

GRAFS: Graphical Faceted Search System to Support Conceptual Understanding in Exploratory Search

MENGTIAN GUO, University of North Carolina at Chapel Hill, USA

ZHILAN ZHOU, University of North Carolina at Chapel Hill, USA

DAVID GOTZ, University of North Carolina at Chapel Hill, USA

YUE WANG, University of North Carolina at Chapel Hill, USA

When people search for information about a new topic within large document collections, they implicitly construct a mental model of the unfamiliar information space to represent what they currently know and guide their exploration into the unknown. Building this mental model can be challenging as it requires not only finding relevant documents, but also synthesizing important concepts and the relationships that connect those concepts both within and across documents. This paper describes a novel interactive approach designed to help users construct a mental model of an unfamiliar information space during exploratory search. We propose a new semantic search system to organize and visualize important concepts and their relations for a set of search results. A user study ($n = 20$) was conducted to compare the proposed approach against a baseline faceted search system on exploratory literature search tasks. Experimental results show that the proposed approach is more effective in helping users recognize relationships between key concepts, leading to a more sophisticated understanding of the search topic while maintaining similar functionality and usability as a faceted search system.

ACM Reference Format:

Mengtian Guo, Zhilan Zhou, David Gotz, and Yue Wang. 2022. GRAFS: Graphical Faceted Search System to Support Conceptual Understanding in Exploratory Search. *ACM Trans. Interact. Intell. Syst.* 37, 4, Article 111 (August 2022), 37 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Exploratory search tasks are common among inquisitive users. Students explore scientific literature to gain knowledge about a topic. Health professionals synthesize medical literature to systematically assess treatments and associated outcomes for a disease. Journalists analyze news articles to link separate events into a coherent story. Intelligence analysts examine case reports to connect disparate evidence that suggests a potential threat. In all these tasks, searchers start with a complex information problem, yet a lack of understanding of the information space [53]. Such an understanding, also called a “mental model” or “schema” in sensemaking literature [34], may include key concepts or aspects that are important to the topic under investigation, and a rough understanding of how these concepts may relate to one another in the context of the topic. The mental model will co-evolve with the exploratory search process: it is updated as more information is encountered during search, and the perceived incompleteness of the model inspires further search activities [3].

Authors’ addresses: Mengtian Guo, mtguo@email.unc.edu, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; Zhilan Zhou, ztl@live.unc.edu, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; David Gotz, gotz@unc.edu, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; Yue Wang, wangyue@unc.edu, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

2160-6455/2022/8-ART111 \$15.00

<https://doi.org/10.1145/1122445.1122456>

Since such a mental model plays a crucial role in exploratory search, it is desirable for systems to assist users in constructing a mental model at the beginning of an exploratory search task.

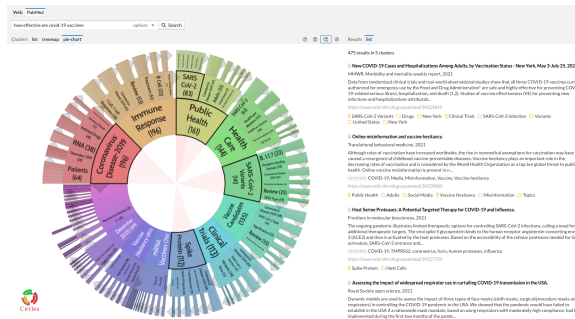
Existing search engines offer varying degrees of support for exploratory searchers to build a mental model of the information space. The classical list presentation (e.g., Google) provides minimal support (if any) as the list of result snippets may or may not contain important concepts or their relations. Document clustering search engines aim to group search results into clusters and automatically summarize clusters using keywords (Figure 1a). They attempt to organize search results into conceptual groups, which may help users build a mental model. However, algorithm-generated document clusters and cluster-labels are not guaranteed to be meaningful even in the eyes of experts. The potential unintelligibility adds to the burden of a user who is exposed to an unfamiliar information space for the first time.

As semantically annotated documents become available through manual or automated annotation approaches, search engines can begin to surface meaningful semantic concepts to aid exploratory searchers. For example, a facet panel groups semantic concepts into different categories, i.e., facets (Figure 1b). It provides the user with an overview of important concepts in the search results, as well as the capability to filter results by specific concepts. However, a facet panel cannot effectively communicate how these concepts *relate* to one another. It only implicitly reveals the connections between a selected concept and other unselected concepts if it shows the volume of search results associated with each concept (e.g., bars of different lengths in the facet panel in Figure 1b).

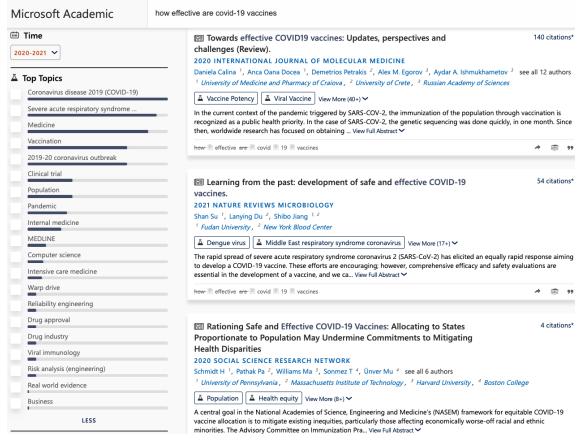
Other search interfaces visualize concepts in a 2D layout, using spatial proximity as a metaphor to communicate conceptual relationships (Figure 1c). This works well if the visualized concepts are few and their relationships are sparse. However, in many search scenarios, the results may contain a large number of concepts that are densely related to each other. In those scenarios, the visualization tends to clutter the interface with many concepts and relations, while also inevitably losing information due to the projection of the complex network of concepts into a two-dimensional representation.

In our work, we envision an intelligent search system that can help exploratory searchers discover key concepts and conceptual relationships in search results through a simple, minimally cluttered interface. This is a meaningful and challenging problem. On the one hand, seeing how key concepts interplay with each other helps the searcher build a more complete mental model of the information space, which can enable more informed and fruitful investigation. On the other hand, the number of relationships between key concepts can be too large, and presenting them may easily overwhelm the user who just started the exploration. An ideal system would start by revealing just enough key concepts and relations needed to construct an initial mental model, and progressively nudge the searcher towards an increasingly nuanced understanding of the information space over the course of search. In this paper, we begin to address this problem through three key contributions.

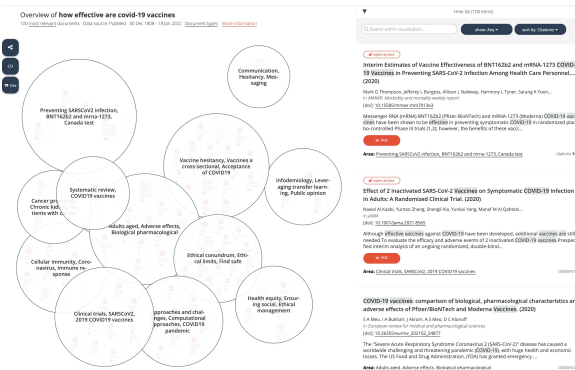
- First, we design a new interface to communicate key aspects of a computationally maintained knowledge subgraph to users during interactive search tasks. The design objectives include: (1) reveal relationships between concepts without cluttering the interface; (2) leverage user's familiarity with existing search interface elements such as a main result listing and a facet panel; and (3) support a mixed-initiative approach where users can make changes to the subgraph as needed. This results in the Graphical Faceted Search system, or GRAFS. In GRAFS, a knowledge subgraph is embedded in a familiar faceted search-like interface to help searchers construct a basic understanding of the information space and encourage them to dig deeper.
- Second, an exploratory searcher's need for a mental model inspires us to formulate a new data model – an initial knowledge subgraph for exploratory search. The goal of this data



(a) Carrot2



(b) Microsoft Academic



(c) Open Knowledge Map

Fig. 1. Examples of current search systems that show semantic annotations as different interface elements. (a) Carrot2 [33] infer a hierarchical cluster structure over search results and visualize the hierarchy in a pie-chart or treemap. (b) Microsoft Academic [50] annotates each article with research topics and uses them as faceted filters on the left side of the results list. (c) Open Knowledge Maps [21] organizes results into semantic clusters and uses spatial layout to show semantic relationship between clusters. Accessed in January 2022.

model is to initiate the construction of a mental model in the searcher's mind. We propose that the subgraph shall meet the following criteria: (1) be relevant to the search context; (2) have wide coverage of the information space to stimulate learning; and (3) be efficiently computable to support real-time interaction. We propose an efficient algorithm for extracting such a knowledge subgraph given semantic search results.

- Third, we provide results from a user study conducted to evaluate the proposed search interface in the context of medical literature exploration tasks. Compared to a classical faceted search interface, GRAFS helps users better identify relations between concepts and develop a deeper understanding of search topics. At the same time, GRAFS maintains the original functionality of a faceted search system with which users are familiar.

2 RELATED WORK

The proposed search system aims to assist users in exploratory search tasks. Its frontend interface leverages faceted filters, result clustering, and information visualization. Its backend data model computes and maintains a knowledge subgraph from semantic search results and user interaction. As such, this work is related to multiple lines of previous work in interactive and intelligent systems as we discuss below.

2.1 Exploratory Search

Exploratory search happens when a user has a complex information problem but insufficient knowledge to clearly express the information need [53]. Answers to such problems often cannot be found in any one document, but instead has to be synthesized over many documents retrieved by a series of revised queries. Exploratory searchers alternate between two modes: learning and investigation [27]. In the learning mode, a searcher tries to make sense of important aspects encountered in the search results and how these aspects relate to one another, i.e., construct a mental model of the information space. Realizing missing pieces in the current mental model, a searcher will enter the investigation mode to collect more information to fill in the gap, leading to the next round of learning. Common web search engines, such as Google and Bing, are primarily designed for fact retrieval, with limited explicit support for learning activities. More elaborate interactive features have been proposed to support exploratory search [52], including facet filters [24, 56], result clustering [7, 57], and information visualization [12, 17]. Our goal in this work is to design search systems that provide better support for exploratory search, especially for the discovery of key concepts and their relations in the learning stage.

2.2 Faceted Filtering and Search Result Clustering

Faceted search systems are commonly used in domains where documents are associated with rich metadata. These search systems group different dimensions of metadata into facets which allow a user to slice and dice search results along different facets and facet-values [46]. The list of relevant facet-values, sometimes each annotated with a corresponding number of associated results, naturally provides an overview of the information space. These powerful capabilities make faceted search systems amenable to exploratory tasks [53]. Faceted search systems have successful applications in digital libraries [14] and e-commerce websites [30], where rich metadata have traditionally been manually assigned to each document. With the help of supervised machine learning and natural language processing techniques, metadata from a pre-defined knowledge graph or ontology can be automatically assigned to documents [39].

Search result clustering systems also aim to organize a large collection of results into subgroups [7]. Each subgroup is often assigned a word or phrase label generated by the clustering

algorithm. This approach can be applied to arbitrary textual search results without manual annotation efforts, as cluster labels and document groupings are automatically generated by text clustering algorithms. However, algorithm-generated clusters and cluster labels can sometimes be difficult to interpret and can lead to user confusion [16].

The approach outlined in this paper draws on the strengths of both faceted search and clustering. We augment the traditional facet panel with relationship links that are interactively surfaced between semantic concepts (facet-values). Moreover, we cluster key concepts such that semantically related concepts are represented closely in the facet panel, nudging users to see connections among related (and therefore nearby) concepts.

2.3 Text Search Result Visualization

Information visualization provides powerful approaches for users to view, search, manipulate, and reason about complex textual information through graphical representations. Information retrieval systems have a long history in employing visualizations to help users obtain an overview of search results [15]. A common approach is to project documents, keywords, and concepts as objects on a two-dimensional canvas or nodes in a network, such that spatially close objects are semantically related [5, 12, 19, 21, 31, 49]. When additional document metadata are available, search results can also be visualized as with additional organization such as topical facets [50], timelines [25], geographic maps [44], and knowledge graphs [37]. However, visualizing nominal, high-dimensional textual information in a low-dimensional space with limited screen resolution will necessarily lead to loss of information. Therefore, to make full sense of visualized search results, users often still need to read associated documents [15].

In the approach outlined in this paper, we visualize not retrieved documents but important concepts and relations in retrieved documents, which constitute a knowledge graph. Recognizing that an exploratory searcher is typically not familiar with all concepts and their relations until after reading associated texts, we keep the visualization simple and progressive. This approach introduces additional detail as an exploratory search task unfolds as opposed to taking up large screen area or visualizing all details of the subgraph upfront.

2.4 Knowledge Graph in Search Systems

Knowledge graphs, or semantic networks, play important roles in search systems. Semantic search technology relies on accurate recognition of concepts in queries and documents to achieve backend capabilities such as query intent understanding [4], automatic query expansion [10], improved result ranking [55], entity retrieval [2], and direct answer retrieval [43]. In the frontend, semantic concepts afford new interactive browsing features in addition to document listings. These include per-result semantic tags [51], concept-based filtering [42, 50], faceted navigation [18, 42], and conversational search [48].

