12-9-2021 2:00 PM

A Visual Analytics System for Rapid Sensemaking of Scientific Documents

Amirreza Haghverdiloo Barzegar, *The University of Western Ontario*

Supervisor: Sedig, Kamran, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science

Abstract

With the rapid growth of scientific documents over the years, researchers must examine large collections of documents to keep up with their research fields. Over the past years, numerous tools have been developed to support researchers in making sense of the documents collection; however, due to the high load and complexity of scientific information, many of these tools have only covered basic tasks or restricted information items. This thesis describes a visual analytics system (i.e., a tool that integrates data visualization, human-data interaction, and machine learning) that helps researchers explore and examine scientific documents thoroughly and rapidly with an especial focus on the textual content of scientific documents. Through a usage and comparative scenario, we illustrated the efficiency and advantages of our system over similar tools. Finally, we discussed possible future extensions and upgrades thanks to the modular architecture of the system.

Keywords

Visual Analytics Systems, Sensemaking, Scientific Documents, Interactive Visualizations, Human-Information Interaction

Summary for Lay Audience

With the rapid growth of scientific documents and interdisciplinary studies conducted in the past years, researchers have found it challenging to remain up to date with their research fields. Search and exploration, filtering, reading, and extracting information items, and comparison are common tasks that researchers perform to make sense of the collection of documents they are working with. Many computational tools have been developed over the years to support researchers in performing these tasks; however, they often support some of these tasks or just specific information items of scientific documents (e.g., exploration of bibliographic information of document collections). On the other hand, due to the complex structure and high load of scientific information, the visualization techniques used in the visual interface of the existing tools require researchers to perform extra interactions with the tool to access their desired information.

This thesis describes a visual analytics system (i.e., a tool that integrates machine learning algorithms with data visualization and human-data interaction) with an innovative visual interface design to afford rapid sensemaking of scientific documents for researchers. By combining an integrated visualization component and advanced text analytical approaches, we have managed to design a system that not only encodes a broad spectrum of scientific information items (ranging from bibliographic information to derived attributes of textual content of scientific documents) but also supports a wide range of activities during sensemaking process (e.g., rapid exploration, skimming, comparison). As a proof of concept, we have provided a scenario in which we examine the efficiency of different components of our system compared to the existing tools. Last but not least, we analyzed the limitations and future extensions of our system.

Acknowledgments

I want to express my gratitude and appreciation for my wonderful supervisor, Dr. Kamran Sedig, for his consistent support and guidance throughout this research project. I am sincerely grateful for your dedication, fatherly pieces of advice, and support of my decisions.

I am also thankful to my parents, Mahnaz and Davood, who helped me be where I am by their selfless sacrifices and vision.

Finally, I offer my deepest gratitude to my fiancée, Mona. She supported me patiently and warmed my heart in each moment.

Table of Contents

List of Figures (where applicable)

Chapter 1

1 Introduction

Scientific documents are a specialized type of literature considering their topics, intended audience, content and presentation methods [1]. They contain original research results or reviews of current results, and the primary reason for their publication is to share knowledge and results with peer scholars to make research and development progress. Scientific documents can be referred to with different titles such as academic/scientific articles, academic/scientific publications, academic/scientific papers, etc. To avoid any ambiguity, we will refer to them only with “scientific documents.”

One of the most common types of scientific documents is “research papers,” which describe significant advancements in a field of research. Originality, novelty, quality of scientific content and contribution to the existing body of knowledge are the critical aspects of a research paper and assessed by peer reviewers. Review articles are another type of scientific documents which provides synthesized summaries of existing research in a specific research domain. Other than these two common types, [2] mentions several other forms of scientific publications.

Academic search engines and bibliographic databases are the familiar places from which we can access scientific documents [3]. We can refer to Google Scholar and Microsoft Academic as examples of large crawler-based search engines which allow academic users to access a significant resource of publicly available documents. It is important to mention that the number of publications has increased exponentially over the last decades [4]. Therefore, it is challenging for researchers to keep up to date with a field of research or find relevant works among the increasing number of topics and documents.

Researchers might need to carry out various activities in order to find their desired scientific documents and conduct their studies. Direct search-based approaches or exploratory searches are considered as one of the initial steps. When researchers are unfamiliar with a research domain or have not planned out their search strategies, their initial step is exploratory searches rather than direct document retrieval [5]. During the process of searching, some implicit minor tasks such as comparison, knowledge acquisition, aggregation, navigation, analysis or evaluation of search results might be performed.

Finding related works and scientific documents is not limited to searching in academic databases. Each of the retrieved and selected search results should be analyzed further by the researchers. To conduct a complete literature study, find a theoretical basis to support a topic, judge the relevancy of certain scientific documents to a specific topic, or aim for other goals, researchers need to drill into the content of each selected scientific document to be able to extract the required information.

