requirements:

- **R3: Provide citation information in the recommended citation list to convey relevance.** The first observation from the first user study suggested that the citation lists in the *Write* view needed to be augmented with information beyond their metadata and abstracts to help the user judge relevance. In particular, participants remarked that they would be interested in seeing citation information of the publication to help decide whether to look at its related publications in the *Explore* view.
- **R4: Preserve context of the proposal-in-progress during elaboration.** The second observation inspired us to design a visual component that 1) enables citation during exploration and 2) provides adequate information about the proposal-in-progress in the *Explore* view to help users reorient and decide what evidence to search for.

In designing the components, we wanted to keep them both visually and conceptually grounded or *embedded* in the text-editing process and the proposal text being created to provide additional information without excessively distracting the user from the writing process.

6.1 Sparkline visualizations of citation patterns to convey publication relevance

In designing the citation pattern visualization for recommended citations, we aimed to convey the citation information while maintaining the focus on individual citations rather than drawing attention to properties of the citation network. Therefore, we designed the citation pattern visualization as a sparkline (Fig. 1c). The sparkline created for each recommended citation shows three types of information: the publication year of the citation, publication years of its references, and the number of times it has been cited per year. Reference counts and citation counts are further divided into "total counts" and "bookmarked counts." The x-axis of the sparkline maps to years. The red dot indicates the publication year of the current publication. Every gray bar preceding the red dot represents the number of references in the current publication that were published that year. The blue bar, on the other hand, indicates how many references from that year have been bookmarked by the user. Similarly, the gray lines to the right of the red dots indicate the number of citations the current publication has every year, while the blue lines show how many of those citations are in the user's bookmarks. As a whole, the citation visualization shows the overall importance of the citation within the research community and also the extent to which it connects to information known to the user. For example, seeing a publication with many citations in total but none bookmarked by the user may suggest that this is an influential publication currently outside the user's radar.

6.2 Paper thumbnail summarizing proposal states

The paper thumbnail is designed to remind the user of her progress and current focus in order to reduce unnecessary transitions between exploration and framing. It (Fig. 2d) has three parts. To the right of the thumbnail is the *text age tracker* showing the contents of the proposal being written, with text transparency indicating its recency: the newest edits are opaque, while the longer a piece of text stays untouched, the more transparent it becomes. We colored the text in green to distinguish the paper thumbnail from the proposal editor and to invoke a freshness metaphor. We observed that users were glancing at the overall structure of and keywords in the thumbnail to remind themselves of key ideas and reference needs, and the choice of color seemed to match well with how the thumbnail was used. The thumbnail's middle section contains a set of *paragraph selectors* that uses the same coloring scheme as the text. Users can click on a selector to jump to the corresponding paragraph or attach a citation to the paragraph. The leftmost section contains a set of *citation containers*, with each yellow rectangle representing a citation for that paragraph. For example, Fig. 2d shows that the user has identified one citation for the first paragraph in the proposal and none for the other paragraphs.

7 User Study 2

After the second design iteration, we conducted another user study to assess how the use of the tool influences the user's research proposal-writing process and outcome. Participants were asked to write short research proposals using either ThoughtFlow (**TF**) or a combination of an academic search engine (Google Scholar or PubMed) and Google Docs (**ASE**). We analyzed characteristics of these research proposals, their subjective estimate of effort spent under the two conditions, and tool usage patterns. Our evaluation of the research proposals' characteristics integrates concept-map analysis [30] into insight-based evaluation [31] to capture global properties of frames established by the user in the proposal. Our study aimed to answer the following questions:

- Do users generate research ideas with more breadth, depth, and supporting evidence using **TF** than **ASE**?

- Does the proposal-writing process (e.g. transitions between **framing** and **elaboration** and approaches to finding related work) differ with **TF** and **ASE**?
- How do users use specific visual components of the tool and how does their usage influence the analysis process?

7.1 Study Design

Five graduate students studying computer science or cognitive science participated in this second study. None of the participants was in the first study, so as to ensure that all participants were new to the tool and also to gather feedback and observations from a broader group of participants. Since we were interested in how researchers approach research proposal writing using different tools, we asked each of them to come up with two research topics to develop into mini-research proposals during the study using two set of tools. To make sure that participants would actively use the tools to search for related work and generate new ideas during the study, we asked them to choose topics of which they had basic knowledge but on which they had not done extensive research. They were also asked to choose two topics that they were equally familiar with. The prompt used in the user study is provided in Appendix A.

During the user study, each participant attended two 1-hour sessions. In each session, the participant worked on one of their two topics, using either **TF** or **ASE**. Three participants used ThoughtFlow in the first session and two used the academic search engine first. Each participant was given a 10-minute introduction on using ThoughtFlow at the beginning of the ThoughtFlow session. The two sessions for the same participant were scheduled on consecutive days to reduce the influence of the topic and information from the first session on the second.

As in the first study, we used a think-aloud protocol and captured screen recordings of the sessions. We also collected user interaction logs. At the end of each session, the participant filled out the NASA Task Load Index (TLX) questionnaire, a multidimensional rating scale for self-reporting cognitive workload. We also interviewed participants to collect their feedback on the overall process as well as on the individual components and functionality of the tool.