The work introduced in this paper focuses on enabling user interactions with a query-specific knowledge graph during exploratory search. Previous work in natural language processing has proposed various approaches to constructing a query-specific knowledge graph, with the primary goal of improving relevance ranking of documents or entities [9, 11, 13, 38]. Although we also construct a query-specific knowledge graph in this work, the primary goal is for front-end presentation as a means to assist exploratory searchers whose goal is not necessarily to identify the most relevant documents or entities, but rather to make sense of an unfamiliar topic and corresponding search results with varying degrees of relevance. Previous search interfaces either present a set of relevant concepts without surfacing their relations [42, 50], or use a large screen area to visualize concepts and relations (in which the main document listing area is substantially reduced, deviating from the familiar traditional search interface) [18, 21, 37]. In contrast, our design objective is to

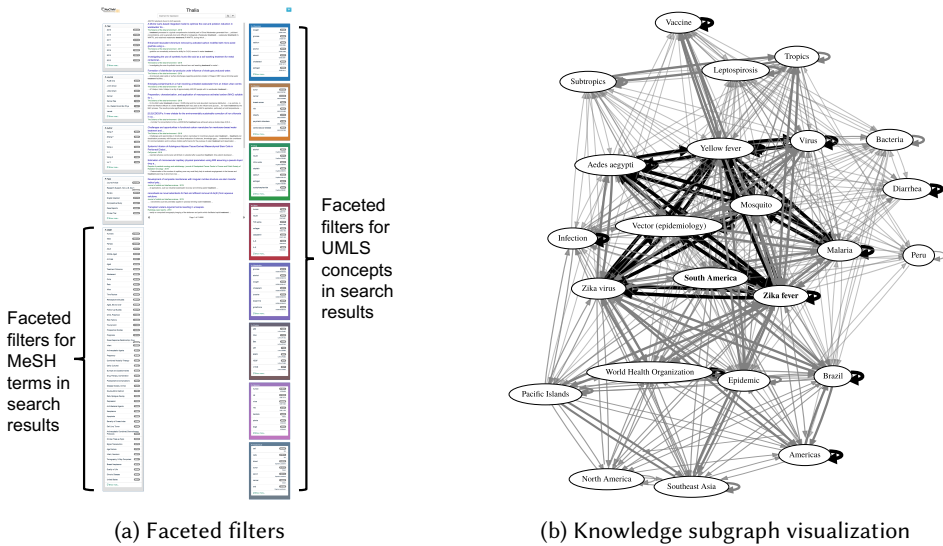


Fig. 2. Challenges in presenting concepts and relations in search results. (a) A screenshot of Thalia search engine’s result page for the query “treatment for depression” over PubMed abstracts [42]. Hundreds of MeSH terms and UMLS concepts mentioned in the search results are used as faceted filters. This leads to exceedingly long lists of faceted filters on two sides of the interface. (b) A visualization of the knowledge subgraph constructed for the query “Zika fever” on English Wikipedia (reproduced with author’s permission) [11]. Each node is a DBpedia concept mentioned in the search results. Two concepts are connected by an edge if they are mentioned in the same context in any search result. Showing all relationship links is visually overwhelming even after pruning the number of concepts down to two dozen.

show relations between concepts while still maintaining key aspects of the traditional and effective document listing design common to most users’ search experiences.

3 GRAPHICAL FACETED SEARCH

In this section, we describe the design and implementation of the graphical faceted search (GRAFS) system. We start by motivating our research problem. We then divide the problem into two parts: a data representation problem and an interactive presentation problem. In each part, we formulate the problem, discuss the general solution framework, and describe our specific implementation of the solution.

3.1 Challenge and Motivation

An exploratory searcher needs to obtain an overview of a large number of potentially relevant documents returned by her query. Instead of reading these documents one by one trying to build a conceptual understanding of the information space, a better strategy is to learn key concepts and relations that show up frequently throughout the search results. In fact, semantic search engines also generate and leverage such information in their backend [35].

A semantic search engine applies entity recognition and linking algorithms to documents (at index time) and queries (at search time). This has become increasingly common thanks to the advances in natural language processing (NLP) and the availability of knowledge graphs (also called standardized vocabularies or ontologies) in different domains. As a result, when a search engine returns a list of documents, it can simultaneously return a set of entities (and their relations) recognized in

these documents. In other words, a semantic search not only returns a set of documents, but also a set of semantic units, a subgraph inside the global knowledge graph. On the one hand, this knowledge subgraph has great potential to help an exploratory searcher obtain a mental model of the information space in question. On the other hand, this knowledge subgraph can be enormous if it is shown in its raw form. It may consist of not only core concepts of the information space but also a large number of peripheral and even non-relevant concepts that happen to be mentioned in the result documents. Therefore, **it is challenging to allow exploratory searchers make use of a query-specific knowledge subgraph in a non-overwhelming manner.**

To illustrate this challenge, consider NaCTeM's Thalia search system for PubMed abstracts [42]. The system uses NLP algorithms to extract UMLS concepts and MeSH terms from abstracts and use them as faceted filters on the result page. As shown in Figure 2a, the query "treatment for depression" retrieves an exceedingly long list of concepts from search results, which resulted in faceted filters vertically spanning multiple screens! Even worse, top-ranked concepts in each semantic category (facet) are most frequent yet non-informative, such as *Human* in *Gene/Species*, *Cell* in *Anatomical Parts*, and *Oxygen* in the category *Chemical*.

To reduce the complexity of the query-specific knowledge subgraph, researchers have proposed selection and ranking algorithms aiming to preserve only a set of core concepts [9, 11, 38]. However, even after reducing the space of concepts, the number of possible relationships among concepts can still be large. Figure 2b shows a knowledge subgraph produced by a state-of-the-art approach when searching Wikipedia articles using the query "Zika fever" [11]. Despite having only 24 concepts, the subgraph is densely connected due to intricate relationships between these concepts in search results. Showing all relationship links upfront can easily overwhelm a user's mind.

The above challenge directly motivates our work. Below we describe our approach to this challenge, including interactive designs for presenting a knowledge subgraph for exploratory search, as well as computational methods for constructing and maintaining the underlying data model. As a specific use case, we construct and present knowledge subgraphs to medical literature searchers in an uncluttered and non-overwhelming manner.

3.2 Interaction Design and Data Model

In this section, we first present our interaction design goals for GRAFS. We then design the underlying backend data model, i.e., a knowledge subgraph, that supports the frontend design goals.

3.2.1 Interaction Design Goals. Informed by the challenges of exploratory search described above, we propose the following design goals (DGs) for the interactive experience to be fulfilled by GRAFS.

DG1 Preserve the simplicity, familiarity, and capability of a search interface. Search engine users are well educated to use the Google-like list presentation and faceted navigation interface. We aim to maintain the usability of the new search interface by leveraging the basic layout and functionality of existing systems.

DG2 Show relations between concepts without cluttering the interface. Visualizing relations between concepts have great potential to help users explore the information space. However, high-density connections can easily clutter the interface and discourage learning. We aim to selectively expose these relations in a sparse, non-cluttering manner.

DG3 Preserve human agency. The visualization of concepts and relations, i.e., a knowledge subgraph, is meant to initiate the user's learning activities. Users should have the ability to make adjustments to the knowledge subgraph based on their mental model or current interest during the sensemaking process.

3.2.2 Data Model Formulation. The above interaction design goals imply a data model that supports user interaction. The data model is a small-scale knowledge graph that contains the most informative concepts and relations embedded in a large number of search results. The goal of this data model is to inspire users to learn about these concepts and relations and form their own mental model of the information space, i.e., a parsimonious and evolving understanding of key concepts in the space and how they relate to one another. The data model aims to lead the user in constructing their mental model, but it is not the mental model itself. Below we formulate this data model.

A semantic search engine returns not only a ranked list of results for a given query but also concepts (or entities) and their relations within those results. Formally, an initial search query q retrieves a list of documents $D_q = \{d_i\}_{i=1}^n$. Within each document d_i , the search system also recognizes a set of concepts $C_i \subset C$ and a set of relations $R_i \subset R$.¹ Here, C is the set of all possible concepts in the global knowledge graph, and R is the set of all possible relations between those concepts. Therefore, in addition to the list of documents D_q , the initial search query q effectively retrieves a query-specific knowledge graph $G_q = (C_q, R_q)$, where $C_q = \bigcup_{i=1}^n C_i$ and $R_q = \bigcup_{i=1}^n R_i$. We assume that concepts and relations in G_q have been pre-extracted by NLP algorithms in the search engine backend [9, 35, 54]. As described in Section 3.1, G_q often contains a large number (hundreds or even thousands) of concepts and relations. Directly presenting the entire graph G_q to users will inevitably cause information overload.

To support the above interaction design goals, we need to construct an **initial knowledge subgraph** $H_q \subset G_q$ that is most helpful at the initial stage of mental model construction. Just as a traditional search interface tackles information overload by showing a small number (e.g., 10) of the most relevant documents on the first page, we posit that the GRAFS exploratory search interface will encourage learning and navigation by showing a small initial subgraph of the most important concepts and relations. As the exploration unfolds, the user will gradually move beyond this initial subgraph.

How to construct such a knowledge subgraph? If our goal is to find a subgraph with k concepts to be shown to the user (where k is small), then it translates into a computational problem of selecting the best k -sized subgraph H_q out of the original graph G_q . Guided by the interactive design goals **DG2** and **DG3**, we propose the following selection criteria (SC):

- SC1 Relevance.** Concepts and relations in this subgraph should be centered around the user’s search interest, as opposed to drifting into peripheral parts of the information space.
- SC2 Coverage.** Concepts and relations in this subgraph should cover diverse subregions of the information space, as exploratory search aims at breadth and learning.
- SC3 Efficiency.** The subgraph selection and update procedures should be responsive enough to support a smooth search experience [1].

We note that these selection criteria are related to search result diversification [36], where the goal is to provide result documents to cover different aspects of an ambiguous query. In the learning stage of an exploratory search, the user’s goal is also ambiguous. The crucial difference is that here our goal is to select or rank knowledge graph concepts instead of result documents.

Even after being pruned to a manageable size, a knowledge graph can still be too abstract to make immediate sense in a user’s eyes. To further support user interpretation of the extracted knowledge subgraph H_q , we augment it with the following data elements to be used in the frontend interface.

¹We use standard set-theoretic notation in this paper. “ $A \cup B$ ” means the union of sets A and B . “ $A \cap B$ ” is the intersection of sets A and B . “ $e \in A$ ” means element e is a member of set A . “ $A \subset B$ ” means A is a subset of B . “ $A \setminus B$ ” means the subset of A that is not in B . “ \emptyset ” means the empty set. “ $\bigcup_{i=1}^n C_i$ ” stands for the union of n sets C_1, C_2, \dots, C_n .

Concept Provenance. For each concept $c \in H_q$, we need a small representative set of context windows (e.g. sentences) in result documents D_q which mention the concept c and the query q . These context windows explain why a concept is relevant in the search context.

Graph Partitioning. The set of concepts in H_q should be partitioned such that concepts in the same partition are densely connected by edges in H_q , and concepts between partitions are loosely connected by edges in H_q . For semantically close concepts in the same partition, we present them visually close to each other and use the same color coding in the interface. This can help with the chunking process in learning [28] and nudge users to think about connections between nearby concepts according to the proximity principle in Gestalt Principles [45]. In other words, visualizing these partitions is an implicit way of surfacing relations between concepts (DG2).

3.3 Data Model and Interactive System Implementation

In this section, we describe the interactive system for delivering the experience that GRAFS aims to achieve and the computational implementation of the underlying data model. Here our description follows the data flow: we first describe the implementation of the data model, followed by the implementation of the interactive system that presents the data model to users.

3.3.1 Data Model Implementation. We implemented the described data model on top of a custom-built search engine for 32.6 million PubMed abstracts, which were downloaded from the National Library of Medicine website. We use an efficient maximal pattern matching-based algorithm [8, 20] to annotate clinical concepts mentioned within each abstract by looking up terms in the SNOMED-CT vocabulary. The initial knowledge subgraph is built specifically for each user-issued search query to ensure *relevance* (SC1). Given a search query q , we first extract the original query-specific knowledge graph $G_q = (C_q, R_q)$ as follows. All concepts mentioned in the set of retrieved documents by the search query D_q form the extracted concept set C_q . We take a generic view of concept relations and assume that a pair of concepts in C_q are related if they co-occur in any retrieved document. This generates a set of concept relations R_q . We do not further consider fine-grained types of relations because, in the context of a specific query, concepts can be related in novel and nuanced ways that are not documented in an ontology (e.g., SNOMED-CT), and accurately extracting such relations from text is a challenging natural language understanding task [26].

In principle, one can apply existing methods for ranking and selecting knowledge graph concepts for search tasks [9, 11, 38]. However, these methods are mainly optimized for batch evaluation settings and their computational cost can be too high to be run at an interactive rate. For example, shortest-path algorithms and random walks on the query-specific knowledge graph G_q have superlinear (e.g., quadratic) time complexity in the number of concepts in G_q . We adopt an efficient implementation instead.

We construct a subgraph $H_q \in G_q$ that contains a user-specified number (e.g., 20) of concepts in C_q . We incorporate the *relevance* (SC1) and *coverage* (SC2) criteria into an efficient subset selection algorithm (SC3). The algorithm is inspired by the maximal marginal relevance [6]. The basic idea is to sequentially select items that simultaneously have high relevance to the query and few relationship links between each other.

Formally, let q be the search query, $C_i \subset C_q$ be the current set of selected concepts with size i (initially $i = 0$, $C_i = \emptyset$), and $C_q \setminus C_i$ be the current set of unselected concepts. We select the next concept c_{i+1} as follows:

$$c_{i+1} \leftarrow \operatorname{argmax}_{c \in C_q \setminus C_i} \left[\lambda \cdot r(c, q) - (1 - \lambda) \cdot \max_{c_j \in C_i} s(c, c_j) \right]. \quad (1)$$

Here, c is a candidate concept in the current set of unselected concepts. $r(c, q) = |\{d|c \in d, d \in D_q\}|$ measures the *relevance* (SC1) of c by the number of retrieved documents containing c . $s(c, c_j) = |\{d|c \in d, c_j \in d, d \in D_q\}|$ measures the relationship strength between c and a selected concept $c_j \in C_i$ by the number of documents where c and c_j co-occur. Intuitively, the term “ $-\max_{c_j \in C_i} s(c, c_j)$ ” promotes *coverage* (SC2) of the next concept c by forcing it to be semantically far from the current selected set C_i . $0 < \lambda < 1$ defines the relative importance of relevance and coverage criteria. We set λ to a static value ($\lambda = 0.5$) during our study to put equal importance on relevance and coverage criteria. k concepts are added into H_q incrementally following Equation (1). The algorithm has a time complexity of $O(k^2|C_q|)$. The algorithm runs *efficiently* (SC3) since k , the number of concepts in the subgraph H_q , is usually small (e.g., 20), and the algorithm has linear (instead of quadratic) complexity in the number of concepts in G_q , which is often very large (e.g., 5000).