To explain the overall interactions between researchers and scientific documents in more detail, we can refer to the process of conducting research synthesis. *Research synthesis* is a general term defined as the process of combining multiple primary research results aimed at one topic or scope of study [6]. There are numerous types and methodologies to conduct such studies; however, we will only mention the common framework used to conduct scoping reviews. *Scoping review* is a relatively new review type that tries to map literature, key concepts, gaps in the research, and types of evidence related to a defined area or field [7], to elaborate on the required interactions with scientific documents. This framework [8] proposes a six-stage process consisting of 1) identifying the research question, 2) searching for relevant studies, 3) selecting studies, 4) charting the data, 5) collating, summarizing and reporting the results, and 6) consultations. It is apparent from the title of each stage that researchers are required to make sense of the content and the results proposed in each study to decide on the inclusion/exclusion of them, extract their data and generate reports in order to conduct such studies. In the next section, we will focus on the general process of sensemaking and sensemaking of scientific documents.

1.1 Sensemaking and Scientific Documents

Sensemaking is a general term that suggests the active processing of information to achieve understanding. Sensemaking activities happen whenever we confront a new, complex problem, and they involve finding information about the problem and require learning about new domains, exchanging knowledge, acquiring situation awareness, and solving ill-structured problems [9], [10].

Sensemaking can be considered as a continuous effort to understand connections among entities such as people, places, or events in order to anticipate their direction and act accordingly. Sensemaking activities do not have a clear beginning and ending all the time; instead, they are usually characterized as ill-structured and open-ended problems where potentially conflicting pieces of information should be gathered and understood. Therefore, sensemaking does not follow the waterfall model of data to understanding (data – information – knowledge – wisdom), nor is it a function of the amount of information where more information could guarantee a better sensemaking process.

To understand the process of sensemaking in more depth, we can refer to the definition proposed by [11], where sensemaking is considered as a process of searching for a representation and answering task-specific questions by encoding data in that representation. This definition consists of two essential components, representations and task-specific questions. Considering the first component, it considers two types of resources involved in the process of sensemaking, internal cognitive resources and external resources for information processing and storage; and it mentions representation as tools by which we can reduce the cost of using these resources during different operations of sensemaking. Regarding the second component, this definition is concerned with problems involving large amounts of information and sensemaking is involved in their information retrieval and processing tasks and subtasks to answer potential questions. In conclusion, sensemaking activities involve establishing some goals, discovering the structure and type of involved information items, generating required questions, and organizing corresponding answers to the raised questions.

Since many academic goals such as conducting a review study also involve large amounts of information, the perspective expressed in the above definition can help analyze the sensemaking of scientific documents in more depth. In a broad conceptualization, the sensemaking process consists of 4 main sub-processes: 1) information gathering, 2) encoding the information in a new representation, 3) gaining insight through manipulation of the new representation, and 4) creation of some knowledge based on the resulted insight. Consequently, we can infer two key features of sensemaking processes: 1) the dynamic flow of data in a sensemaking process in a sense that the initial external data items might potentially be removed or reformed during the process, and 2) iterative nature of the sensemaking process in which the sub-processes can come with many back loops and occur one after the other to guarantee the appropriateness of examined information and the developed mental model. Considering the abovementioned features, we should examine the sensemaking sub-processes in the context of scientific documents thoroughly.

Initially, researchers search through external data sources, filter relevant documents, and store them for further processing to gather the required information. Next, they read the stored documents and extract desired information items to find relevant evidence to draw inferences, support, or reject a theory. Depending on the size and complexity of extracted evidence, in the next step, researchers re-represent the information in their mental space or using an external tool (e.g., computational tools). Generating hypotheses, considering biases, and presenting the resulted theory are the follow-up activities.

According to the above examination of sensemaking sub-processes, it can be concluded that through the process of sensemaking, researchers seek information. However, sensemaking activities are different from information-seeking activities in which the goal is to find a specific body of information. In a sensemaking process, desired information items are distributed among documents and data sources, and researchers have to extract the information items, find their relationships and make sense of these information pieces [12].

As shown above, the sensemaking of scientific documents consists of various sub-activities that can be analyzed to better understand this process. Searching and filtering retrieved scientific documents from external data resources, reading scientific documents and extracting desired information, discovering the relations between extracted information items and their corresponding documents, and collating and synthesizing extracted information scattered in a corpus of documents are the major sub-activities carried out during the process of sensemaking.

As mentioned earlier, academic databases and search engines are the familiar data sources through which researchers search for their desired scientific documents. Many exploration systems provide an interface to a specific repository or an integration of multiple sources of scientific documents, provide rankings for authors, publications, or other aspects of the documents, or categorize the documents into research topics by extracting keywords. However, supporting researchers in searching and navigating through different dimensions (e.g., topic, co-author, etc.), discovering research trends, and relating authors and scientific documents semantically have remained a challenge for many of these systems [13]. By using these databases, researchers can retrieve scientific documents by searching for specific keywords, research domains, publications, authors, etc. Afterwards, they should drill into each retrieved document to judge its relevancy, decide its inclusion/exclusion for further analysis, and extract key information.