For all quantitative comparisons below, we report effect sizes measured using Cohen's d. Effect size is an important measure [32] [33] because, unlike p-value, it is not confounded by sample size and better captures practical significance [34]. It is often preferred to p-value when the analysis goal is to interpret results from a small sample without intending to generalize them to a broader population [35]. As we intended to use the quantitative comparisons to drive and complement the qualitative analysis, effect size was more appropriate than p-value, with the caveat that the reported differences cannot be generalized to all potential users of the tool.

7.2 Case Study

We describe one participant session below to demonstrate the types of workflows and insights afforded by the tool, labeling key operations with their corresponding requirements (R1 or R2, as detailed in Sec. 3.2). The participant was a second-year graduate student in cognitive science who wanted to write a proposal on the effects of anxiety on cognitive control as a potential topic for her preliminary exam paper. After outlining the aim of "investigating the effect of trait anxiety on executive control using brain imaging and analysis of functional activity", she switched to the *Explore* view to look for papers containing the keywords "frontal lobe connectivity", since the frontal lobes play an important role in regulating executive control. Once the user selects a paper from the search result, the term topic lists update to highlight the paper's topic and corresponding terms, which cover some of the frontal lobe's functions, such as "orientations" and "social" [R2]. The user browsed and bookmarked a few papers from the cluster.

The user then switched back to the *Write* view and continued to elaborate on the motivation for investigating the effect of anxiety on executive control from the connectivity angle. While she was writing, the citation recommendations updated [R1]. She browsed the recommendations and cited three papers, commenting that the she found the recommendation really helpful because the papers were all related to but focused on different types of anxiety and connectivity analysis methods, and multiple queries would probably be needed if she had to rely on keyword-based search. While browsing, she also used the citation sparklines to compare papers with similar topics based on their publication time and citation counts and to choose those with higher and more consistent numbers of citations over the years. She then updated her proposal's motivation accordingly to discuss the implication of having these different types of anxiety and connectivity analysis methods, referencing the most cited papers in each category [R1].

The user also noticed a recommended citation about how gender difference mediates the effect of anxiety on decision making. She bookmarked and clicked on the citation to be directed to the *Explore* View, which highlighted the topic cluster with the selected paper and terms such as "hormone" and "maturation" [R1, R2]. She bookmarked some papers about gender and age difference on anxiety and commented that she had "always been interested in gender difference but didn't think of connecting that with anxiety". She then went on to write about the significance of the proposal, mentioning the impact of understanding individual differences in anxiety levels. In the post-study interview, the participant commented that the topic cluster view let her quickly determine there was adequate related work on individual difference and anxiety by glancing at paper abstracts and keywords on this topic. Without the recommendation and topic cluster view, she might not have discovered this angle, and even if she had, she would have had to spend much more time evaluating its feasibility and might have been distracted from the main theme.

Finally, the user started writing about the feasibility of the proposal. She outlined existing applicable techniques and switched to the *Explore* view to look for references using each of the connectivity analysis techniques. She cited a few papers using the paper thumbnail, and then switched back to elaborate the feasibility section while glancing at the cited papers [R1]. In the post-study interview, the user mentioned that the paper thumbnail was especially useful for the feasibility section because it reminded her to enumerate the techniques. It also helped reduce distraction since she didn't have to switch constantly between searching for papers and

summarizing the technical results.

The user finished the proposal by, again, browsing citations recommended for the significance section, to see if she could identify additional areas that could benefit from the proposed research. She noted and cited one paper on eating disorder, which had many total citations but few from the ones she had already collected during the session, commenting that the visual pattern stood out and the paper was important and represented a less familiar research area [R2]. It took the user 45 minutes to finish the proposal, of which around 16 minutes were spent on framing, i.e. editing the content of the proposal.

The workflow observed in this case study exemplifies the two-phase sensemaking frame that we characterized in Sec. 3. We observed, for example, that the user transitioned into *elaboration* driven by information needs on frontal lobe connectivity and transitioned to *reframe* the proposal based on recommended citation on gender difference and anxiety. The active use of and positive feedback on the citation sparklines and the paper thumbnail suggest that these two components helped reduce context switch and address **R1**. The use of citation sparklines to identify work outside the cluster of currently cited papers addresses the challenge of overreliance on citation relationships and demonstrates the benefit of multiple entry points for finding related work (**R2**).

7.3 Analysis of Proposal Characteristics

Our analysis of the proposals was partly inspired by the insight-based evaluation methodology. We consider each proposal, together with all relevant papers that have been cited or bookmarked, as an agglomeration of all insights that the user has accumulated or activated throughout the user study session. While it is impractical to accurately assess the influence of the user's prior knowledge of the topic on each final proposal, our pre- and post-study interviews suggest that they indeed had minimal knowledge of each topic before the user study. While most participants were aware of one or two pieces of potentially related work before the study, all reported that they could not write anything more than a speculative outline without further literature search and analysis. Therefore, we believe it is safe to assume that participants' prior knowledge had equal and minimal effect on the final proposals and the analysis process.

We then characterized proposal quality and insights generated in each session on the following aspects:

- **Amount of supporting evidence**: numbers of related studies that the user has cited to support arguments in the proposal or has bookmarked for further analysis
- **Conceptual structure**: properties of the network of concepts and relationships elaborated in the proposal

7.3.1 Amount of Supporting Evidence

The amount of supporting evidence was obtained directly from analyzing user-created contents and reflects the efficiency of information foraging during the proposal-writing process. We consider both cited and bookmarked articles as supporting evidence. Cited articles are those explicitly referenced by the participant in the proposal, and bookmarked articles are those saved by the participant (by copying and pasting when using Google Docs or by using the bookmark feature in **TF**) but not explicitly used as references.