We noticed that some selected concepts are frequent but not informative. For instance, the concept “COVID-19” appears in more than half of the documents retrieved by a COVID-19 related query, and therefore has a large relevance term $r(c, q)$. However, such a concept contains little new information about the topic. We, therefore, omitted concepts with $r(c, q) > |D_q|/2$, treating them as equivalent to “stop words” (frequent functional words) in the context of the current search results D_q .

Concept Provenance. For each selected concept c in H_q , we extracted three representative sentences containing both concept c and query q to show how c is used in the context of search results. The selection algorithm is also inspired by maximal marginal relevance. First, all sentences containing c were extracted from D_q to form the candidate sentence set S_q . Let $S_i \in S_q$ be the current set of selected sentences with size i (initially $i = 0, S_q = \emptyset$). The algorithm selects the next sentence s_{i+1} as follows:

$$s_{i+1} \leftarrow \operatorname{argmax}_{s \in S_q \setminus S_i} \left[\lambda \cdot v(s, q) - (1 - \lambda) \cdot \max_{s_j \in S_i} v(s, s_j) \right]. \quad (2)$$

Here, $v(s, q)$ measures the relevance of sentence s with respect to query q , while $v(s, s_j)$ measures the similarity between the candidate sentence s and a selected sentence s_k . The function $v(\cdot, \cdot)$ represents queries and sentences as bag-of-words vectors with TFIDF weights and computes cosine similarity between two vectors. We empirically set $\lambda = 0.5$ to balance the relevance and diversity of the set of selected sentences.

Graph Partitioning. This step generates groups of concepts in H_q based on their relations. Instead of choosing a fixed number of partitions (as in algorithms like k -means or k -medoids), we performed agglomerative hierarchical clustering over the concepts in H_q . First, each concept starts in its own cluster. At each subsequent step, the two closest clusters are merged. The distance between two clusters is the distance between the farthest concepts (complete-linkage clustering). The distance between two concepts is calculated as the symmetric difference between the two sets of documents that contain each concept: $d(c_i, c_j) = |D_i \cup D_j| - |D_i \cap D_j|$, where $D_i = \{d|c_i \in d, d \in D_q\}$. $d(c_i, c_j) = 0$ when $D_i = D_j$, i.e., when c_i and c_j always co-occur in retrieved documents and are therefore very close semantically. Therefore, even if two concepts c_i, c_j frequently co-occur ($|D_i \cap D_j|$ is large), their distance is still far if their union is much larger than their intersection. This avoids the problem where a concept is viewed to be close to many other concepts simply because it appears in many documents (e.g., concepts in the query that are prevalent in search results).

The hierarchical clustering algorithm produced a tree structure where each leaf corresponds to a concept. To generate partitions, we cut the tree such that each sub-tree is as large as possible but has no more than one-third of all concepts (leaf nodes) in the original tree. This method is able to produce partitions with relatively balanced sizes regardless of the original tree structure. Each partition contains semantically related concepts. This approach is different from the commonly

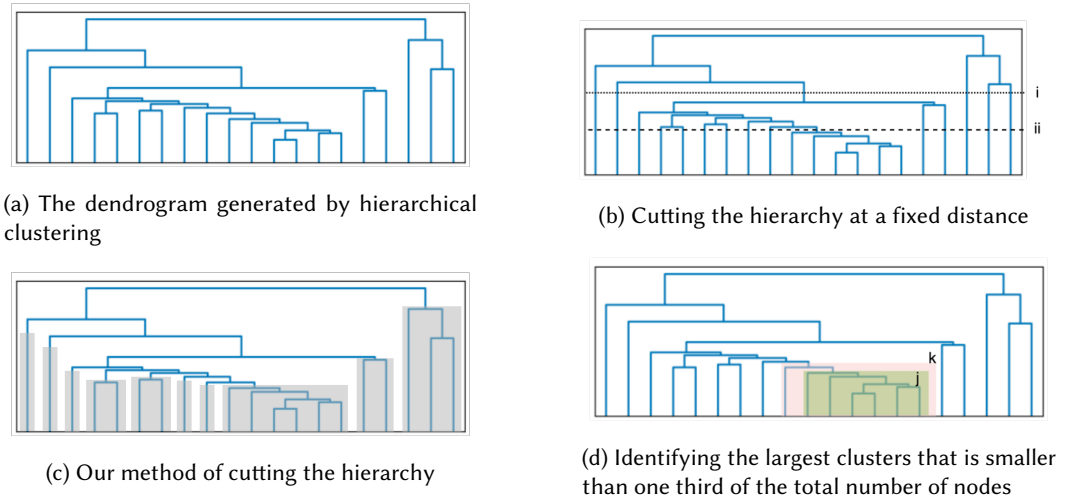


Fig. 3. (a) The hierarchical clustering algorithm produced a tree structure. (b) Cutting the hierarchy at a fixed distance level can easily generate one very large cluster (i) or numerous small clusters (ii). (c) We cut the tree such that each sub-tree is as large as possible but has no more than one third of all leaf nodes. (d) Cluster j is a valid cluster as including one more node will result in a cluster (Cluster k) with seven nodes, exceeding one third of the total number of leaf nodes (20 nodes in total).

used method of cutting the tree at a fixed level, which may produce one very large partition of loosely related concepts or numerous small partitions if the original tree structure is highly skewed. Figure 3 illustrates a concrete example of our partitioning method.

3.3.2 Interactive System Implementation. We implement GRAFS as an interactive web application. Its frontend elements leverage the JavaScript library D3.js. Its backend document indexing and search functions are based on Apache Solr. At the beginning of a search, a user can issue her search query of interest in the same manner as in a typical search engine. Simple keyword queries or complex Boolean queries are supported. The user can also specify the number of documents ($|D_q|$) to retrieve. We set the default number of $|D_q|$ to be 1000, since we aimed to assist users in the exploration of a large and complex information space. After retrieving a set of documents, the system backend will compute the data model (knowledge subgraph, concept provenance, and graph partitioning) on the fly, which will serve as the input to the frontend interface. The user can specify the desired number of concepts in the knowledge subgraph, which by default is set to 20. The default number 20 was determined empirically during prototype development to provide an informative set of concepts while also not overwhelming the user with too much information.

Interface Overview. A demonstration video showing all features of the GRAFS interface is [available online](#)². Figure 4 shows the interface rendered for the search topic “treatment for depression” with 20 selected concepts. The corresponding search query is shown in the search box at the top of the interface, where the user can construct or edit search queries.

²<https://tinyurl.com/m3nabas3>

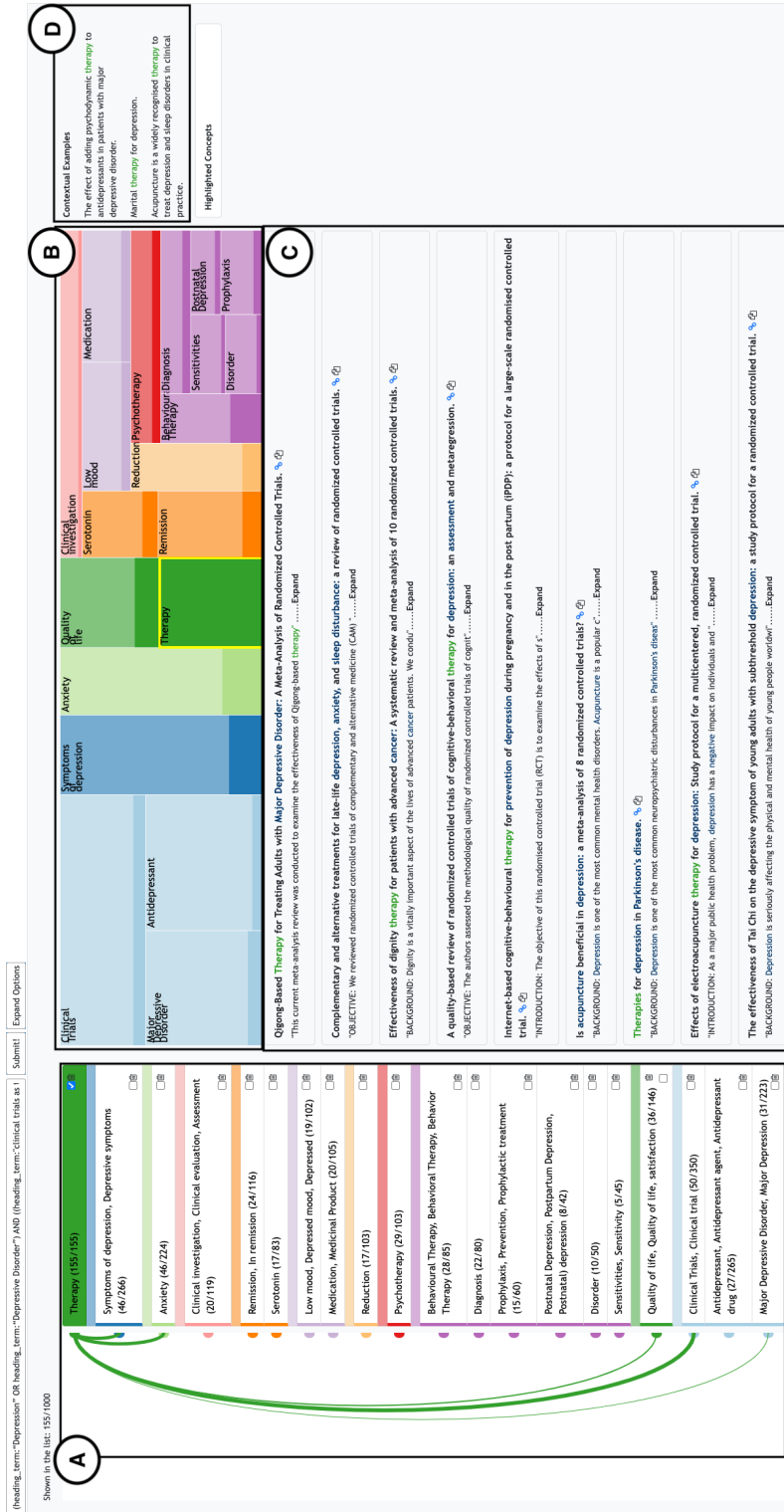


Fig. 4. GRAFS for the topic “treatment for depression”, including (A) the graphical facet list containing key concepts and their relations represented by arcs, (B) the graphical facet treemap showing the hierarchical structure of concepts based on the graph partitioning results, (C) the document list showing filtered articles ranked by their relevance to the search query, (D) the concept provenance that lists three examples of the selected concept.

To support exploratory search, GRAFS user interface (UI) follows the structure of a faceted search interface with a *facet list* (Figure 4(A)) on the left and a *document list* (Figure 4(C)) in the middle (DG1). In Figure 4(A), each facet is a concept/node of the initial knowledge subgraph, which allows users to learn major concepts and filter articles by those concepts. The *facet treemap* (Figure 4(B)) shows a space-filling treemap [40] of concepts, which visualizes the prevalence and relationship with each other. Given a selected concept (“Therapy” in Figure 4), the concept provenance sentences are listed on the right (Figure 4(D)), providing a succinct summary of the concept.



Fig. 5. Facet list transitions. (a) The user starts with an initial view. (b) After the concept “Therapy” is selected, it is moved to the top of the facet list. Arcs are drawn to connect the selected concept to the five most related concepts. The arcs have the same color as the selected concept. (c) A second concept “Quality of Life” is selected, and it is also moved to the top of the facet list. Relations between the intersection of the selected two concepts (at the top) and the other concepts are again shown by arcs.

1D Graphical Facet: Facet List. Figure 5(a) shows an initial view of the facet list. The one-dimensional facet list contains the concepts/nodes in the knowledge subgraph. Based on the result of graph partitioning, the order of concepts follows the order of leaves given by the cluster hierarchy and the color of each concept indicates its group identity.

Users can select a concept to filter the retrieved document set so that only documents containing the selected concept will be listed in the document list. For instance, the user might be interested

in “Therapy.” Upon selection, the concept (“Therapy” in Figure 4 and 5(a)) will appear at the top of the list. The rest of the concepts maintain the same order as the initial facet list to maintain the concept grouping information.

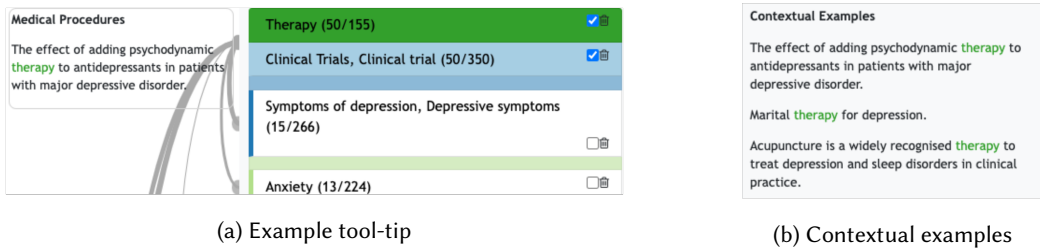
The visualization component in the facet list was also designed to facilitate users’ exploration of the knowledge subgraph and construction of their own mental model. To achieve this, arcs are drawn to connect the selected concept to the top five most related concepts in the knowledge subgraph. The choice of “top five” was determined empirically during prototype development as a compromise between the desire to provide information and the need to avoid excessive visual complexity that could reduce usability. Though the users in our evaluation study appeared satisfied with the choice of five arcs (no negative feedback was provided regarding this design choice), other thresholds or heuristic-based approaches are possible alternatives. The top arc connections allow users to inspect a small portion of the per-query knowledge subgraph given the current search focus and get a visual hint about possibly interesting concepts to explore. For instance, using the co-occurrence frequency to indicate relationships between concepts, the five concepts most related to “Therapy” are connected with arcs drawn with a thickness that is proportional to the corresponding co-occurrence frequency. For example, “Therapy” and its most frequently co-occurring concept “Clinical Trials” appear together in 50 documents. This is reflected by the thickest arc in the example which is drawn to connect these two concepts.