The process of making an initial judgement of a document's relevancy to the researcher's information needs is referred to as "document triage" [14]. During document triage, researchers examine a document and determine its relevancy. However, multiple revisions might occur during the process, and the perceived relevancy of one document might rise or fall as the researcher comprehends more pieces of the document. Previous studies indicate that abstracts and titles of scientific documents play an important role in the relevance judgement of researchers during document triage. It is also reported that section headings, emphasized text (e.g., italicized paragraphs, bullet points, figure captions), images and figures are the other document elements that researchers focus on during the triage [15]. In conclusion, document triage can be seen as a form of visual search where documents are explored and assessed in short timespans and commonly read linearly. Therefore, in the case of long documents, researchers tend to read only the overview sections at the top, which results in neglecting possible matching content with lower visibilities.

Depending on the information needs and research objectives (e.g., looking for the definition of a concept, finding supporting and testing data of an idea, or comparison of scientific conclusions of several documents), researchers might perform visual searching tasks to find essential features of a scientific document such as its title, section headings, and abstract, skim a document in a quick fashion, or read it more thoroughly to look for inner textual information. It is also important to mention that researchers often work on different scientific documents simultaneously to search, compare, arrange, link, annotate and analyze their content [16]. Furthermore, they are generally looking for specific sections of a scientific document (e.g., definitions, hypothesis, algorithms, measurements) that can be found in any part of the document’s text instead of its entire content.

Once a collection of scientific documents is retrieved and filtered and the required information items are extracted, researchers can discover the relationships within the corpus. The relationships between two scientific documents can be categorized into two major groups: 1) bibliographic and 2) content-based semantical relationships. If two documents are connected bibliographically, they are likely semantically close to each other; however, the semantic closeness of two documents cannot guarantee their citation-based relationships. Current academic databases provide bibliographic information of scientific documents; in addition, citation-based recommendation and analytics systems (e.g., [17], [18]) support researchers to drill into the citation patterns, discover trends, and track the development of a research domain. On the other hand, extracting semantical relationships are more difficult for several reasons:

1. The information needs of one researcher vary from the other depending on the researchers' academic tasks or experience.
2. Any part of the content of scientific documents can potentially answer a researcher's information requirements, unlike bibliographic information located in a specific section of documents.
3. Covering all the content of a scientific document, extracting key information and matching it with another document is much more time-consuming than extracting bibliographic information.

Summarization, keyword extraction, sentence extraction and other NLP techniques are examples of proposed solutions to support researchers in making sense of documents' content and discovering inter-document links.

In conclusion, as mentioned before, the subprocesses during the sensemaking activities are iterative. To cite an example, a researcher can search for a different key term, extract a new set of scientific documents, and add to the existing corpus after triaging the initial set of documents and figuring out the need for more supporting documents for a subtopic. Furthermore, it is important to mention that external representations play an important role in sensemaking activities. As working memory has some limitations which restrict researchers from processing a large number of information items, the relationships among them and hypotheses derived from them simultaneously, offloading some of the information onto external representations can expand the capacity of working memory [19]. However, in the case of scientific documents, they might contain implicit and internal layers of meaning, patterns and structures that cannot be encoded in these external representations easily [12]. Therefore, we will study the role of external representation and their interactivity during the sensemaking process of scientific documents in more depth in the next section.

1.2 Sensemaking and External Representations

During the process of sensemaking, external representations can enhance cognitive power in various ways. They are considered as sharable objects of thought which enable re-representation of information and construction of complex structures. In addition, they facilitate the computation of explicit encoding of information and consequently reduce the cost of controlling thought [20]. Depending on a specific task, sensemaking can be considered as a process of developing more sophisticated representations to organize information, and if the information does not fit into one representation, a search is undertaken to find a better representation for encoding the information. From this perspective, sensemaking involves a constant exploration of new representations [11], [21]. Although external representations enhance the process of sensemaking and their coupling with internal representations from an integrated cognitive system [22]; however, it is through **interactions** that they are able to support sensemaking.

Interactions with external representations can be categorized into two major groups: 1) the set of operations performed when **using** external representations (e.g., rearranging retrieved scientific documents), and 2) the set of operations performed to prepare for using external representations (e.g., retrieving a set of scientific documents). Interactions can be seen as means to reduce the cost of projection and imagining by creating external structures and supporting more complex projection and more advanced computation, leading to deeper and broader sensemaking [20]. In conclusion, external representations and interactions are two important pillars of the sensemaking process.