While coding the number of citations and bookmarks, we verified that each publication was indeed relevant and would help strengthen the proposal. Participants would often bookmark instead of citing an article if it was considered potentially useful in proposal revision or another proposal on the same topic. Therefore, we consider the number of bookmarked articles as an indicator of the amount of evidence collected that could potentially lead to reframing or new frames of the participant's research plan if the participant were given more time.

On average, participants cited 4.6 articles for every proposal with both interfaces (SD=1.14 for **ASE**, SD=1.52 for **TF**). Participants bookmarked more articles when using **TF**: an average of 0.8 articles (SD=1.1) using **ASE**, and 3.2 articles (SD=1.8) using **TF**. The difference in the number of bookmarked articles has a large effect size (Cohen's d) of 1.61 (95% CI [.12, 3.04]), suggesting that participants could discover more evidence for potentially reframing and generating new research ideas using **TF**.

7.3.2 Conceptual structure of the proposal

We analyzed the conceptual structure of the proposal to investigate whether the use of ThoughtFlow influenced the properties of frames constructed by the participants during sensemaking. We hypothesized that topic-based exploration would lead to more divergent thinking and users would explore more alternative angles when elaborating the core aims. This hypothesis is grounded in cognitive theories about creativity, which suggest that presenting external stimuli to a user can lead to unexpected insights by introducing potential analogies that might not be immediately available from prior knowledge [36].

To test the hypothesis, we constructed a concept map for each proposal. A concept map is a graphical knowledge representation that depicts a set of concepts and relationships among them as a node-link diagram [30]. Concept maps are often used in education to study learners' knowledge of a domain and have also been used for usability evaluation [37], but this is the first attempt (as far as we know) to apply concept mapping to evaluating visual analysis tools. Appendix B shows example concept maps of proposals created by the same participant using **ASE** and **TF**, respectively (Fig. 5 and Fig. 6). Two coders created the concept maps separately, and then resolved differences in the two maps created for the same proposal through discussion.

We defined the measures below for each concept map:

- **Total number of concepts**: a *concept* is a noun (phrase) that is key to the main ideas expressed in the proposal
- **Total number of relationships**: a *relationship* describes connections among multiple concepts, e.g. "Parkinson Disease medication increases the level of striatal dopamine"
- **Normalized degree of the core-concept group**: a *core concept* is a concept that appears in a one-sentence description of the proposal provided by the participant at the beginning of each session; this measure was computed by counting the number of all non-core concepts connected to at least one core concept and then dividing the count by the total number of core concepts

	P1		P2		P3		P4		P5	
	ASE	TF	ASE	TF	ASE	TF	ASE	TF	ASE	TF
concepts	18	18	13	27	9	8	16	14	17	20
relationships	19	21	13	28	10	7	20	17	16	23
core-concepts degree	1	1.7	1.5	4.5	1.5	4	2.7	2.5	2	5
map diameter	4	3	2	3	3	3	2	4	2	3

TABLE 2: Concept-map-based measures of individual proposals. Of the five participants, two of them (P2 and P5) created proposals higher in all four measures using ThoughtFlow. The other participants show no consistent differences in the four measures between their two proposals.

- **The diameter of the concept map**: the longest among all shortest distances between each non-core concept and any core concept

The first two measures are analogous to insight characteristics "observations" and "generalizations" used in previous studies (e.g., [38] and [39]). The last two measures have not been applied in insight-based evaluation before. The degree of the core-concept group is used to capture the number of different angles the user has explored to elaborate the significance, novelty, and feasibility of the proposal. The diameter of the concept map is meant to capture the depth of an argument about a core concept. We believe these concept-map-based measures help capture global properties of insights generated by the participants and are complementary to other characteristics commonly used in insight-based evaluation.

Individual measures for each proposal are shown in Table 2. On average, proposals created using **TF** are higher in all four measures. On the other hand, while two participants created proposals with consistently higher measures, the effects for the other three are much smaller and limited to certain measures.

7.4 Analysis of Proposal Writing Process

Given the observation that participants created proposals with richer concepts and relationships around the core concepts using **TF**, we analyzed users' proposal-writing processes and subjective cognitive load to investigate whether the difference in proposal structures could be partially attributed to reduced context switch and other factors manifested in the sensemaking process.

7.4.1 Sensemaking Activity Patterns

We focused on the transitions between writing and related-work search and observed the following:

Starting state is a personal preference and does not change depending on the interface. Participants started the sessions by either writing the outline of the proposal or searching for related work to get more background knowledge. We observe that each participant always started from the same state in both sessions, i.e. some participants started by writing an outline first and others started by searching for related work first, regardless of what tools they were using. Therefore, whether to start with framing or elaboration seems to be an individual preference and seems not to be influenced by the integrated environment provided by the tool.