The interface also allows users to further narrow down the search results by applying multiple filters. When more than one concept is selected, only documents that contain all the selected concepts will be included in the interface. For instance, the user may observe that “Quality of Life” was clustered into the same group as “Therapy”, but their connection is not as strong as other concepts such as “Clinical Trials”. To answer the question, the user applied “Quality of Life”, which allows the user to examine the documents where the two concepts co-occur. Further reading helped the user find several articles that mention “Quality of Life” as an important secondary outcome for measuring the effectiveness of “Therapy”. The relation between the two selected concepts and the rest concepts are again shown by arcs, where the thickness indicates co-occurrence frequency, as shown in Figure 5(c). Based on those arcs, the user may hypothesize that “Symptoms of Depression” and “Anxiety” can also serve as measures of the effectiveness of “Therapy”, and these measures may be used in “Clinical trials”. They can test their hypothesis by further selection and reading. Therefore, these arcs in the facet list visually nudge the user to narrow down their interest by selecting the next concept.

When the user hovers over a concept in the facet list, a tooltip with the concept’s example sentence appears to the left of the facet list as shown in Figure 6 (a). At the same time, all the arcs that are not connected with the hovered concept turn grey to highlight only the arcs connecting to the hovered concept. When multiple concepts from the same cluster are selected, the arcs have the same color as the cluster (Figure 5(c)). When multiple concepts from different clusters are selected, the arcs are grey, a color that is not used by any cluster (Figure 6 (a)).

Concept Provenance. Concept provenance sentences are included as (1) the tooltip shown to the left of the facet list when the user hovers over a concept and (2) contextual examples listed on the right. The tooltip only includes the most related sentence extracted, aiming at providing a quick explanation of the concept as the user goes through the concept list, as shown in Figure 6(a). Users can refer to all three extracted representative sentences in contextual examples, as shown in Figure 6(b).

2D Graphical Facet: Facet Treemap. To provide a spatial overview of the selected concepts, the GRAFS UI also includes the facet treemap, which directly shows the hierarchical structure given by graph partitioning. Each rectangle represents a concept. The rectangles are colored to match the concepts’ group identities, and the size of the rectangle reflects the prevalence of the



(a) Example tool-tip

(b) Contextual examples

Fig. 6. Concept provenance. In (a), as the user hover over a concept in the facet list, the type of the concept and an example are shown in a tooltip next to the concept. In (b), when the user selects a concept, three examples are shown to provide a brief summary of how the concept is used in context.

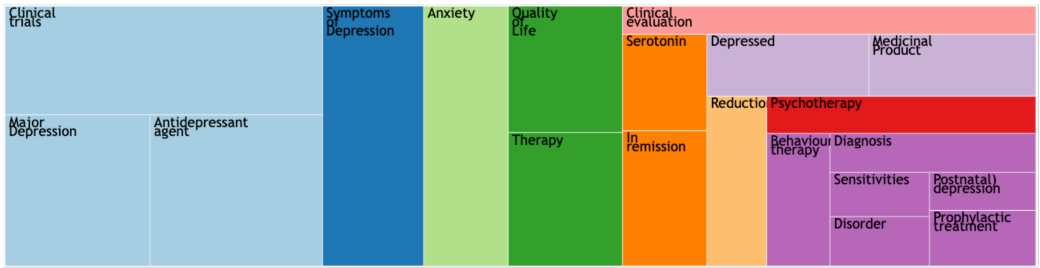
corresponding concept. Figure 7 illustrates the transition of the facet treemap through an example. A treemap representation was selected for two reasons. First, it naturally encodes the clustering of concepts in a way that communicates relative frequencies of occurrence. Second, it has a relatively simple visual structure even when large numbers of concepts are included. This is in contrast to alternatives such as node-link diagrams which can be difficult to interpret as the number of nodes and edges increase.

In addition to communicating concept prevalence, the facet treemap also provides an alternative way for users to explore concept relations. The position of concepts gives clues to the distance between concepts. Moreover, when the user selects concepts, a histogram is included in each concept's rectangle to show the percentage of documents containing the selected concepts. For instance, Figure 7 (b) shows the facet treemap after selecting "Therapy", and the darker area in "Clinical Trials" shows the percentage of documents mentioning "Therapy" among all the documents containing "Clinical Trials." The additional use of color does add some visual complexity to the visualization. However, the interactive nature of this feature, which links a user's clicks to the corresponding color changes, reduces the chance of confusion. Participants in the evaluation study did not have any difficulty interpreting the visual representation of this feature.

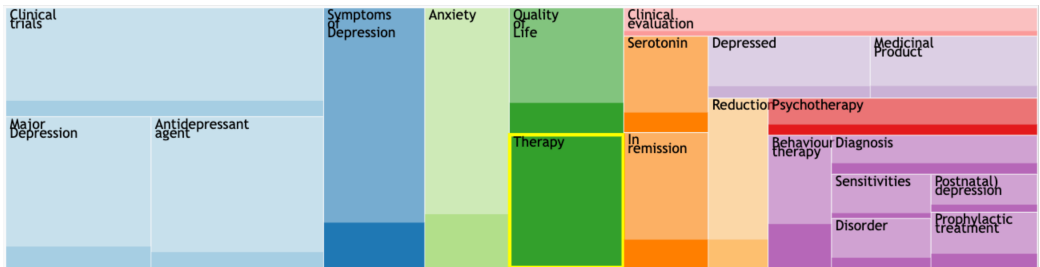
Different from the arc connections indicating the co-occurrence frequency, those percentage histograms reflect concept relations relative to concept prevalence (DG2). This property helps surface interesting relations that might be hidden by arcs, especially for less frequent concepts. For instance, after selecting "Prophylactic treatment", the darker area fills a relatively large percentage of the rectangle of "Postnatal depression", indicating that "Postnatal depression" might be closely related to "Prophylactic treatment" (Figure 8 (a)). And this relation seems to be stronger compared to "Symptoms of Depression", since the percentage covered by the darker area in the "Symptoms of Depression" rectangle appears to be smaller than in the "Postnatal depression" rectangle. This relation can be easily ignored if the user only looks at the arcs, where the connection of "Prophylactic treatment" with "Symptoms of Depression" seems to be much stronger than with "Postnatal depression" based on the absolute number of co-occurrences (Figure 8 (b)).

As the facet list and the facet treemap contains the same set of concepts, we use brushing techniques to synchronize interactions on either of them. Users can select a concept by directly clicking on it through the facet treemap, and the facet list will be re-rendered to incorporate the selection. Hovering over a concept on the facet treemap also has the same effects as hovering over the facet list.

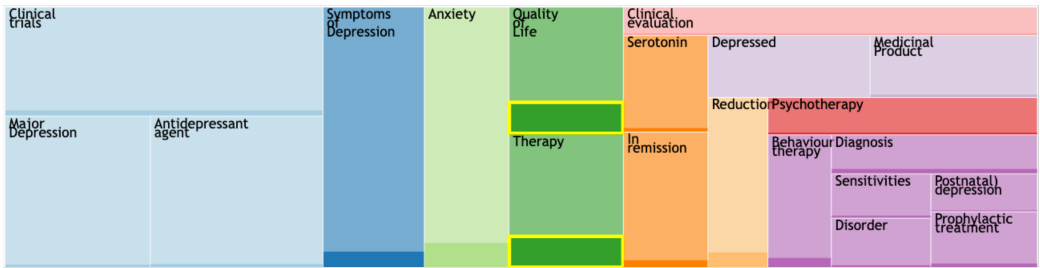
Editable List. Users may want to make changes to the facet list based on their domain knowledge or new discoveries during their exploration. As shown in Figure 9, users can add new concepts or remove existing concepts from the list. Deletion can either be done through the facet list or the facet



(a) Initial view



(b) Select "Therapy"



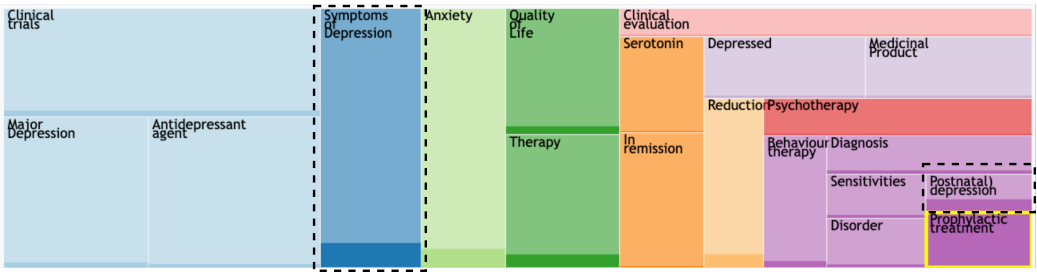
(c) Select "Therapy" and "Quality of Life"

Fig. 7. Facet treemap transitions by doing the same sequence of actions as in Figure 5. The user first selects “Therapy” from the initial view (a). In (b), the dark bar in each rectangle shows the overlap with the selected concepts. The user continues to select “Quality of Life”, giving rise to (c).

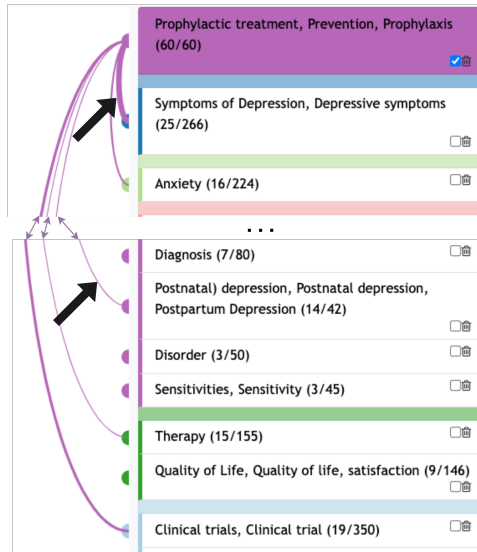
treemap. As shown in Figure 9 (c), the deleted concepts will be listed separately in the interface at the top of the facet list in case the user wants to add the concept back. Users may encounter interesting concepts in articles during exploration and want to gain a deeper understanding of them. To facilitate this, all the recognized SNOMED-CT concepts are annotated in articles and can be added to the facet list. The adding and removal of concepts will trigger the data model to re-calculate the knowledge subgraph and graph partitioning given the edited facet list (DG3).

4 EXPERIMENTAL EVALUATION

We conducted a controlled user study to evaluate the usability of GRAFS and its ability to facilitate exploratory search and learning. We compared GRAFS to a typical faceted search interface with the same set of facets and documents to understand the effects of the newly introduced knowledge subgraph and visualization components.



(a) Select "Prophylactic treatment" in the facet treemap



(b) Select "Prophylactic treatment" in the facet list

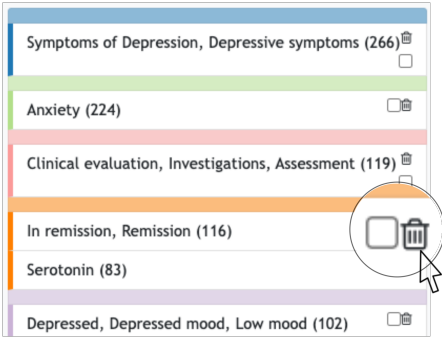
Fig. 8. The facet treemap may help surface interesting relations hidden by arcs. The strong relation between “Prophylactic treatment” and “Postnatal depression” shown in (a) is not obvious in (b).

4.1 Hypotheses

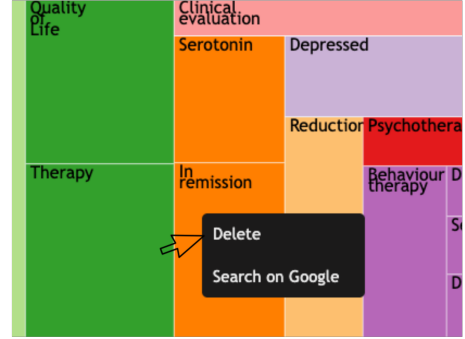
The user study was designed to test the following hypotheses:

Hypothesis 1. With the help of the knowledge subgraph, GRAFS positively influences users’ conceptual understanding and sensemaking activities.

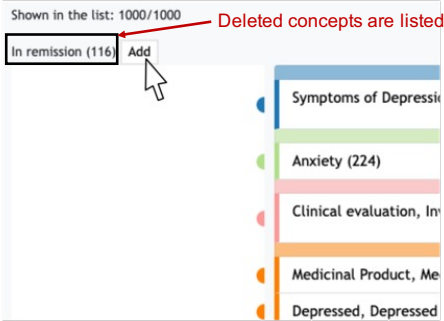
- 1.a We expect GRAFS to help users develop a better understanding of concept relations compared to the baseline system. The graphical *facet list* and *facet treemap* present relations between concepts in the current searching context and an overview of the large information space, which may otherwise require a substantial effort to build mentally via manual search and inference.
- 1.b Seeing concept relations helps users gain a deeper understanding of the search topic during sensemaking. Knowing concept relations should help users build a better conceptual understanding and mental model of the topic, and encourage users to explore more deeply into the results.



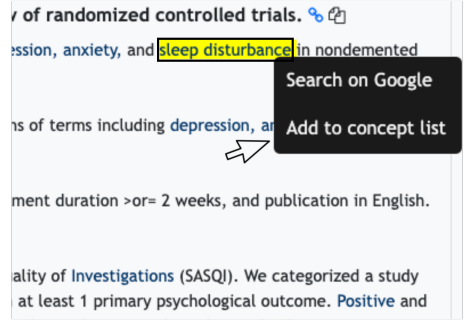
(a) Delete a concept from the facet list



(b) Delete a concept from the facet treemap



(c) Add a deleted concept back to facet list



(d) Add a concept from document

Fig. 9. Deletion can either be done through the facet list (a) or the facet treemap (b). The deleted concepts will be listed separately in the interface at the top of the facet list, and users may add them back by clicking on the "Add" button as in (c). (d) Users can add recognized SNOMED concepts to the facet list from documents.