Visualizations are one of the most common means of representing and interacting with information. In general, visualization is the process of transformation of data *D*, according to the specifications *S*, into a time-varying image *I(t)* [23]. Each element of the above definition should be considered in its broadest sense. Therefore, a spectrum ranging from a blinking LED to a complex virtual-reality setup can be referred to as visualizations. Visualization can be categorized into two major groups of static and interactive visualization; however, static visualizations can support simple tasks including a limited amount of information, as perceptual analysis of visual attributes within a static visualization does not guarantee complete sensemaking of information [24]. Since most academic tasks and their corresponding sensemaking processes involve complex and large amounts of scientific data and documents, interactive visualization systems are the potentially suitable external representations to support sensemaking activities.

The combination of automated analysis techniques with interactive visualizations constructs "visual analytics systems," which enable effective understanding, reasoning, and decision making based on complex large data sets [25]. In the context of scientific documents, there has been an increase in the development of visual analytics systems in recent years that support the search and analysis of documents. Various data types can be extracted from a corpus of scientific documents ranging from the textual content, which is the central component and encodes the main scientific contribution in natural language, to the metadata of each document and citation-based inter-document connections. Many of the current visual analytics systems only focus on a subset of these data types or activities during the sensemaking process. To cite an example, [26] only visualizes each document's derived topics and concepts in a 2-D visualization canvas. Therefore, textual content, figures and tables, bibliographic information and other important information of the documents should be extracted from other sources. This lack of ability to penetrate the inner layers of information may result in gaps between inner mental processes and external representations [27].

On the other hand, encoding a large subset of information derived from a corpus of scientific documents may result in researchers' inability to navigate properly through the visualization(s) and answer the information needs during the sensemaking process. Reading textual content of documents, discovering semantical relationships within a corpus, comparing several documents and articulating a hypothesis about the studied scientific documents are time-consuming activities and challenging to be supported by visual analytics systems in this domain. Visual analytics systems (VASes) should remove distracting information and provide precise and clear visualizations with low latency to enable researchers to explore inter-document relationships and drill into each document within the corpus. However, overload of scientific information, the closeness of scientific literature and complex semantic analysis, and limitations of the devices supporting VASes are the main challenges for providing such experience for researchers [28].

Developing a citation network of a corpus, single document and multi-document textual summarization, topic extraction, and document clustering are examples of techniques used to support the analysis of scientific document collections by VASes. Depending on the visualization(s) representing this processed information, researchers can conduct their academic studies and make sense of the involved scientific documents in a quicker way.

1.3 Research questions

Many VASes provide multiple coordinated visualizations in separate windows or tabs to cover different tasks with their system and link different information items within a corpus to each other. Therefore, there is a possibility that researchers forget about the context of the corpus when they try to focus on the provided information about a single document or perform any other required navigation to discover detailed information. The gap between multiple visualizations may cause a disconnection between researchers' internal representation and external representations provided by a VAS and delays the process of sensemaking. The abovementioned challenges were the main motivations to examine the possibility of developing a VAS that can support rapid sensemaking of scientific documents by covering a large subset of sensemaking sub-activities based upon the scientific information provided in a corpus encoded in an innovative integrated visualization.

The research questions that this thesis examines are as follows:

1. Is it feasible to integrate machine learning and natural language processing techniques, data visualizations, human-data interactions and web API infrastructures to develop an online VAS that supports the sensemaking of scientific documents?

2. Can such a VAS cover a large subset of scientific information and sensemaking sub-activities within a corpus in a rapid fashion and still avoid researchers' confusion during their interaction with the system?

3. Can such a VAS be scalable in the sense that different machine learning techniques approaches be plugged in to support a specific functionality?

4. What are the essential design considerations to develop such a system?

Chapter 2

2 Background

In the previous chapter, we examined scientific documents and their importance in conducting academic research. We also analyzed the sensemaking process of scientific documents and the corresponding sub-activities during the process. Finally, we elaborated on the role of external representations and visualizations in supporting the process of sensemaking and explained some of the challenges in developing suitable visual analytics systems (VASes) in the domain of scientific documents.

There has been a significant increase in the number of scientific publications recently. This rapid growth can be observed in the large amounts of information that scientific databases and online libraries hold. To exemplify, Elsevier and Scopus offer more than 16 million documents from 2,500 journals. In addition, more and more scientific documents are published online as “open-access” content [28], [29]. The emergence of cross-disciplinary sources and the growing number of scientific documents results in a challenging situation for researchers to follow the research trends and recognize the key documents. Therefore, adopting traditional methods like reading numerous documents is time-consuming and subjective and causes researchers’ inability to address their research questions and information needs. As mentioned earlier, encoding large amounts of scientific information using visualizations can enhance researchers’ cognitive power during the sensemaking process. Thus, this thesis examines the possibility of developing a VAS that supports the rapid sensemaking of scientific document collections and tackles the previous challenges of similar systems.