Transition patterns between elaboration and framing differ between two interfaces. While the high-level process under both interface conditions can be described as a cyclic transition pattern between framing and elaboration, the number of transitions differs between the two conditions for some participants. Participants on average transitioned between the two phases nine times during each **ASE** session, and an average of seven times with **TF**. However, the standard deviation of number of transitions is much larger with **TF** (SD=6.38) than **ASE** (SD=1.83). Looking more closely at the data, we found that two participants made only one or two transitions with **TF** and seven or eight transitions with **ASE**. The other participants made similar numbers of transitions (between 10 to 13 per session) under the two interface conditions. We found from the video that the number of transitions reflects two different overall research ideation strategies. In the session with only one transition, the participant did not start writing the proposal until enough related work had been collected. In the session with two transitions, the participant started by writing a detailed draft proposal, went on to find supporting evidence, and came back to revise the proposal near the end of the session. In all other sessions, participants would often make immediate edits to the proposal when they found new related work. These observations suggest that ThoughtFlow reduced context switch for at least some of the participants.

7.4.2 Cognitive Load Analysis

The perceived overall effort, performance, and frustration were approximately the same in **TF** and **ASE**. On the other hand, participants reported slightly higher physical demand (3.8$\pm$0.84 for **ASE**, 3.3$\pm$0.67 for **TF**, d=0.66, 95% CI [-.64, 1.92]) and mental demand (4.4$\pm$0.55 for **ASE**, 3.8$\pm$0.57 for **TF**, d=1.07, 95% CI [-.30, 2.39]) when using **ASE** (d is Cohen's d), commenting that it was more fun and easier to keep track of references using **TF**, and this may help lower both kinds of demand. However, some also felt that the overhead of learning the system increased the mental demand.

Interestingly, while the average temporal demand was close in the two conditions, two participants reported feeling a bit more rushed when using **TF** because of the amount of relevant information available to them. These two participants were those who created proposals with higher core-concepts degrees and map diameters using **TF**, suggesting that in this case perceived temporal demand might be a reflection of the amount of perceived valuable information made available by the tool.

7.5 Analysis of Use of Individual Components

7.5.1 Paper search by keywords complements topic-based exploration but requires unified entry point

Participants generally liked being able to select a topic cluster associated with a specific paper. Some participants

used keyword search to locate papers they had read before, and then explored other papers from the same topic cluster. One participant compared this type of exploration with direct selection of topics, and felt that in the former case, the paper **exemplified** the semantics of the topic cluster and helped her make sense of the topic.

On the other hand, the coexistence of two search options – topic term search and paper keyword search – was confusing to some participants. Sometimes the user would mistake the term-search bar for the paper-query bar. Another source of the confusion was the different interface responses after the two search actions. One participant asked why the document panel was not updated when she selected a term in the term-topic visualization, expecting term selection to return an immediate list of results, as in title search.

This confusion can be partly attributed to a mismatch between users' mental models of the literature search process and the mixed-entry-point model used in ThoughtFlow. Three participants explicitly mentioned that it took them a while to learn the different ways of finding publications with ThoughtFlow's interface because they were used to keyword-based search with regular search engines. The term-topic visualization, on the hand, asks the user to perform a two-level search – first locating a term and then a topic – and this may take a while to get used to.

7.5.2 Effectiveness of topic-based exploration depends on the analysis process

The use of the term-topic visualization can be classified into two categories: *term-based search* and *topic exploration*. *Term-based search* involves using the search bar or browsing the list of terms to find terms that might lead to relevant topics. *Topic exploration* can happen after either initial *term-based search* or after locating the topic and top terms associated with a paper, and involves further selecting and reordering terms to filter and examine the topics.

The separation of the two types of interactions with the topic model led to the following observations.

Observation 1: term-based search was more effective during later stages. Early in the sessions, participants preferred and tended to have more success with directly searching for publications with specific search terms in the title and abstract. Term-based search usually led the user to a list of publications loosely related to the search terms, and users tended to give up quickly when they could not find publications that closely matched the specified terms.

However, topic-based exploration became more favored towards the end of the sessions, when the participants felt they had exhausted publications with the keywords they had in mind. For example, a participant writing about cognitive effort started to work with the term-topic visualization after using keyword-based search almost exclusively for about 30 minutes, stating that she would like to *"check if I have missed anything."* By selecting the term "effort" and reordering the term list, she noticed that the term "fatigue" ranked high in the list after reordering, which led to two relevant papers that would otherwise have been missed.

Fig. 4 shows a subset of actions from three sessions that exemplify this observation. In all three, participants tried term-based search ("search w/ TM", row 1) early in the session, but these searches did not lead directly to any citation or bookmark (row 5). These participants tried term-based search again after finding some related work using paper-search and writing for a while, and were able to find relevant publications the second time around.

To assess whether this behavior can be attributed to topic quality, we analyzed user think-aloud data to gauge perceived topic quality, since conventional objective measures, such as held-out predicted likelihood, may not reliably capture the interpretability of the inferred topics [40]. Participants were asked to report every time they noticed a topic that seemed difficult to interpret, as well as to comment on inconsistencies between the perceived themes of documents in a topic and the themes implied by top terms in a topic. Overall, all participants noticed one or more topics with interpretability issues or semantic incoherence between topic and documents. However, they commented that topic quality did not seriously hinder their exploration of related work using the topic models.