Hypothesis 2. The additional complexity introduced in GRAFS has no negative influence on users' searching activity.

- 2.a We expect that the introduction of the additional features in GRAFS will not hurt people's performance in facet filtering tasks. GRAFS preserved the layout and functionality of a faceted search interface, so users should be able to perform filtering tasks as in a typical faceted search interface.
- 2.b We expect the complexity of additional information provided in GRAFS will not hurt the overall usability of the system. The proposed data model and interface design present the concept relation information in a simplified way that facilitates users' learning and understanding of the complex information.

4.2 Baseline Faceted Search Interface

We compared GRAFS to a typical faceted search interface with a list of facets on the left and a ranked list of documents taking the major part of the page as shown in Figure 10. To minimize the difference in the information conveyed by the two systems, we applied the same concept selection method as in GRAFS. In the baseline system, however, concepts are organized by pre-defined categories. The knowledge subgraph or graph partitioning are not surfaced in the baseline interface.

(heading_term:"Depression" OR heading_term:"Depressive Disorder") AND (heading_term:"clinical trials as shown in the list: 155/1000

Submit Expand Options

A

- Diseases & Disorders
- Anxiety (46/224)
- Major depression, Major Depressive Disorder (31/223)
- In remission, Remission (24/116)
- Depressed mood, Depressed, Low mood (19/102)
- Disorder (10/86)
- Sensitivities, Sensitivity (5/45)
- Postnatal depression, Postpartum Depression, Postnatal depression (8/42)
- Signs & Symptoms
- Medical Procedures
- Clinical trials, Clinical trial (50/350)
- Therapy (155/155)
- Clinical Investigation, Investigations, Clinical evaluation (20/119)
- PSYCHOTHERAPY (29/103)
- Reduction (17/103)
- Behavioral Therapy, Behavior Therapy, Behavioural Therapy (28/85)
- Diagnosis (22/80)
- Prevention, Prophylaxis, Prophylactic treatment (15/60)
- Medications

B

Qigong-Based Therapy for Treating Adults with Major Depressive Disorder: A Meta-Analysis of Randomized Controlled Trials. [% 2](#)
"This current meta-analysis review was conducted to examine the effectiveness of Qigong-based therapy".....Expand

Complementary and alternative treatments for late-life depression, anxiety, and sleep disturbance: a review of randomized controlled trials. [% 2](#)
"OBJECTIVE: We reviewed randomized controlled trials of complementary and alternative medicine (CAM).....Expand

Effectiveness of dignity therapy for patients with advanced cancer: A systematic review and meta-analysis of 10 randomized controlled trials. [% 2](#)
"BACKGROUND: Dignity is a vitally important aspect of the lives of advanced cancer patients. We concl.....Expand

A quality-based review of randomized controlled trials of cognitive-behavioral therapy for depression: an assessment and meta-regression. [% 2](#)
"OBJECTIVE: The authors assessed the methodological quality of randomized controlled trials of cognit.....Expand

Internet-based cognitive-behavioural therapy for prevention of depression during pregnancy and in the post partum (IPPP): a protocol for a large-scale randomised controlled trial. [% 2](#)
"INTRODUCTION: The objective of this randomised controlled trial (RCT) is to examine the effects of.....Expand

Is acupuncture beneficial in depression: a meta-analysis of 8 randomized controlled trials? [% 2](#)
"BACKGROUND: Depression is one of the most common mental health disorders. Acupuncture is a popular.....Expand

Therapies for depression in Parkinson's disease. [% 2](#)
"BACKGROUND: Depression is one of the most common neuropsychiatric disturbances in Parkinson's disease.....Expand

Effects of electroacupuncture therapy for depression: Study protocol for a multicentered, randomized controlled trial. [% 2](#)
"INTRODUCTION: As a major public health problem, depression has a negative impact on individuals and.....Expand

The effectiveness of Tai Chi on the depressive symptom of young adults with subthreshold depression: a study protocol for a randomized controlled trial. [% 2](#)
"BACKGROUND: Depression is seriously affecting the physical and mental health of young people worldw.....Expand

Efficacy of brief dynamic interpersonal therapy in patients with major depressive disorder: a prospective, multicenter randomized controlled trial protocol. [% 2](#)
"BACKGROUND: In China, psychodynamic psychotherapies are widely used as a treatment for depression. H.....Expand

1 2 3 4 5 6 7 8 9 10 next page

C

Contextual Examples

The effect of adding psychodynamic therapy to antidepressants in patients with major depressive disorder.

Martial therapy for depression.

Acupuncture is a widely recognised therapy for depression and sleep disorders in clinical practice.

Highlighted Concepts

Fig. 10. Baseline search interface. It contains (A) facet list where concepts are organized by pre-defined categories, (B) document list showing filtered articles ranked by their relevance to the search query, (C) concept provenance that lists 3 examples of the selected concept.

4.3 Study Design

4.3.1 Overview. We adopt a within-subjects design ($n = 20$) where each participant was exposed to two Web-based search interface conditions, GRAFS and the baseline, working on two topics (“COVID-19 Diagnosis” and “Treatment for Depression”). As illustrated in Table 1, 20 users were randomly split into four groups (i.e., five users per group). Each user completed two sessions. In each session, they worked on one topic using one system.

	Session 1		Session 2	
	Topic	System	Topic	System
Group 1	COVID-19	GRAFS	Depression	Baseline
Group 2	Depression	Baseline	COVID-19	GRAFS
Group 3	Depression	GRAFS	COVID-19	Baseline
Group 4	COVID-19	Baseline	Depression	GRAFS

Table 1. Experimental design. Participants were split into four groups. We counter-balanced the order of using the two systems and the topics used for testing the two systems.

To control for the difference between topics (e.g., topic difficulty and user familiarity), we counter-balanced the topics used for testing GRAFS and the Baseline. And to control for learning and fatigue effects, we counter-balanced the order of interface conditions. To control for differences in search results, we pre-specified search queries for each topic and did not allow participants to make edits to the queries. In principle, however, GRAFS does allow users to enter arbitrary search queries which are not pre-specified.

4.3.2 Participants. We recruited 20 participants (10 males, 10 females) via campus-wide mailing lists. All participants either were pursuing or had attained a graduate degree in a STEM field. Six participants had a biomedical background or professional experience, such as working as a Molecular Biology researcher. Nine participants’ field of study is information science, four participants’ field of study is computer science, and one participant’s field of study is analysis and management. Participants were randomly assigned to different conditions.

4.3.3 Search Tasks. Participants were asked to imagine being tasked to write a survey paper on either COVID-19 Diagnosis or Treatment for Depression. Given a search system and a query, such as GRAFS and COVID-19 Diagnosis, participants needed to complete two tasks: (1) produce an outline for the hypothetical paper (30 minutes), and (2) answer questions about specific aspects of the topic by facet filtering (10 minutes).

Task 1: Outline Generation. To test **Hypothesis 1**, the user study task should be a realistic exploratory search task that requires conceptual understanding and sensemaking. Prior works on exploratory search and task design suggest the following desirable characteristics of exploratory search tasks: “uncertainty, ambiguity, discovery, be an unfamiliar domain for the searcher, provide a low level of specificity about how to find the information, and be a situation that provides enough imaginative context for the participants to relate and apply the situation” [23]. We designed our task to meet the above-mentioned characteristics. We adopted the scenario of developing an outline of a survey paper on biomedical topics. The process of generating an outline requires the user to make sense of the large information space of the search results, understand important concepts within the space, and organize pieces of information into a structure based on their understanding of how those pieces are related to each other. The task requires the user to quickly build a mental model of the search topic and present it in the format of an outline. At the same time, the included

concepts and the structure of the participants-generated outlines surface their mental model of the topic and allow us to objectively evaluate the exploratory search outcomes.

Participants were asked to gather information and generate an outline of the hypothetical paper during the provided 30 minutes. Given the time limit, we suggested to participants that they should produce outlines in the form of bullet points with indentations that indicate the structure rather than organized, well-written paragraphs. Participants were instructed to provide sub-topics, major arguments or short descriptions for each sub-topic, and document references in their outlines to reflect the overall structure of the hypothetical paper. They can create and edit the outline at any point during the 30 minutes session.

Task 2: Question Answering. The outline task evaluates users' exploration of a broad topic, which is often not how faceted search is primarily used. To test **Hypothesis 2.a**, we designed a question-driven task where participants focus on looking up answers for the following specific questions through facet filtering.

- *COVID-19 Diagnosis*: Find evidence that antibody can be used to detect asymptomatic COVID-19 infection.
- *Treatment for Depression*: Find if there are any clinical trials that study the effects of antidepressant medication.

Participants were given 10 minutes to find one or two articles that can answer the given question. They were not required to summarize the found articles.

Rationale for Selecting the Search Topics: As many of our participants are not from the biomedical domain, well-known health-related topics with relatively easy terminologies are favorable. Both COVID-19 and Depression are fairly familiar topics for laypeople, and they also involve professional concepts in biomedical literature that require learning and exploration. Therefore, we chose these two topics so that participants could start their exploration quickly and have the potential to learn unfamiliar concepts.

4.3.4 Outcome Measures. We evaluated the system using both subjective feedback from users and an objective evaluation of the generated outlines. We first list the different types of measures used in the study and then connect the measures and metrics to each hypothesis.

User Perceptions: In the post-task questionnaire after each sub-session, participants provided their level of agreement with nine statements listed in Table 2, on a Likert scale from 1 (strongly disagree) to 7 (strongly agree).

Objective Evaluation of the Generated outlines: We evaluated the quality of participant-generated outlines according to six rubric items as listed in Table 3. Three graders were asked to select the better outline between the two generated by each user on all of the criteria, and ties are not allowed. The graders were not informed of the system used to generate the outlines and did the grading independently. In the end, we counted the total number of votes given to each system on each criterion.

Time Spent on Task 2: During the question-answering task, we measured the time spent by participants to find articles that they believed can exactly answer the given question.

Qualitative User Feedback: We asked for detailed feedback using a semi-structured exit interview with the questions in Table 4. We aimed to obtain open-ended findings besides testing our hypotheses regarding the general influence of GRAFS on exploratory search and the design insight of each interface component.

User Interactions: As an auxiliary measure of qualitative user feedback, we counted the number of participants that used *Facet List*, *Facet Treemap*, *Editable List*, and *Concept Provenance* based on their action logs automatically collected during the user study. This piece of information helped

us focus our qualitative analysis of a component's usability on the feedback provided by those participants who actually engaged with the component.

To keep a natural flow of user study, some measures (e.g., post-task questionnaire and exit interview) collect information to test multiple hypotheses. We highlight the connection between the measures and the hypotheses below.

Hypothesis 1

- *User Perceptions.* In the post-task questionnaire, we listed four items (Q1-Q4 in Table 2) regarding how GRAFS helps the participants generate the outlines. Q2 aims to test **Hypothesis 1.a** by directly asking the participants whether GRAFS helped them learn concept relations. Q1, Q3, and Q4 are system qualities that can help users develop a complete and deep understanding of the topic, thus testing **Hypothesis 1.b**.
- *Objective Evaluation of the Generated Outlines.* We obtained objective quality evaluations of the outlines according to the six rubric items as listed in Table 3. The six rubrics assess an outline's quality from different aspects. A1, A2, and A3 focus on the system's effect on sensemaking and learning, while A4 and A5 focus on the effect on people's searching and foraging. A6 measures the overall quality of the outline. If GRAFS helps participants gain a deeper understanding of the topic, participants should be able to generate outlines with higher quality when using GRAFS compared to using the baseline. We compared the number of votes given to GRAFS and the baseline to test **Hypothesis 1.b**. A1 also helps test **Hypothesis 1.a**, since the structure of the outline reflects a participant's understanding of how concepts are related to each other.
- *Qualitative User Feedback of the system.* In the exit interview, we asked the participants to select the better outlines from the two they generated and a preferred system (I1 and I4 in Table 4). Participants' selections reflect the overall effectiveness of the system in the exploratory search task, thus helping test **Hypothesis 1**. We compared the number of votes given to GRAFS and the baseline. Furthermore, we did a qualitative analysis of participants' answers during the exit interview to gain insights into how GRAFS facilitates the exploratory search process (Table 4).

Hypothesis 2

- *User Perceptions.* To test **Hypothesis 2.a**, we include one question in the post-task questionnaire regarding the easiness of doing the task (Q5 in Table 2). To test **Hypothesis 2.b**, we include four items (Q6 - Q9 in Table 2) related to the overall usability of the system.
- *Time Spent on Task 2.* The time spent on looking up answers during Task 2 measures the difficulty for the participants to conduct the facet filtering task using the system, thus helping test **Hypothesis 2.a**.
- *Qualitative User Feedback and User Interactions.* Participants' selections of the preferred system (I4) reflect the overall usability of the system, thus helping test **Hypothesis 2.b**. The qualitative analysis of participants' feedback I4 and I5 also helps test the usability of GRAFS and its individual components.

4.3.5 Procedure. User study sessions were conducted either in-person or remotely using Zoom video conferencing, and each lasted for roughly two hours. After providing informed consent, participants completed a demographics questionnaire that asked about their background and prior experience with literature search. A study session contained two sub-sessions that followed the same sequence of steps and a participant completed the outline task and question answering task on one topic using one interface design in each sub-session. At the beginning of a sub-session, we guided the participant through a hands-on tutorial of the search interface using an example topic.

		Question	Hypothesis
Outline generation	Q1 Overview	The faceted search tool helped me gain an overview of key concepts in the search results.	1.b
	Q2 Relation	The faceted search tool helped me see how key concepts are related in the search results.	1.a
	Q3 Discover	The faceted search tool helped me discover interesting aspects to explore further.	1.b
	Q4 Decision	The faceted search tool helped me decide which aspects to focus on and which to ignore.	1.b
Question answering	Q5 QA difficulty	I find it difficult to use the faceted search tool when I complete the question answering task.	2.a
Usability	Q6 Easy to use	The faceted search tool is easy to use.	2.b
	Q7 Manageable	The faceted search tool is manageable.	2.b
	Q8 Stimulating	The faceted search tool is stimulating.	2.b
	Q9 Well-organized	The faceted search tool is well-organized.	2.b

Table 2. Post-task questionnaire. Participants provided their level of agreement on a Likert scale from 1 (strongly disagree) to 7 (strongly agree). Each question is associated with a hypothesis.