Since our proposed VAS focuses on the rapidity of the sensemaking process, in section 2.1, we will study researchers’ reading behaviors which is the most time-consuming activity during sensemaking and helps researchers extract key information, link, and compare scientific documents. In section 2.2, we will dissect VASes, cognitive activities, and human-data interaction basics. Visual approaches used in visualizing scientific information will be reviewed in section 2.3. In the last section of this chapter, we will analyze some of the related works in this domain.

2.1 Reading behavior during sensemaking process

Researchers spend an important proportion of their working time on reading scientific documents, and studies show that the number of documents a researcher reads annually has increased impressively [30]. The methods adopted for reading scientific documents vary depending on researchers' career stage, familiarity with scientific language, the medium by which the documents are read (paper-based or digital), area of research, and other factors. To define the reading behavior of researchers in a domain, it is important to rely on surveys; therefore, we will review the result of surveys conducted in previous studies. Although printed documents are still commonly used by researchers for direct annotation, portability, tangibility, navigation, and comprehension [31], we will focus on reading behavior and needs in digital environments as our proposed VAS will also provide scientific information in a digital environment. Furthermore, relying on electronic sources has increased thanks to easier access and interactive possibilities recently.

In terms of the motivations, researchers read scientific documents to discover other studies in their domain, find additional sources of references for their work, remain informed, compare different studies, figure out different research methodologies, and find additional ideas for their research [16]. By breaking down scientific documents into their fragments (e.g., introduction, methodology, result, conclusion, etc.), we can analyze researchers’ reading behavior more thoroughly. When researchers try to answer a specific question related to a concept, they will have to find all the definitions and descriptions of that concept across different fragments in different documents, read them, classify them, compare them, and make implicit links between them. It is also studied that perception of scientific documents shifts throughout academic careers. While early-career researchers find fragments related to methodology and results challenging to understand, senior researchers focus more on those fragments [16], [32].

Focusing on specific fragments of a scientific document leads to a non-linear type of reading which is the typical academic reading workflow. As scientific documents often have similar structures, researchers expect to find their information needs in specific fragments and navigate to them directly [31]. This strict structural order of text allows researchers to adopt their own reading plans, jump, skim, read in a discontinuous fashion, and still make sense of the text [33]. In general, digital environments support this type of reading more than linear and concentrated type as the links in web and hypertext literature provide multiple options to navigate in text. Thus, screen-based reading is often centered around browsing and scanning, locating keywords, and one-time, selective, and non-linear reading rather than in-depth and concentrated reading [34].

Based on the previous surveys, it is important to mention some design and development considerations of scientific document digital reading applications to enhance researchers' experience. Besides non-linear reading, which can be supported by content indexing, hyperlinking, and hierarchal navigation, annotation is also a fundamental component of the academic reading process. Annotations are goal-oriented activities that occur within the context of a document without the need of switching to other tools and enable researchers to scan for them by visual searching [35]. Automatic annotation and affording proper interactions to annotate scientific documents can enhance researchers' reading experience effectively. Furthermore, reflowable adaptive layout using a different presentation than the original one (e.g., a single-column scrollable text stream) and contextual literature where researchers can interact with a collection of documents are other demands mentioned in the survey conducted by [31].

Parallel reading, active reading, and skimming are specific terms used to refer to different reading behaviors when working with scientific documents and need further inspection here. Parallel reading refers to the strategy where researchers work with different documents in parallel. On the other hand, active reading refers to reading with a critical thinking and learning approach, which is an important aspect of education [36]. Finding related resources and navigating through them is vital for active reading; however, in many digital reading applications, researchers have to pause their reading activity to navigate through other documents within the corpus, which results in either losing focus on the document under examination or losing the context of the corpus [35], [36]. At last, skimming is a rapid form of reading that is selective and non-sequential and helps researchers get a gist of a scientific document, discover its structure and look for specific information in a quick fashion. Although skimming is commonly used to support covering the increasing amount of scientific information, its efficiency is heavily dependent on researchers' visual interactions with text to spot keywords, navigate appropriately through the document and find relevant sentences [37].

In conclusion, we reviewed some of the reading strategies and skills, the importance of fragments of scientific documents, general reading behaviors, and shortcomings of existing digital reading applications in this section. In the next section, we will examine visualization patterns and techniques, typology of scientific information, and visual approaches for encoding this information in more depth.

2.2 Visual approaches for encoding scientific information

Information encoding is defined as the process of converting information from one state to another. However, in the context of information visualization, mapping information items from an information space to a representation space (i.e., the space of perceptual visual forms of information items) refers to information encoding [38]. In order to review the existing visual approaches for encoding scientific information, it is important to clarify the differences and definitions of some overlapping concepts like visualization techniques and patterns. For this purpose, we will refer to the framework proposed by [38] in which visualization techniques are considered as "methods or templates that can be used in specific visualization design contexts" such as node-link diagrams, which are used to encode the connections within a network of entities in most cases. In contrast, visualization patterns are not confined to specific contexts of use as they exist in a consistent level of abstraction independent of any particular technology, platform, medium, or domain. In other words, visualization patterns are abstractions rather than visualizations (e.g., the concept of branches that can be used to represent diverging and converging features of information).