Observation 2: mismatches between generated topics and initial user targets may lead to unexpected insights. Participants reported occasional mismatches between the topic clusters that the user hoped to see and the actual topic clusters presented. However, such mismatches sometimes led to unexpected insights. For example, one participant wanted to find documents about the use of *color* in the *design* of *fitness* apps. The system did not provide any topic with that set of terms, but did show a topic cluster about color and productivity. The participant found the topic unexpected but useful, since *"[productivity] is also about behavioral change, which is what I'm really interested in. [...] I might be able to use some of the methodologies from these studies."*.

7.5.3 Citation sparklines led to more examination of recommended citations

All participants viewed the citation recommendation functionality favorably. Four participants bookmarked at least one citation recommended by the system, and one of them used recommended citations as the starting point to explore other papers from the same topic cluster. The citation recommendation was used much more frequently in the second user study than in the first, suggesting that the sparklines encouraged more consideration of recommendations.

Participants commented that it was helpful to see the temporal distributions of paper citations and references, especially for recognizing *out-group* and *new and noteworthy* papers. *Out-group* papers are those with many citations overall but few from papers that the user has bookmarked. In some cases, such papers turned out to be seminal or review papers on relevant findings from a remotely related research area: *"This paper is interesting because it reviews the application of color in education. I would not have thought about searching for related work in this area and it is pretty cool the tool recommends it to me."* With the sparkline, newly published papers with a high citation rate but small total number of citations are also more likely to be noticed.

On the other hand, some participants also questioned whether citation information alone was adequate for judging the relevance of a paper. Some participants requested adding to the sparkline author and publication venue information, which they would often use to assess the credibility of a paper. Another perceived limitation of the sparkline, as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2873011, IEEE Transactions on Visualization and Computer Graphics

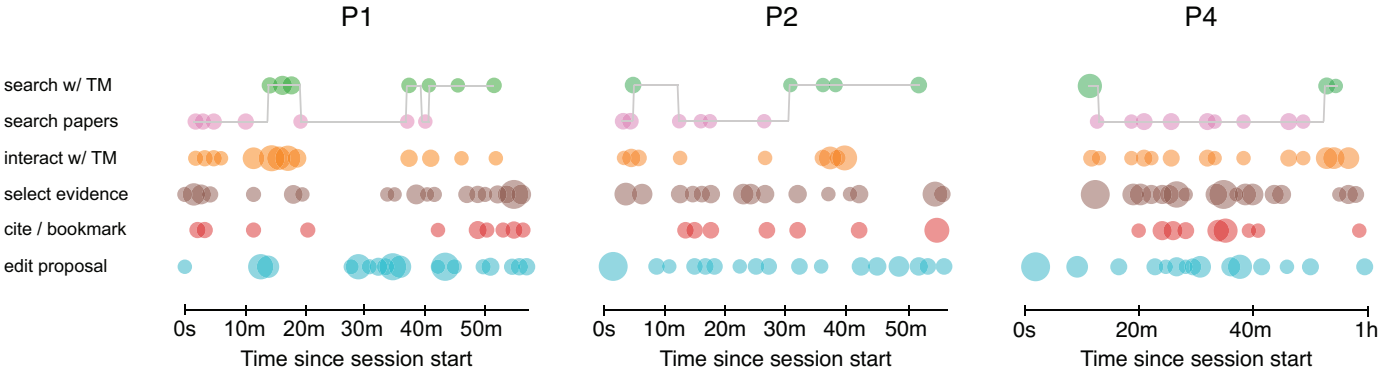


Fig. 4: Interaction logs of selected actions from three user sessions. We connected all *search w/ topic model* and *search papers* with lines to signify the transitions between the two types of searches. In all three sessions, participants shifted from paper search to search with topic models as the session progressed. *Search w/ topic model* early on in the session led to few citations or bookmarks.

remarked by one participant, is that it might distract the user from a paper's content: *"I sometimes would skip over papers with no citations, [...] I wonder if I have missed papers that are relevant and just don't have lots of citations yet"*.

7.5.4 Preference for in-place citation with paper thumbnail depends on the consistency between evidence and frame

Participants found the paper thumbnail helpful, especially for making immediate citations without switching back to the *Write* view. Two participants made all their citations through paper thumbnail, and two participants used a mix of paper thumbnail and bookmarks in the *Write* view to add citations. They all appreciated the ability to cite a paper during literature search, which saved time and reduced cognitive load vs. switching back to the proposal and reexamining the bookmarked evidence. One participant commented that the text age tracker was a useful reminder when he needed to decide what related work to search for. On the other hand, two participants felt that the paper thumbnail was unnecessarily dense, since they were relying on topic sentences and keywords from each paragraph in the thumbnail instead of entire sentences to remind them of each paragraph's content.

One participant who cited articles through bookmarking commented that he liked having that option because sometimes it was not immediately clear how best to fit the citation into the current frame. In other words, the preference for in-place citation over a separation of citation collection and citation linking may depend on whether the citation can be readily incorporated into the current proposal frame. Thus, a system needs to provide both options.

7.6 Revisit the four challenges

Here we summarize anecdotal evidence that suggests ThoughtFlow can help address some of the four challenges outlined in Sec. 3.2.