		Description	Hypothesis
Sensemaking/ Learning	A1 Structure	Which outline is more logically organized/structured, as opposed to a list of randomly ordered points?	1.a, 1.b
	A2 Interpretation	Which outline contains more interpretation written by users?	1.b
	A3 Topical depth	Which outline covers deeper issues of the topic, as opposed to superficial issues?	1.b
Search/ Foraging	A4 Documentation	Which outline collects more papers to document their findings?	1.b
	A5 Topical diversity	Which outline covers more diverse issues, as opposed to being narrowly focused?	1.b
	A6 Overall	Which outline is overall the better one of the two?	1.b

Table 3. Outline assessment rubric. Each rubric item aims to test one or more hypotheses.

Participants were then informed of the topic to work on and asked to rate their level of familiarity with the topic on a scale from 1 (unfamiliar) to 7 (familiar). Then, participants completed the outline and question answering task. After the two tasks, participants finished a post-task questionnaire to

	Description	Hypothesis
I1 Subjective outline evaluation	Which outline are you more satisfied with?	1.a, 1.b
I2 General strategy	Could you describe your strategy for approaching the task?	-
I3 Benefit of each system	Did either of the faceted search tools help you in any way to accomplish the task? If so, how? If not, why not?	-
I4 GRAFS vs baseline	If you were to do the same task again, which of the two interfaces would you prefer to use? Why?	1.a, 1.b, 2.b
I5 Per-component usability	Did you find {concept arcs, treemap, example tooltip, deleting, adding} useful? Why?	2.b

Table 4. Exit interview questions. Some of the questions aim to test one or more hypotheses.

provide feedback for the search interface. Finally, after completing the two sub-sessions, participants provided additional feedback about their experience using the tools through an exit interview.

4.4 Data Analysis Methodology

Below we describe the data analysis methods we applied to each outcome measure in Section 4.3.4.

User Perceptions.

- *Analysis of system difference.* The average ratings of the systems are compared to understand the difference between systems. Fisher’s randomization test was used to test for significant effects due to system differences.
- *Analysis of factors other than the system difference.* Participants’ perceptions of both GRAFS and the baseline can be influenced by a variety of factors including their prior experiences and the order of using the systems. In our analysis, we split participants into groups considering the order in which they used the two systems (G_{order}), whether they have biomedical domain knowledge (G_{domain}), their experience with conducting a literature review ($G_{experience}$), and topic familiarity ($G_{familiar}$). We evaluated the effects of these group identities on the questionnaire responses with a mixed ANOVA model. The group identities were treated as between-subject factors and the system (GRAFS or baseline) was treated as a within-subject factor. We recorded the F statistic, proportion of variance η^2 , and p value for each factor.

Objective Evaluation of the Generated Outline. The number of votes given to the two systems by the graders is compared to understand the difference between the systems.

Time Spent on Task 2. The averages of the time in seconds spent on Task 2 when using the two systems were compared. Fisher’s randomization test was used to test for significant effects due to system differences.

Qualitative User Feedback.

- *Analysis of votes given to each system.* The number of votes given to the two systems by the participants when they answered I1 and I4 is compared to understand the perceived overall effectiveness and usability of the systems.

- *Qualitative analysis.* We conducted qualitative analyses on the detailed feedback regarding how the systems and system components influenced participants' activities and factors that influenced their performance on the tasks.

User Interactions. We analyzed the interaction logs of each user and computed whether they selected any concept in the facet list to view the arcs, hovered over any concept to view the concept provenance tooltip, selected any concept in the facet treemap, deleted any concept from the facet list or treemap, and added any concept from the search results.

5 RESULTS

Overall, the results from our study support **Hypothesis 1.a** and **Hypothesis 2**. We observed less clear results for **Hypothesis 1.b**. Though there is some evidence supporting **Hypothesis 1.b**, our study also illustrates certain challenges in confirming the effects of different systems on exploratory tasks. This section presents the results and feedback obtained from our study and a number of insights uncovered during our analysis. Fisher's randomization test was used to test for significant effects due to system difference [41]. We interpret an effect as *significant* if $p < .05$, and *weakly significant* if $.05 \leq p < .1$.

5.1 Overview

The presentation of the study's results and our key observations are organized according to the hypotheses defined in Section 4.1. Here we provide an overview of the main findings, with more details provided in the sections to follow.

Hypothesis 1 Results: With the help of the knowledge subgraph, GRAFS positively influenced users' conceptual understanding and sensemaking activities.

- 1.a** We found that GRAFS made it easier for users to see the relations between concepts (Q2) (Sections 5.2.1, 5.2.3)
- 1.b** We received mixed results for Hypothesis 1.b. The objective evaluation of user-generated outlines showed that participants were able to generate outlines with more depth (A3) when using GRAFS, supporting our hypothesis (Section 5.2.2). The majority of participants also subjectively preferred to use GRAFS for the given tasks (I4), indicating that GRAFS facilitated participants' sensemaking activity (Section 5.2.3). This also supports the hypothesis. However, no difference was observed for the overall quality of user-generated outlines (A6) across the two systems, and participants did not indicate in the questionnaire (Q1, Q3, Q4) that GRAFS helped them develop a deeper understanding of the search topic (Sections 5.2.1, 5.2.2). Our results also showed that the baseline system helped participants collect more documents (A4) to support their arguments, and more participants believed the outlines they generated using the baseline is better (I1) (Sections 5.2.2, 5.2.3).

Hypothesis 2 Results: The additional complexity introduced in GRAFS had no negative influence on users' searching activity.

- 2.a** We found GRAFS did not make it more difficult to do the question-answering task evaluated subjectively by participants (Q5) (Section 5.3.1) and objectively by the time participants spent on the question-answering task (Section 5.3.2).
- 2.b** We found GRAFS was rated similar to the baseline system on the usability metrics (Q6 - Q9) (Section 5.3.1). GRAFS is more stimulating (Q8) and less manageable (Q7) than the baseline system (weakly significant). The majority of participants (14 out of 20) selected GRAFS for the future task (I4) (Section 5.3.3). Overall, participants provided positive feedback for the newly introduced components (I4, I5) (Section 5.3.3). However, the *concept treemap* and the

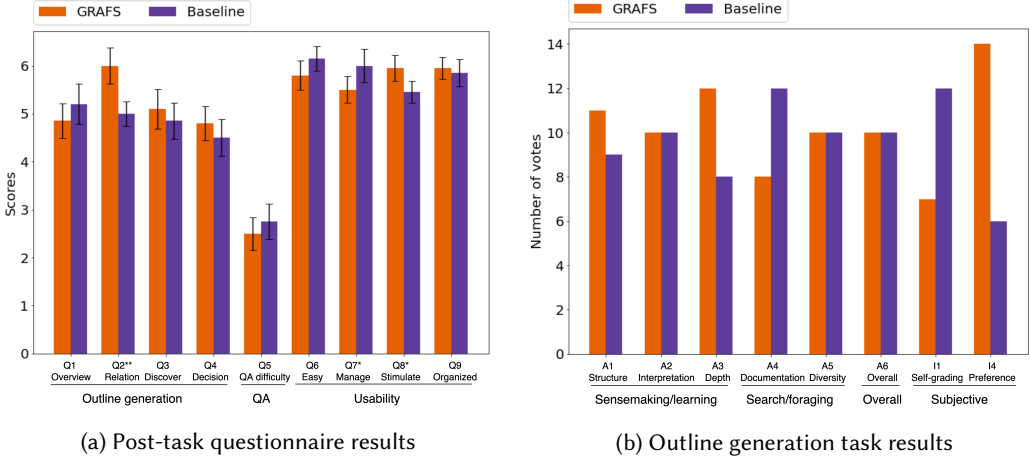


Fig. 11. Results. (a) Mean ratings given by participants in the post-task questionnaire. Error bars show the standard error. There is a significant difference between GRAFS and the baseline in Q2 (**), and a weakly significant difference in Q7 and Q8 (*). (b) Numbers of votes given to each system by graders (A1-A6) and participants (I1, I4). GRAFS gets more votes in A3 Topical depth and I4 System preference, while the baseline gets more votes in A4 Documentation and I1 Self-grading. Overall, there is no difference between the outlines generated using the two systems (A6).

function of deleting and adding concepts were both more beneficial for users with a relatively deeper understanding of the topic compared to *concept arcs* and *concept example tooltips*.

5.2 Hypothesis 1

Below we present detailed results regarding **Hypothesis 1** organized by relevant outcome measures.

5.2.1 User Perceptions.

Analysis of system difference. As shown in Figure 11(a), participants thought the two systems were different in helping them see the relation between concepts ($Avg_{GRAFS} = 6.0$, $Avg_{Baseline} = 5.0$, $p = 0.033$). This provides support for **Hypothesis 1.a** that GRAFS helps users see the relations between concepts better.

In contrast, the search system did not have significant effects on questions related to Q1 overview, Q3 discover, and Q4 decision. These results failed to show evidence supporting **Hypothesis 1.b**.

	F	η^2	p
G_{order}	0.323	0.018	0.577
$System$	6.406	0.262	0.021
$G_{order} \times System$	3.139	0.148	0.093

Table 5. Effect of order on a user's rating of Q2 Relation

Analysis of factors other than the system difference. Topic familiarity, literature search experience, and domain knowledge did not have any measurable effect on participants' responses on Q1-Q4. A weakly significant interaction effect between the order and system was found for participants' rating on Q2 Relation ($p = 0.093$, Table 5). As shown in Figure 12(a), when rating whether the system helps to see concept relations, users who worked on GRAFS first tended to give

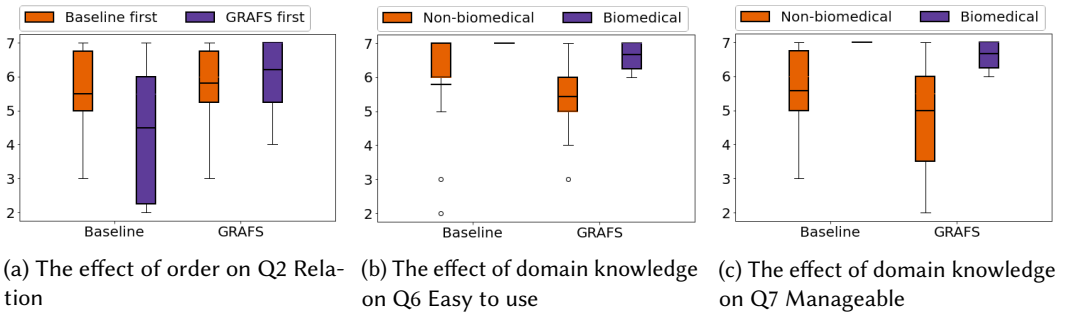


Fig. 12. The effects of the order of using the two systems and participants' domain knowledge on participants' perception of the system. (a) Participants that worked on GRAFS first gave lower scores to the baseline in the system's ability to help them see the relation between concepts. (b) Biomedical domain experts gave higher scores to Q6 Easy to use compared to participants without domain knowledge. (c) Biomedical domain experts gave higher scores to Q7 Manageable compared to participants without domain knowledge.

lower scores to the baseline system compared to the case when the user started from the baseline system. This phenomenon suggests that users did use some of the provided visual representations in GRAFS to perceive the relations between concepts. When exposed to the baseline system as the second interface, users appeared to have an increased awareness of the difficulty in seeing concept relations without the GRAFS features. This further supports our **Hypothesis 1**.

5.2.2 Objective Evaluation of the Generated Outlines. Figure 11 (b) shows graders' votes for the two systems (A1-A6). For A3 Topical depth, graders gave 12 votes to GRAFS and 8 to the baseline, supporting **Hypothesis 1.b**. In terms of A4 Documentation, 12 votes went to the baseline system and 8 went to GRAFS, indicating that being exposed to relation information may have negative effects on people's searching or foraging behavior. Therefore, A4 fails to support **Hypothesis 1.b**. One possible explanation is that when using GRAFS, participants may spend more time exploring concept relations and developing a deeper understanding, with less time left for collecting supporting documents. GRAFS and the baseline system received the same or the similar number of votes in A1 Structure, A2 Interpretation, A5 Topical diversity, and A6 Overall quality, which fails to support **Hypothesis 1.b**. Overall, A3 supports **Hypothesis 1.b** and the remaining criteria disagree with it in terms of the objective evaluation of outlines.

5.2.3 Qualitative User Feedback.

Analysis of Votes Given to each system. Figure 11 (b) shows participants' votes for the two systems (I1 and I4). When asked which system they would use if they were to do the task again, the majority of the participants ($n = 14$) selected GRAFS (I4), supporting **Hypothesis 1**. When evaluating the generated outlines, a majority of the participants ($n = 12$) felt that they did a better job producing their outlines using the baseline system (I1), failing to support **Hypothesis 1**.

Combined with questionnaire responses, one possible explanation is that being exposed to concept relations (Q2) increases complexity such that the user interface becomes less manageable (Q7, Section 5.3.1). However, participants still found that relation information was useful for the exploratory task. Therefore, users expressed a preference for GRAFS because, if given more time on task, it would support a deeper understanding of the topic during exploratory literature search. In this way, participants' preference for GRAFS can be viewed as evidence for **Hypothesis 1**.

Benefits of GRAFS and Baseline As mentioned previously, the majority of participants preferred GRAFS over the baseline system (14:6) when answering I4. During the exit interview, we

asked participants to indicate the reason for their preference. In general, participants' feedback shows that the new information and functions provided by GRAFS facilitated users' exploratory activity while introducing some complexity to the use of the system.