The scientific information space contains various entity types, including scientific documents, authors, venues, institutions, events, and fields of study [39] and the goal of visualization is to represent the relations between these entities [40]. Prior to encoding scientific information, it is important to make sure corresponding data items are correctly extracted. The first set of data extracted from academic data sources is the raw data containing the information about authors, title, abstract, keywords, venue, publisher and other metadata about the scientific documents, which is important for supporting their management and developing related system [40]. In the next stage, the extracted raw data should be filtered, processed, and transformed into clean and structural data to meet the requirements for visual encoding [41].

Processing raw data of a scientific document, constructing its derived attributes, and encoding them give researchers insights that require a lot more cognitive load to gain without the help of visualizations. To cite an example, by performing text processing and mining techniques such as tagging, text clustering, dimension reduction, and topic modeling and representing the extracted topics from scientific documents, visual analytics systems can help researchers gain insight into the evolution trend of a topic in a research domain [41]. Other than the bibliographic information of scientific documents, which help researchers discover the relationships within a corpus, follow research trends, and identify important documents, the textual content is the central component of a scientific document and is usually stored in PDF, XML, or HTML formats in scientific databases. Since making sense of the textual content of documents and extracting information from them are time-consuming and affect the rapidity of the sensemaking process the most, we will focus more on the visual approaches adopted for encoding textual content or related derived attributes.

It is studied that approaches toward text analysis and visualization have increased recently, unlike the previous decade where visualizations were more focused on the authors and bibliographic information [42]. Many approaches for encoding scientific information integrate visualization techniques with natural language processing techniques, which will be studied thoroughly in the next chapters. Encoding textual content of scientific documents can be used to support different activities ranging from finding and comparing documents to extracting patterns and relations. We will refer to some existing visualizations and study the techniques they used below:

- CyBis [43] is a 3D analytical interface for a bibliographic visualization tool with the objective of enhancing the scientific document searching experience. It uses **metaphors** for representing both documents and terms, where relevant documents to a search query are scattered as circles in a **3D cylinder** based on their publishing year and closeness to a specific term. It also supports standard 3D interactions to deal with possible occlusions.

- Rexplore [13] is a tool for exploring scientific data which supports visualizing trends, finding the semantic relationships between authors, and multi-dimensional search. **Node-link graph** is the primary technique used in this tool to encode the connection between authors, and the interactivity of the system allows its users to change the type of links, ranking criteria, and filters.

- ParallelTopic [44] is another tool that integrates Latent Dirichlet Allocation (LDA) topic modeling with interactive visualization to support the sensemaking of large text corpora. Other than the **scatterplots**, which show the number of topics each document contains, **parallel coordinate metaphor** is also utilized in this tool to present the probabilistic distribution of a document across topics. Furthermore, **word clouds** corresponding to each topic and **streamgraphs** for depicting the topic evolution over time are the techniques used in this tool.

- The visualization method proposed in [45] is based on a hierarchal topic model which extracts topics of three different domains and their correlation. It utilizes a **Sankey diagram** for presenting the topics, a **scatter plot** for showing the topics in 2D space, **word clouds** to show the corresponding terms of selected topics and subtopics, and a **stream diagram** to illustrate the topic trend.

- The visual exploration approach proposed in [46] uses **maps**, a popular visualization **metaphor** for scientific information, to encode the scientific documents in the computer science domain. It encodes the words and phrases extracted from titles to cities in the map; therefore, countries are created based on their similarity. It also uses the **heatmap** technique to visualize the profile of researchers, research institutions, or conferences over the base map.

- TileBars [47] is a visualization that utilizes term distribution information for Boolean search results. In the first place, it uses the text tiling technique to partition text into coherent thematic segments. In the next step, it represents a list of documents encoded into **rectangular bars** consisting of several squares that correspond to text segments. While the length of a bar shows a document's length, the darkness of a square indicates the frequency of search terms in a segment.

- The visualization proposed by [48] displays search results in a **matrix view** such that each document in the results set belongs to a single **cell** and is represented by an **icon**. Through interaction, the axis of the matrix view can be set based on different aspects of the document (e.g., author, year of publication, or relevancy).

- Sparkler[49] is a prototype system that visualizes the relevancy of multiple queries to a collection of documents. It uses a **circular spatialization scheme** to afford the comparison of multiple queries where resulted documents are distributed on a circle split into one segment per query. Documents are encoded by colored glyphs according to the query, and queries are encoded with triangle icons positioned in the center of the visualization. Furthermore, distance from the center encodes the relevancy to a query and documents with similar relevance values are positioned the same distance from the query icon but spread along an arc, resulting in the formation of a simple **histogram**.