- **Context switch**: Analysis of the transition patterns between *Write* and *Explore* views suggests that some participants experienced less context switching with ThoughtFlow. In particular, it seems that much of this reduction in context switch can be attributed to the use of the paper thumbnail, which is a reminder of the proposal's outline and anchors for placing newly discovered related work.
- **Evidence imbalance**: We observed no cases in which participants discovered evidence inconsistent with the original hypotheses, which could be partially attributed to the bias towards publishing positive findings in research. However, we saw multiple cases in which participants discovered new areas of related work (e.g. the individual-difference aspect in the case study), which is also an example of reframing.
- **Disciplinary barriers**: We observed a few examples in which the user found relevant publications outside familiar research areas using the topic model visualization. In one case, the user started with a paper about "progressive visualization" and, from the same topic cluster, found a paper on "item sampling and information structure" from a human-computer interaction journal with which he was unfamiliar with. The user later referenced this paper in his proposal, since it describes a sampling scheme to achieve balanced coverage of an information space that he could potentially adapt for a query scheme to power progressive visualizations, strengthening the feasibility of his proposed work. This paper would be unlikely to discover through keyword-based search, but its relationship with progressive visualization was captured in a topic cluster with terms such as "sampling", "hierarchy", and "navigation" that are common to working with an information space with hierarchical organization and sampling.
- **Citation-based clusters**: As described in the case study, some participants explicitly used the citation sparklines to look for and find references outside the circle of publications already cited, suggesting that the citation sparklines could motivate some proposal writers to extend the coverage of citations.

8 Discussions

8.1 Design Implications

8.1.1 Effectiveness of a data-exploration approach may depend on the sensemaking state

In the second user study, we found that all participants used both keyword-based search and topic-based exploration to locate related work. However, preferences for the two types of approaches tended to shift as the proposal became more developed, and topic-based exploration was

generally more useful when an initial frame of the proposal had been established. This observation suggests that topic-based exploration and keyword-based search are complementary and suitable for different ideation stages. Thus, it is beneficial to support both types of information foraging in visual analysis tools designed for open-ended sensemaking tasks with text data, which usually involve gradually developing knowledge and can benefit from unexpected findings that arise from recognizing connections among terms and concepts. This observation also has an implication for evaluation. Since the effectiveness of topic-based exploration seems to depend on the completeness of the user's frame about the problem space, it might be under- or over-estimated in an evaluation that uses tasks localized to certain sensemaking stages. Evaluation of such exploration methods should consider all sensemaking stages in which such methods could be deployed.

8.1.2 Augmenting Outputs, not Inputs

Supporting multiple information navigation methods, e.g. term-based search and paper keyword search, often leads to increased interface complexity for both input controls and output display. This tradeoff between information-foraging flexibility and interface simplicity is common in many visual analysis applications. Results from this study suggest that, in such situations, the interface may be more intuitive if the input controls remain as simple as possible while the outputs are augmented. In this study, it was unnecessarily burdensome to ask the user to recognize and explicitly distinguish two types of search available in the system. Both types of search can be performed using the same type of user input (a query string), and the distinctions between the two need be emphasized only when presenting the search results to the user. Thus, the interface can provide a unified search bar that accepts a string, and update both the term-topic visualization and paper-search results accordingly. In general, visualization designers might consider consolidating multiple input controls that take similar parameters (e.g. strings or geographical regions) while displaying heterogeneous search results modularly.

8.1.3 Support framing with visualization

Unlike information foraging and elaboration, framing activities are usually purely mental and do not map onto observable user actions. However, we have identified some gaps related to framing and ways in which a visual-analysis tool can help bridge the gaps.

First, as discussed in Sec. 7.6, the paper thumbnail seems to help reduce context switch by capturing the proposal outline during exploration. More generally, this suggests that the availability of persistent visual structures that capture user-constructed data frames may help reduce context switch during data exploration, especially in systems that support a range of analysis activities and thus inevitably require view switches.

Second, research on confirmation bias and creativity suggests that users might **not question established frames** and look for evidence inconsistent with the current frames as actively as they should. With ThoughtFlow, we observed instances where the use of citation sparkline and topic-cluster exploration led to unexpected discovery of related work or of more concepts connected to the core concepts in some of the final proposals, as well as more bookmarked evidence that could stimulate reframing or creation of new frames. This suggests that supporting data exploration and recommendation based on similarity metrics that might introduce more data inconsistent with the user's initial frames can potentially lead to serendipitous insights.

8.2 Concept-Map Analysis for Insight-based Evaluation

Our experience with concept-map analysis in this work suggests that such analysis can be useful for insight-based evaluation, allowing evaluators to summarize and compare how individual insights together form the user's task-relevant knowledge. Meanwhile, we have also identified a few limitations in how we applied the analysis.

First, conventional concept-map analysis does not distinguish among types of concepts and relationships. When applying it in insight-based evaluation, however, it may be worthwhile to further categorize concepts and relationships and assign weights accordingly. In our case, we have identified different types of relationships such as "examples" and "analogies". Depending on the analysis domain, some relationships may be considered more insightful than others. Second, concept-map analysis by default does not capture concept qualifiers. For example, "PBWM model is a complicated, yet intuitive model of working memory" is represented in the same way as "PBWM is a model of working memory". Certain qualifiers may represent insightful observations of the data and could be worth capturing.

The concept-map analysis can also be augmented to capture the temporal evolution of the insights. A basic approach would be to track when each concept and relationship is added to the map and compute metrics that characterize the map's temporal properties, e.g. growth rate.