Participants described the benefits of GRAFS related to (1) relations, (2) specific components, and (3) general experience. Several users ($n = 4$) mentioned that they loved the fact that they can learn relations between concepts in GRAFS (**Hypothesis 1.a**). Many participants ($n = 7$) mentioned some components that they particularly liked, including facet treemap ($n = 4$), editable list ($n = 3$), and facet list arcs ($n = 3$). Four participants selected GRAFS for a better overall experience. Specifically, they mentioned GRAFS is interesting and colorful ($n = 2$), and helps with overview ($n = 1$) and new ideas ($n = 1$).

Most of the participants that preferred the baseline system thought the baseline system is clearer ($n = 4$), while two mentioned that they liked the pre-defined categories ($n = 2$).

Participants' Feedback on Factors Affecting the Generated Outlines. During the exit interviews, participants provided detailed reasons for why they preferred the outline they generated with one system over the other when answering I1. Participants' comments indicate that the interface design did affect their activity when completing the task, but the influence of the search topic on a user's outline proved to be larger. This may help explain why we did not observe major differences in the quality of people's outlines in A1, A2, A5, and A6 (Section 5.2.2).

Reflecting this, most of the participants ($n = 12$) mentioned the influence of topic difference on the outline quality. This includes the technicality of a topic, participants' familiarity with a topic, and whether a topic is evolving or well-established. Some participants ($n = 6$) thought "Treatment for Depression" was easier than "COVID-19 Diagnosis" in terms of vocabulary and the topic's structure. Five participants mentioned that they were more familiar with one of the topics, and those participants felt that they did a better job on the familiar topic. Some participants ($n = 4$) mentioned that the two topics were different from each other in the work required to perform the task, which in turn influenced the quality of the final outcomes. Three out of the four participants commented that they found depression to be a well-established field while COVID-19 related research was newer (and therefore harder to approach due to the more scattered and heterogeneous set of related concepts).

A number of participants ($n = 6$) mentioned that the features of the interface affected the quality of their outline. Four users said that GRAFS helped them generate good summaries as it enabled them to edit the concept list ($n = 2$), see relations between concepts ($n = 1$), and was easier to use ($n = 1$). In contrast, one participant thought the features in GRAFS were distracting and negatively influenced his work. One participant liked the categories given in the baseline system and reported that they helped her to write a good summary.

Overall, we found that topic differences and participants' backgrounds may have a bigger impact on outline quality than the system. On the one hand, these factors may exaggerate the number of votes given to one system versus another due to the small sample size of the study. On the other hand, these factors may also obscure system differences (i.e., if participants always perform better on one topic, then GRAFS and Baseline will always get equal votes).

5.3 Hypothesis 2

Below we present detailed results regarding **Hypothesis 2** organized by relevant outcome measures.

5.3.1 User Perceptions.

Analysis of system difference. As shown in Figure 11(a), the questionnaire results generally support **Hypothesis 2**. There is no significant difference in Q5, indicating that the additional features in GRAFS did not make the question answering task more difficult (**Hypothesis 2.a**).

The system usability questions (Q6-Q9) provide supporting evidence for **Hypothesis 2.b**. There is no significant difference in Q6 Easy to use and Q9 Well-organized, while the differences in Q7 Manageable ($Avg_{GRAFS} = 5.5$, $Avg_{Baseline} = 6.0$, $p = 0.099$) and Q8 Stimulating ($Avg_{GRAFS} = 5.95$, $Avg_{Baseline} = 5.45$, $p = 0.092$) are only weakly significant.

	F	η^2	p
G_{domain}	6.170	0.255	0.023
$System$	1.797	0.091	0.197
$G_{domain} \times System$	0.002	0.000	0.967

Table 6. Effect of domain knowledge on a user's rating of Q6 Easy to use

	F	η^2	p
G_{domain}	8.160	0.312	0.010
$System$	3.635	0.168	0.073
$G_{domain} \times System$	0.173	0.010	0.682

Table 7. Effect of domain knowledge on a user's rating of Q7 Manageable

Analysis of factors other than the system difference. Participants with biomedical background gave significantly higher scores for Q6 Easy to use ($p = 0.023$, Table 6, Figure 12(b)) and Q7 Manageable ($p = 0.010$, Table 7, Figure 12(c)) for both GRAFS and the baseline system. The effect of G_{domain} is more significant than the effect of the system on the responses for these two questions. This suggests that people's judgment of a system's usability is highly influenced by people's domain knowledge. Topic familiarity, literature search experience, and the order of using the two systems did not have any measurable effect on participants' responses on Q5-Q9.

5.3.2 Time Spent on Task 2. Participants indicated their perception of the difficulty of the question answering task through the post-task questionnaire. As mentioned in Section 5.1, there is no significant difference between the ratings given to the two systems regarding the task's difficulty ($p = 0.510$). We further evaluated participants' performance based on the time spent on the task. On average, participants spent 181.1s on the question answering task using GRAFS and 238.2s using the baseline system, and the difference is not significant ($p = 0.208$). This shows that participants performed the question answering task at least as well on GRAFS as on a typical faceted search system, which supports **Hypothesis 2.a**.

5.3.3 Qualitative User Feedback.

Analysis of Votes Given to each system. When asked which system they would use if they were to do the task again, the majority of the participants ($n = 14$) selected GRAFS (Figure 11 (b), 14), implying that participants perceived GRAFS to be usable overall. This result thus helps support **Hypothesis 2.b**.

Benefits of GRAFS and Baseline. During the exit interview, we asked participants to describe why they prefer one system over the other. Participants described the benefits of GRAFS related to (1) relations, (2) specific components, and (3) general experience. The detailed comments are presented in Section 5.2.3. These comments reflect that a majority of the participants found the newly introduced components usable and were able to apply that new information to their tasks, providing support for **Hypothesis 2.b**. However, some participants also pointed out that the

baseline system is clearer ($n = 4$), implying that the newly introduced features in GRAFS add complexity to the use of the system.

Usability of Interface Components. During the exit interview, participants were asked whether a component is useful or not for accomplishing the given tasks. We note that there is a difference between (1) a participant not using a feature, and (2) a participant finding the feature not useful after using it. More specifically, we noticed that some participants did not use certain features in GRAFS due to time limitations, lack of domain knowledge, and limited experience with a new interface design. Therefore, we report both (1) the number of participants that used each feature based on the action log, and (2) the number of participants that found each feature to be useful.

Overall, the facet list arcs were the most used feature. The facet treemap was used less often, in part due to its higher complexity. The functions of deleting and adding concepts were the least used. Though they all add complexity to the traditional faceted search interface, the components all received positive feedback from participants who have used them. The concept provenance tooltip was found to be easy to use, but not as useful as other components.

Facet List Arcs: Action logs show that all 20 participants made at least one selection from the facet list, which triggers the display of facet list arcs. The majority of participants thought the feature of concept arcs to be useful ($n = 11$). Some participants ($n = 6$) mentioned that arcs helped inform them about what to look at or which concepts to click next (**Hypothesis 1.b**). Five participants mentioned that arcs helped them understand relations between concepts (**Hypothesis 1.a**), which in turn helped them better understand the topic and adjust their selection strategy (**Hypothesis 1.b**).

For participants that did not state that arcs were useful, a majority of them said that they ignored this feature ($n = 5$). For instance, User 4 said that he preferred to just read articles and User 12 mentioned that due to time limitations, he was not able to use that information. Some participants ($n = 2$) used alternatives to the arcs to get similar information, such as the facet treemap, or the number of documents listed near each concept.

Facet Treemap: We considered a participant to have used the facet treemap if they clicked at least once on a concept in the visualization. 13 participants meet this criterion. Participant feedback illustrates that the facet treemap can be difficult to make sense of due to its unfamiliarity, and it introduces complexity to using the system. Yet while it may take time for users to learn how to use this feature, nine out of 20 participants thought that the facet treemap was useful.

Most of the participants who found it useful ($n = 8$) reported that the facet treemap helped them see the relations between concepts (**Hypothesis 1.a**). Because both the facet treemap and the facet list arcs communicate the connections between concepts, some participants ($n = 7$) compared the two. Five participants indicated that they preferred facet treemap compared to facet list arcs, while the other two participants preferred arcs. Those two felt that the arcs were more straightforward.

Interestingly, among people that preferred the facet treemap, two participants commented that the facet treemap was not straightforward to understand and that it took time for them to find it useful. For instance, User 20 mentioned that he didn't find the treemap to be useful immediately. However, during the exploration, he found that thick arcs in the facet list always seemed to connect to broader concepts such as infectious disease. He, therefore, started to look to the treemap for a more nuanced view of the relations and, eventually, found the treemap to be more useful. User 14 mentioned that at the beginning of the session, she felt the information in the facet treemap is cluttered and the structure seemed to be quite complex. As a result, she assumed that it would be hard to understand the treemap and ignored it. Only after using the facet treemap did she realize that it is easy to use.

A majority of the participants ($n = 11$) did not report the facet treemap as being useful. Three participants felt the information was cluttered and overwhelming, while some participants ($n = 2$) thought it occupied too much screen space. Three participants mentioned that it was difficult to make sense of the information provided by the facet treemap, and two participants said they were not used to information presented in that way. Many of these comments suggest unfamiliarity as a key hurdle.

Concept Provenance Tooltip: All participants used this feature when hovering over items in the facet list. Eight participants mentioned that the facet list tooltip was useful. Some participants ($n = 4$) mentioned that the tooltip helped them gain a quick understanding of what the concept meant. Two participants expressed the wish that they could navigate directly to an original article from the given examples.

For people that did not report that the facet list tooltip was useful, four mentioned that they did not need extra explanations of the concepts either because they had good background knowledge or they only used familiar concepts during their exploration. Some participants ($n = 3$) preferred to understand concepts using other methods, such as “googling” or reading the actual articles. One participant mentioned that the usefulness of the tooltip was highly dependent on the content of the sentence shown in the tooltip. Overall, the concept provenance tooltips were found to help people understand concepts during exploration, but only when users were faced with new concepts.

Editable List: GRAFS allows users to make adjustments to the automatically generated knowledge subgraph by deleting or adding concepts. Overall, this feature was less used by participants due to task time constraints, limited familiarity with the topic, and limited training with this way of interacting. However, participants provided positive feedback on this feature, and we found some interesting use cases that indicate the benefits of preserving human agency in GRAFS.

In total, seven participants used the function of either deleting or adding concepts. Out of these seven, four participants used both, two participants only deleted concepts, and one participant only added concepts.

We examine the feedback regarding deleting concepts and adding concepts separately. Six of the 20 participants used the function of deleting concepts, and six participants (four of the six participants that used concept deletion, plus two who did not use concept deletion) mentioned that the function of deleting concepts was useful for removing irrelevant or equivalent variants of concepts. This helped users focus on a shortened concept list ($n = 2$) and find important relations faster ($n = 1$).

Five of the 20 participants used the function of adding concepts, and four participants (three of the five participants that used concept addition, plus one who did not use concept addition) found adding concepts to be useful. Two people used this function when they came across unlisted concepts that they felt were important based on their reading. An interesting use case was provided by User 2. She described that when she worked on “Treatment for Depression,” she frequently came across the concept of “Exercise.” Therefore, she added “Exercise” to the concept list to investigate deeper. When she felt she had done enough reading of related materials, she then removed the concept from the list. One participant added the concept “COVID-19,” which was removed as a per-query stop word, to remind herself about the major topic. One participant suggested the interface should enable users to add new concepts to the facet list by directly typing in concept names in addition to selecting concepts in documents.

For participants that did not use the add or remove functions, four participants mentioned that time limitation was a major reason. Some participants ($n = 3$) mentioned that the lack of familiarity with the topic is another important factor.

6 DESIGN IMPLICATIONS

6.1 The Effect of Showing Concept Relationships

The major difference between GRAFS and the baseline faceted search interface is the focus on revealing relationship information between key concepts. The results from our experiments show that this additional information can positively contribute to a user's conceptual understanding of a search topic. Showing these relationships as arcs nudges users to build more a sophisticated mental model of the information space as reflected by conceptually deeper outlines. Notably, these benefits were achieved without substantially impacting system usability. In user experience terms, the relationship arcs serve as *signifiers* that make conceptual relationships more *discoverable* by users [32]. These signifiers may not be needed in relatively simple exploratory tasks such as comparative shopping, but can be particularly valuable in research-oriented tasks where users need to explore, discover, and learn about an unfamiliar and complex information space. A future research direction is to characterize task scenarios where it is most beneficial concept relationships to augment a faceted search interface.

The mixed results for Hypothesis 1.b (“seeing relationships helps users gain a deeper understanding of the search topic”) indicate that objective and subjective evaluation of learning outcomes may not align. Although GRAFS helped participants construct objectively deeper and more organized outlines, more participants favored the outlines generated using the baseline system that had a simpler logical structure but contained more papers. Such results imply a tension between *learning* and *satisfaction* in exploratory search. Learning activities on GRAFS may have not only slowed participants down in terms of getting more papers into their outline, but also exposed them to a larger sphere of knowledge and made them think their outline was not thorough enough. In other words, learning made users aware of what they do not know and feel less satisfied with what they have already known. In psychology research, studies also observed that people's self-evaluations of test results diverge further from objective evaluations when they are less knowledgeable on the test subject [22]. The mixed results for Hypothesis 1.b have implications for future research in two aspects. First, information retrieval systems that support learning and sensemaking should explore methods for computationally estimating users' learning progress so that they can inform users of their progress. Initial work in this direction is recently explored in search-as-learning literature [47]. Second, such systems' interface should communicate user's learning progress in a positive tone and encourage them to explore further with a sense of achievement. An inspiring line of related work is in online news consumption diversification, where the goal is to encourage users to discover and read news from diverse political viewpoints [29].

6.2 The Value of User Agency

Our study suggests that it is important to preserve human agency when the system takes the initiative to recommend an imperfect data model. In GRAFS, we allow users to adjust the automatically generated data model by deleting and adding concepts. Though this feature was less used in the study, it is one of the most mentioned components when participants talked about the benefits of GRAFS. The use cases provided by participants show that they adjust the concept list based on their exploratory focus, such as concepts they want to ignore, keep track of, and look deeper into. We believe this feature will be more widely used when users have more time in actual exploratory search tasks, as many participants mentioned that they did not edit the concept list because of either time limitations or lack of familiarity with the topic.