- PaperVis[50] is a visualization that affords quick grasping of complex citation-reference structures among a scientific document collection. **Radial Space Filling**, **Bullseye View**, and **node-link graphs** are used in this visualization to depict the relationships between documents.

Aside from the abovementioned techniques, it is important to mention the visualization techniques and approaches that researchers adopt to address the challenges regarding complex datasets. Due to the complexity and high load of scientific data, visualization designers might utilize concepts such as *overview+detail*, *focus+context*[51], or multiple coordinated views[52] aiming to help users interactively change the perspective on the data. **Interactive lenses**, tools that solve a localized visualization problem by temporarily altering a part of the visual representation of data[53], are an important visualization approach to support interactive complex data exploration. Interactive lenses can be conceptualized as functions, which determine what is to be processed by them and how the result should be integrated with the visualization. The key characteristic of interactive lenses is their transient effect, which means that the visualization can return to its original state once the lens is dismissed.

Designing visualizations to encode scientific information and documents is not limited to the techniques utilized in the abovementioned examples. Based on the sensemaking sub-activities and academic tasks a visualization aims to support, different visualization techniques and patterns can be blended to encode required information items. Many other existing visualization techniques applied to encode textual content are mentioned in a web-based interactive visual survey[1] proposed by [54]. This survey has provided around 440 visualization techniques and samples spanning from 1976 to 2019 and categorized them according to different criteria such as supported analytics tasks, supported visual tasks, types of encoded data, and utilized visualization techniques.

To summarize, we had an overview of the basic concepts of visualizations (e.g., visualization techniques, patterns, representation space, etc.) and described different entities of scientific information and the importance of preprocessing before mapping them into representation space. At last, we provided some examples of the common visualization techniques used in this domain. In the next section, we will study the role of

[1] The online survey browser is available at: http://textvis.lnu.se/

interactions and interactivity of a VAS in supporting cognitive activities in more depth and conceptualize the structure of a VAS.

2.3 Interactions, Interactivity, and VAS structure

Although adopting suitable visualization techniques to encode specific scientific information of VAS impacts its efficiency and usability; however, it is through interactions that researchers can adjust visualization and address their contextual and cognitive challenges. It is also important to clarify the distinction between the concepts of interaction and interactivity as interaction refers to actions and reactions between a VAS and its user while interactivity refers to the quality or condition of interaction; therefore, if the quality of the interactions within a VAS is bad, it will not support cognitive activities of researchers effectively in consequence[55], [56]. The interactivity of a VAS can be analyzed with regards to structural elements of each interaction or the combination of interactions together. Diversity of interactions, harmonious relationships among interactions, appropriateness of interactions to support cognitive activities, and flexibility are some of the interactivity factors that VAS designers should take into consideration[57]. While researchers interact with a VAS, some of the information processing occurs in the internal mental space of researchers, some is offloaded onto the computational processing units of the VAS, and some occur through interactions with the VAS. Therefore, the interaction within a VAS creates a coupling between the VAS and the researchers using it to reduce their cognitive effort and allow them to develop a mental model of the scientific information they are working with [38], [55].

In conclusion, as the design of a VAS, its interaction, and the quality of its interactions determine the distribution of processing load between researchers and the VAS and its efficiency to support researchers’ cognitive activities, interaction design and interactivity should be considered central components of design and development of a VAS. [58] provides a hierarchal perspective of the complex cognitive activities supported by VASes and breaks them down into their constructing sub-activities, tasks, sub-tasks, actions, and events in its proposed framework, EDIFICE-AP. It also provides a catalogue of action patterns in the level of actions and reactions with a VAS, with each pattern characterized in terms of its utility in supporting complex cognitive activities, which can aid VAS designers and developers in the process of designing their VAS's intractability.

Considering the previously mentioned characteristics of scientific information items, the need for processing them before visualization, visual approaches to encode them, the role of interactions and interactivity in developing a VAS, and the analysis of sensemaking process as a complex cognitive activity which researchers carry out to conduct their academic research, we can divide the conceptual structure of VASes into five spaces (based on the conceptual visual analytics structure proposed by [55]):

- **Information space** which contains concrete or abstract sources of scientific information (e.g., available scientific documents on academic databases)
- **Computing space**, which is concerned with storing, processing, and encoding the items from information space (e.g., extracting keywords of a scientific document)
- **Representation space** is an interface that connects the human mind to the information space and contains visualizations of the VAS
- **Interaction space** where users of a VAS can perform actions and receive its reactions
- **Mental space** where internal mental events and operations occur (e.g., memory encoding, induction, deduction, etc.).

Although we referred to some of the existing visualizations and visual analytics systems within the domain of scientific information and documents in 2.2; however, in the next section, we will analyze some related works in this domain which try to cover more sub-activities during the sensemaking process of scientific documents in more depth.