8.3 Open Questions and Challenges

8.3.1 Quantifying benefits of topic-based exploration

With the small sample sizes in these studies, it is difficult to draw conclusions about whether and when ThoughtFlow can provide a better proposal-writing environment than the standard approach, though our observations and user feedback suggest that the effectiveness of the tool might be mediated by individual cognitive traits, such as plasticity and willingness to adapt a current mental model to information available through the system. More controlled experiments with larger sample sizes could be conducted with fixed proposal topics to quantify differences in the depth and breadth of proposals created using ThoughtFlow versus conventional academic search engines while measuring cognitive traits and the user's initial mental model of the problem space through pre-study interviews. In addition, the inclusion of various visual components like citation sparklines and paper thumbnails can also be controlled in future experiments to isolate and measure the influence of these components on proposal properties. Controlled experiments can also be used further to quantify to what extent citation sparklines increase the rate of examining recommended citations and considering more disconfirming evidence. We can also test whether a preference for in-place

citation indeed correlates with consistency between the evidence and the existing frame. If such correlations exist, it provides an opportunity to automatically infer framing activities from use of components like the paper thumbnail.

8.3.2 Topic quality and similarity measures

The studies in this paper demonstrate the value of surfacing topic-based document similarity through visualizations during proposal idea generation and information foraging processes, but leave open many questions related to topic quality or alternative approaches to identifying semantically relevant publications. For example, we did not optimize topic model parameters (e.g. the number of topics) or measure the perceived quality of the topic models. More controlled experiments could be conducted to further investigate how each user's perception of topic quality and relevance mediates such factors as trust and information fatigue to affect the utility of document similarity information during proposal writing. In addition, there are a few families of techniques for capturing text semantic similarity (e.g. word embedding [41]). It is worth investigating how these different characterizations of text similarity affect the research idea generation process. For instance, word-embedding techniques usually base similarity on shared local context and vector-space distances and are not designed to capture higher-level interactions among words such as document-level co-occurrence or asymmetric similarity relations [42]. Thus, the use of word embedding could increase keyword-level relevance of recommended references, but the result similarity ranking may have more overlap with citation-based similarity and promote fewer serendipitous insights during exploration. In the end, the effectiveness of a similarity measure may well depend on which sensemaking task the similar content is meant to support.

8.3.3 Serving more experienced users

This study focuses on how users interact with ThoughtFlow to develop proposals on relatively new topics, leaving open the question of how researchers with more experience and familiarity with a topic might use and benefit from the tool. Anecdotal feedback from an initial interview with a faculty member suggests that senior researchers may be more inclined to develop a complete proposal frame based on their prior knowledge of a topic. Reference recommendations are, in this case, more likely to lead to incremental changes to a proposal instead of more actively affecting the proposal's framing, as was observed in this study. This implies that for a researcher with more experience or working on familiar topics, the proposal-writing process may be best supported by a different type of tool.

8.3.4 Other potential applications of citation sparklines

User feedback and use of the citation sparklines suggest that they can enhance user engagement with recommended citations. The sparklines could also be used in other situations where citations are displayed, e.g., beside texts of a publication. However, that the sparkline draws the user's attention to information inherent in the citation instead of to its relevance to the citing article could be either an advantage or source of distraction. More controlled experiments are needed to further quantify the effects of showing citation sparklines.

9 Conclusion

We have presented results from an empirical study of supporting research-proposal writing – a sensemaking activity that involves generating and framing research ideas and supporting the arguments with evidence from the literature – using a visual analysis system. Through two design-evaluation iterations, we have identified the following ways in which visual analysis tool designs can facilitate both framing and elaboration in proposal writing:

- Provide a combination of conventional paper-search and topic-based exploration to keep users grounded in familiar information-foraging processes while also stimulating serendipitous insights
- Recommend references accompanied by citation sparklines to facilitate discovery of related work based on both semantic similarity and citation properties
- Use a paper thumbnail to enable in-place citation and provide an anchor for the current frame to reduce context switch

Our results show that by integrating proposal writing and related work search and providing visual components that target challenges arising during specific sensemaking activities, ThoughtFlow enabled users to develop more elaborated proposals, as assessed by concept-map analysis, than when using academic search engines.

Acknowledgments

References

[1] 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*. ACM, 2006, pp. 801–810.

[2] N. Elmqvist and P. Tsigas, "Citewiz: a tool for the visualization of scientific citation networks," *Information Visualization*, vol. 6, no. 3, pp. 215–232, 2007.

[3] P. Bergström and D. C. Atkinson, "Augmenting the exploration of digital libraries with web-based visualizations," in *Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on*. IEEE, 2009, pp. 1–7.

[4] M. Dörk, N. H. Riche, G. Ramos, and S. Dumais, "Pivotpaths: Strolling through faceted information spaces," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2709–2718, 2012.

[5] F. Heimerl, Q. Han, and S. Koch, "Citerivers: visual analytics of citation patterns," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, no. 1, pp. 190–199, 2016.

[6] F. Beck, S. Koch, and D. Weiskopf, "Visual analysis and dissemination of scientific literature collections with survis," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, no. 1, pp. 180–189, 2016.

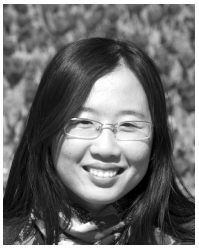
[7] B. Cronin, "Tiered citation and measures of document similarity," *Journal of the American Society for Information Science*, vol. 45, no. 7, pp. 537–538, 1994.