To increase user agency, a future improvement opportunity is to allow users to directly add concepts to the facet list by typing the concept name, and the system can assist the user through auto-completion. This would allow users who have prior knowledge to directly specify concepts

they are interested in without having to find them in the result list. Even for users who learned a new concept on the task, such a feature would allow them to add a newly learned concept by recalling it from memory without having to refind a document mentioning the concept and then visually search for the concept inside the document.

6.3 When Less Is (Not) More In Surfacing Complex Information

Many design elements in GRAFS aim to minimize information overload, reflecting the minimalist motto “*less is more*.” First, the original query-specific knowledge graph containing hundreds of concepts is reduced to a much smaller initial subgraph with 20 or so key concepts. Second, among all relationships in the smaller subgraph, only five arcs connecting the most related concepts are shown at any time. Third, the original densely connected subgraph is reduced to a tree structure (facet treemap). The facet list further flattens this tree structure into a list and is better received by our participants than the tree. All of these elements selectively expose users to a small but informative portion of the underlying information, balancing complexity and usability. In the case of relationship arcs, the exposure is both selective and progressive – new arcs are gradually surfaced as the user explores different combinations of concepts.

However, not all data reductions were perceived as useful by participants. For example, our concept provenance features (hover-over tooltip on the facet list and a textbox on the right side of the screen) were designed to explain why a concept was considered relevant given the search context. The features showed example sentences (instead of whole documents) that mentioned both the concept and the current query. Participants’ qualitative feedback suggested that the provided sentences were not helpful, or even viewed as a distraction for some participants who already had a good background. In addition, some participants preferred more comprehensive explanations from a separate search or a larger context, such as the original articles. As the need for explanation varies between users and contexts, a possible solution is to only show concept provenance upon request, and to ease the navigation from provenance sentences to their original documents.

These phenomena imply that showing a small part of a larger data structure can help reduce information overload *if that data structure is self-similar*. For example, a subgraph is structurally similar to the larger knowledge graph; a subset of relationship arcs is structurally similar to the set of all arcs; a low-dimensional projection of a network is structurally similar to the original network representation. In these cases, users can still make sense of the smaller part as it preserves the “syntax” of the whole. However, fragments of natural language data (concepts and sentences) are integral parts of larger contexts (documents). Showing them out of context may result in “information underload” due to syntactic incompleteness, semantic ambiguity, and lack of coherence. This is especially true in scenarios where users are unfamiliar with the information space and may encounter difficulty in interpreting extracted concepts and sentences without seeing the original context. Therefore, “less is more” applies in some information presentation scenarios but not others.

7 LIMITATIONS AND CHALLENGES

7.1 Limitations of the System

The proposed data model and visualization elements have two key limitations: (a) they introduce complexity to users’ search when trying to present richer information; and (b) they are less familiar to users compared to a typical faceted search interface. To cope with these limitations, we built our system on top of a typical faceted search interface and only added visualizations to the marginal areas of the interface. We attempted to preserve the simplicity and familiarity of the system, and gave users the freedom to ignore additional components. Reflecting this we observed participants that used GRAFS as a Google-like search engine or who ignored new parts of the interface such

as the arcs and treemap. Our study suggests that the added interface complexity in GRAFS did not interfere with basic information retrieval tasks (e.g., looking up specific articles). In terms of subjective perception, participants indicated that GRAFS was less manageable compared to a typical faceted search interface. The same users rated GRAFS and the baseline similarly in “easy to use.” In retrospection, a potential solution would be to allow users to manage the visual complexity by turning certain features “on” or “off” (e.g., through toggle switches).

7.2 Limitations of the User Study

Two main limitations of the user study were (a) the large impact of the search topic and participants’ background; and (b) the time constraint. First, participants’ background and differences between topics may have a much larger influence on the study outcomes than the system, especially for a challenging task involving searching, reading, and writing skills. A within-subjects design may not solve this problem because even the same participant may have different background knowledge in different topics. Second, it takes time for users to make sense of the visualized information. Under the time pressure of a formal study session, participants might be discouraged to use new tools and instead follow their familiar approach to doing the task. Even when participants were willing to use the visualization components, it also required time and effort to fit the new information these new tools provided into their traditional workflow. In future work, it would be valuable to study the system’s usability by observing people using the system as they work on longer-term real-world tasks.

Two other aspects of the study design could serve as potential limiting factors. First, due to the timing of this study which overlapped with the COVID-19 pandemic, some study participants took part in the study virtually via Zoom while others participated in a face-to-face session. While we did not identify any specific impacts of participant modality in our analysis of the results, the mode of participation was a potential confounder in the execution of the study. Second, the post-task questionnaire (see Table 2) included a series of agree/disagree Likert scale questions. The phrasing of the questions as agree/disagree questions could potentially lead participants to record more positive ratings than they would provide if asked for similar feedback using alternative question formats.

8 CONCLUSION

This paper proposed graphical faceted search (GRAFS), a novel interactive approach designed to help exploratory users more effectively construct a mental model of an unfamiliar information space during exploratory search. GRAFS leverages an intelligent backend computational model that extracts a small but essential set of key concepts and their relations present in semantic search results to form a knowledge subgraph. The frontend search interface leverages this subgraph to organize and visualize search information in a faceted search-like interface that users are familiar with. Users are guided by the computed subgraph, which is itself updated based on user search activity. We conducted a user study that compared the proposed GRAFS approach against a baseline faceted search system in the context of exploratory literature search. Experimental results show that the proposed approach can effectively help users recognize relationships between key concepts, leading to a more sophisticated understanding of the search topic while maintaining similar functionality and usability as a classical faceted search system.

REFERENCES

- [1] Ioannis Arapakis, Xiao Bai, and B. Barla Cambazoglu. 2014. Impact of Response Latency on User Behavior in Web Search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information*

- Retrieval (Gold Coast, Queensland, Australia) (SIGIR '14). Association for Computing Machinery, New York, NY, USA, 103–112.
- [2] Krisztian Balog, Marc Bron, and Maarten De Rijke. 2011. Query modeling for entity search based on terms, categories, and examples. *ACM Transactions on Information Systems (TOIS)* 29, 4 (2011), 1–31.
 - [3] Nicholas J Belkin. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science* 5, 1 (1980), 133–143.
 - [4] Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. 179–188.
 - [5] Nan Cao, Jimeng Sun, Yu-Ru Lin, David Gotz, Shixia Liu, and Huamin Qu. 2010. FacetAtlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1172–1181.
 - [6] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–336.
 - [7] Claudio Carpineto, Stanislaw Osiński, Giovanni Romano, and Dawid Weiss. 2009. A survey of web clustering engines. *ACM Computing Surveys (CSUR)* 41, 3 (2009), 1–38.
 - [8] Manhong Dai, Nigam H Shah, Wei Xuan, Mark A Musen, Stanley J Watson, Brian D Athey, Fan Meng, et al. 2008. An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics* 21 (2008).
 - [9] Jeffrey Dalton and Laura Dietz. 2013. Constructing query-specific knowledge bases. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*. 55–60.
 - [10] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. 365–374.
 - [11] Laura Dietz. 2019. ENT Rank: Retrieving entities for topical information needs through entity-neighbor-text relations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 215–224.
 - [12] Michelle Dowling, Nathan Wycoff, Brian Mayer, John Wenskovitch, Leanna House, Nicholas Polys, Chris North, and Peter Hauk. 2019. Interactive visual analytics for sensemaking with big text. *Big Data Research* 16 (2019), 49–58.
 - [13] Faezeh Ensan and Feras Al-Obeidat. 2019. Relevance-based entity selection for ad hoc retrieval. *Information Processing & Management* 56, 5 (2019), 1645–1666.
 - [14] Jody Condit Fagan. 2010. Usability studies of faceted browsing: A literature review. *Information Technology and Libraries* 29, 2 (2010), 58–66.
 - [15] Marti Hearst. 2009. Information Visualization for Search Interfaces. In *Search User Interfaces*. Cambridge University Press, Chapter 10.
 - [16] Marti A Hearst. 2006. Clustering versus faceted categories for information exploration. *Commun. ACM* 49, 4 (2006), 59–61.
 - [17] Orland Hoerber. 2018. Information visualization for interactive information retrieval. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 371–374.
 - [18] Tom Hope, Jason Portenoy, Kishore Vasani, Jonathan Borchardt, Eric Horvitz, Daniel S Weld, Marti A Hearst, and Jevin West. 2020. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. *arXiv preprint arXiv:2005.12668* (2020).
 - [19] Fahd Husain, Rosa Romero-Gómez, Emily Kuang, Dario Segura, Adamo Carolli, Lai Chung Liu, Manfred Cheung, and Yohann Paris. 2021. A Multi-scale Visual Analytics Approach for Exploring Biomedical Knowledge. In *2021 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*. IEEE, 30–35.
 - [20] Clement Jonquet, Nigam H Shah, and Mark A Musen. 2009. The open biomedical annotator. *Summit on Translational Bioinformatics 2009* (2009), 56.
 - [21] Peter Kraker, Christopher Kittel, and Asura Enkhbayar. 2016. Open knowledge maps: Creating a visual interface to the world's scientific knowledge based on natural language processing. *027.7 Zeitschrift für Bibliothekskultur / Journal of Library Culture* 4, 2 (2016), 98–103.
 - [22] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.
 - [23] Bill Kules and Robert Capra. 2009. Designing Exploratory Search Tasks for User Studies of Information Seeking Support Systems. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (Austin, TX, USA) (JCDL '09)*. Association for Computing Machinery, New York, NY, USA, 419–420. <https://doi.org/10.1145/1555400.1555492>
 - [24] Bill Kules and Ben Shneiderman. 2008. Users can change their web search tactics: Design guidelines for categorized overviews. *Information Processing & Management* 44, 2 (2008), 463–484.
 - [25] Yafeng Lu, Hong Wang, Steven Landis, and Ross Maciejewski. 2017. A visual analytics framework for identifying topic drivers in media events. *IEEE Transactions on Visualization and Computer Graphics* 24, 9 (2017), 2501–2515.

- [26] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics* (2022).
- [27] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.
- [28] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.
- [29] Sean Munson, Stephanie Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of The International AAAI Conference on Web and Social Media*, Vol. 7. 419–428.
- [30] Xi Niu, Xiangyu Fan, and Tao Zhang. 2019. Understanding faceted search from data science and human factor perspectives. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 1–27.
- [31] Arlind Nocaj and Ulrik Brandes. 2012. Organizing search results with a reference map. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2546–2555.
- [32] Don Norman. 2013. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books.
- [33] Stanislaw Osiński, Jerzy Stefanowski, and Dawid Weiss. 2004. Lingo: Search results clustering algorithm based on singular value decomposition. In *Intelligent Information Processing and Web Mining*. Springer, 359–368.
- [34] Peter Pirolli and Stuart Card. 2005. 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. McLean, VA, USA, 2–4.
- [35] Ridho Reinanda, Edgar Meij, Maarten de Rijke, et al. 2020. *Knowledge graphs: An information retrieval perspective*. Now Publishers.
- [36] LT Rodrygo, Craig Macdonald, and Iadh Ounis. 2015. Search result diversification. *Foundations and Trends in Information Retrieval* 9, 1 (2015), 1–90.
- [37] Bahareh Sarrafzadeh and Edward Lank. 2017. Improving exploratory search experience through hierarchical knowledge graphs. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 145–154.
- [38] Michael Schuhmacher, Laura Dietz, and Simone Paolo Ponzetto. 2015. Ranking entities for web queries through text and knowledge. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 1461–1470.
- [39] Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (2014), 443–460.
- [40] Ben Shneiderman. 1992. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics (TOG)* 11, 1 (1992), 92–99.
- [41] Mark D Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 623–632.
- [42] Axel J Soto, Piotr Przybyła, and Sophia Ananiadou. 2019. Thalia: semantic search engine for biomedical abstracts. *Bioinformatics* 35, 10 (2019), 1799–1801.
- [43] Huan Sun, Hao Ma, Wen-tau Yih, Chen-Tse Tsai, Jingjing Liu, and Ming-Wei Chang. 2015. Open domain question answering via semantic enrichment. In *Proceedings of the 24th International Conference on World Wide Web*. 1045–1055.
- [44] Benjamin E Teitler, Michael D Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. 2008. NewsStand: A new view on news. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 1–10.
- [45] Dejan Todorovic. 2008. Gestalt principles. *Scholarpedia* 3, 12 (2008), 5345.
- [46] Daniel Tunkelang. 2009. Faceted search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1, 1 (2009), 1–80.
- [47] Kelsey Urgo and Jaime Arguello. 2022. Learning assessments in search-as-learning: A survey of prior work and opportunities for future research. *Information Processing & Management* 59, 2 (2022), 102821.
- [48] Svitlana Vakulenko, Ilya Markov, and Maarten de Rijke. 2017. Conversational exploratory search via interactive storytelling. *arXiv preprint arXiv:1709.05298* (2017).
- [49] Nees Jan Van Eck and Ludo Waltman. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 84, 2 (2010), 523–538.
- [50] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.
- [51] Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research* 47, W1 (2019), W587–W593.
- [52] Ryen W White, Bill Kules, and Ben Bederson. 2005. Exploratory search interfaces: categorization, clustering and beyond: report on the XSI 2005 workshop at the Human-Computer Interaction Laboratory, University of Maryland. In *ACM SIGIR Forum*, Vol. 39. ACM New York, NY, USA, 52–56.

- [53] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1, 1 (2009), 1–98.
- [54] Travis Wolfe, Mark Dredze, and Benjamin Van Durme. 2017. Pocket knowledge base population. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 305–310.
- [55] Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th International Conference on World Wide Web*. 1271–1279.
- [56] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. 2003. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 401–408.
- [57] Oren Zamir and Oren Etzioni. 1999. Grouper: a dynamic clustering interface to Web search results. *Computer Networks* 31, 11-16 (1999), 1361–1374.