[8] J. Nicolaisen, "Citation analysis," *Annual review of information science and technology*, vol. 41, no. 1, pp. 609–641, 2007.

[9] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. J. Gao, H. Qu, and X. Tong, "Textflow: Towards better understanding of evolving topics in text," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2412–2421, 2011.

[10] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian, "Tiara: Interactive, topic-based visual text summarization and analysis," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, p. 25, 2012.

[11] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "Paralleltopics: A probabilistic approach to exploring document collections," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011, pp. 231–240.
[12] M. Drk, D. M. Gruen, C. Williamson, and M. S. T. Carpendale, "A visual backchannel for large-scale events." *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1129–1138, 2010.
[13] S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei, "Online visual analytics of text streams," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 11, pp. 2451–2466, 2016.
[14] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: a full picture of relevant topics," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE, 2014, pp. 183–192.
[15] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 2012, pp. 74–77.
[16] B. Gretarsson, J. Odonovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth, "Topicnets: Visual analysis of large text corpora with topic modeling," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, p. 23, 2012.
[17] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 1992–2001, 2013.
[18] Y. B. Shrinivasan and J. J. van Wijk, "Supporting the analytical reasoning process in information visualization," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2008, pp. 1237–1246.
[19] D. Gotz and M. X. Zhou, "Characterizing users' visual analytic activity for insight provenance," *Information Visualization*, vol. 8, no. 1, pp. 42–55, 2009.
[20] T. M. Green, W. Ribarsky, and B. Fisher, "Building and applying a human cognition model for visual analytics," *Information visualization*, vol. 8, no. 1, pp. 1–13, 2009.
[21] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card, "The cost structure of sensemaking," in *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, 1993, pp. 269–276.
[22] 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, pp. 2–4.
[23] G. Klein, B. Moon, and R. R. Hoffman, "Making sense of sensemaking 2: A macrocognitive model," *Intelligent Systems, IEEE*, vol. 21, no. 5, pp. 88–92, 2006.
[24] P. Zhang and D. Soergel, "Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking," *Journal of the Association for Information Science and Technology*, vol. 65, no. 9, pp. 1733–1756, 2014.
[25] J. S. Yi, Y. ah Kang, J. T. Stasko, and J. A. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 6, pp. 1224–1231, 2007.
[26] M. E. Gorman, "Error, falsification and scientific inference: An experimental investigation," *The Quarterly Journal of Experimental Psychology*, vol. 41, no. 2, pp. 385–412, 1989.
[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
[28] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *KDD'08*, 2008, pp. 990–998.
[29] N. Most. metapub. [Online]. Available: https://pypi.python.org/pypi/metapub/0.3.15
[30] J. D. Novak and A. J. Cañas, "The theory underlying concept maps and how to construct and use them," 2008.
[31] P. Saraiya, C. North, and K. Duca, "An insight-based methodology for evaluating bioinformatics visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 4, pp. 443–456, Jul. 2005. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2005.53
[32] L. Wilkinson, "Statistical methods in psychology journals: Guidelines and explanations." *American psychologist*, vol. 54, no. 8, p. 594, 1999.
[33] J. Cohen, "Statistical power analysis for the behavioral sciences . hilsdale," *NJ: Lawrence Earlbaum Associates*, vol. 2, 1988.
[34] G. M. Sullivan and R. Feinn, "Using effect sizeor why the p value is not enough," *Journal of graduate medical education*, vol. 4, no. 3, pp. 279–282, 2012.
[35] P. D. Ellis, *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press, 2010.
[36] F. Müller-Wienbergen, O. Müller, S. Seidel, and J. Becker, "Leaving the beaten tracks in creative work-a design theory for systems that support convergent and divergent thinking," *Journal of the Association for Information Systems*, vol. 12, no. 11, p. 714, 2011.
[37] R. G. Bias, B. M. Moon, and R. R. Hoffman, "Concept mapping usability evaluation: An exploratory study of a new usability inspection method," *International Journal of Human-Computer Interaction*, vol. 31, no. 9, pp. 571–583, 2015.
[38] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw, "A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 51–60, 2016.
[39] Z. Liu and J. Heer, "The effects of interactive latency on exploratory visual analysis," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2122–2131, 2014.
[40] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading tea leaves: How humans interpret topic models." in *Nips*, vol. 31, 2009, pp. 1–9.
[41] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010, pp. 384–394.
[42] A. Nematzadeh, S. C. Meylan, and T. L. Griffiths, "Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words," in *Proceedings of the 39th annual meeting of the cognitive science society*, 2017, pp. 859–864.



Hua Guo received her PhD degree in computer science from Brown University. She is a Data Visualization Scientist at Twitter. Her research interest is in informing and automating visualization design and evaluation using human-centered approaches.

David H. Laidlaw received the PhD degree in computer science from the California Institute of Technology, where he also did post-doctoral work in the Division of Biology. He is a professor in the Computer Science Department at Brown University. His research centers on applications of visualization, modeling, computer graphics, and computer science to other scientific disciplines. He is a fellow of the IEEE and the IEEE Computer Society and recipient of the 2008 IEEE VGTC Visualization Technical Achievement Award.