2.4 Related works

As mentioned before, the sensemaking of scientific documents is a complex cognitive activity that requires researchers to analyze both the collection of documents and the inner content of each document to be able to develop a mental model of the involved information. Therefore, a suitable VAS should be able to allow researchers to drill into each scientific document in addition to encoding the whole corpus, inter-document relationships, documents' arrangement, and other collective features to support sensemaking and related sub-activities effectively. In this section, we will review some of the VASes in this domain that meet these criteria:

- Action Science Explorer (ASE) [59] is a prototype system[2] that utilizes reference management, statistics, citation text extraction, single and multiple documents summarizations, document clustering, interactive filtering, and network visualizations. In order to afford rapidity of exploration and sensemaking processes, it relies on Natural Language Processing techniques such as textual summarization and network visualizations; in addition, it provides multiple coordinated windows (e.g., reference management, citation text, multi-document summary, and citation networks window) that researchers can use to adjust the visualizations. Concerning the textual content of each document, it provides a full-text view in addition to the citation context feature using colored highlighting technique. Occlusion and possible high density of document nodes in the network visualization and distributed attention on different windows are some of the factors that might affect researchers' sensemaking process and interacting experience.

- Jigsaw [60] is a VAS that integrates multiple text analysis algorithms (e.g., document summarization, document similarity calculation, document clustering, sentiment analysis, entity extraction, and related entities recommendation) with interactive visualizations to allow researchers to explore scientific documents while sensemaking [61]. It provides scientific information of documents and their entities through multiple distinct visualizations such as 1) a List View including lists of entities and color-encoded relationships between them; 2) a Graph View, which uses a node-link diagram to display the connections between entities and documents; 3) a Document View which displays the textual content of a document with highlighted entities, and 4) a Document Cluster View that represents all the documents within the collection along with the partitions

[2] Further information is available at: http://www.cs.umd.edu/hcil/ase/

generated by manual or automatic document clustering. Scrolling the list views and switching between different views to develop a mental model about the scientific information within the system are some of the factors affecting researchers' sensemaking process.

- Paper Forager [62] is a web-based system that allows researchers to explore scientific documents rapidly. Its visual interface is composed of two main components, interface controls, including some controlling options such as search field or authors' list and main display area, which provides single and multiple documents views. The tooltip technique utilized in this system allows researchers to focus on one scientific document and still be aware of the whole collection. The inner content of the documents has remained in its original format and used both as thumbnails to represent documents within the corpus and "page view" to encode documents' content. Although this technique allows researchers to access detailed content rapidly, resized small thumbnails cannot provide such rapidity when the number of documents increases. Furthermore, text processing techniques and encoding key sections of documents could have played an important role in this system which are currently missing.

- The visual analysis prototype developed by [63] supports the exploration of suspected plagiarism cases. Its visualization is based on bipartite graphs and Sankey diagrams, and it uses a three-tiered approach for exploring cases: 1) an overview that represents the distribution of finding spots across the document, their length, categorization, and their relation to the source document, 2) a glyph-based visualization which demonstrates the relation between the source document and the finding spot, and 3) a side-by-side view that represents the actual text fragments of the source and reviewed documents for required comparisons. Although enabling users to drill into latent information in case of requirement enhances this system's efficiency, using NLP approaches for automatic detection and categorization of plagiarism cases could have significantly impacted the system.

- Literature Explorer[64] is a visual analytics suite that supports interactive document retrieval. It uses a topic mining method to uncover "thematic topics" from a corpus and integrates this underlying topic detection with a set of visual components. It uses a graph view to display the connection between topic keywords, a list view to display relevant documents, a theme river view that uses a stacked area graph to display the evolution of topics over time, and a paper view that shows the metadata of a selected document. Although the simplicity and clarity of the visual components afford easy access and reading of information, lacking textual content of documents to show the context of topics and scrolling structure of documents list are some of the limitations of this system.

- PatViz[65] is a VAS that supports interactive search, exploration, and analysis of patents information. PatViz provides various visual windows (11 views) for encoding the patent information (e.g., *world map* to encode the distribution of patents, *text view* to encode the textual content of the patents, *term cloud* to encode the most frequent terms, etc.). Switching between different views and activating all the views in the system leads to complexity of the system and might affect users' experience.

CiteSpace II[66], CiteRivers[17], UTOPIAN[67], VOSviewer[68], Citeology[69], and CitNetExplorer[70] are some of the other notable works in the literature, and we have considered their advantages and limitations before designing our VAS. By reviewing the literature of the related VASes, it can be concluded that node-link diagrams are a common visualization technique used to encode scientific documents, their relationships, and citation networks. However, the occlusion of nodes and a high number of documents force researchers to perform additional navigational actions to be able to scan the documents or discover specific information they are looking for. Therefore, they might lose the context of documents or focus on a document while navigating through the network of document nodes.

In the next chapters, we will conceptualize the architecture of our VAS, analyze the information space, computing space, representation space, interaction space, and mental space thoroughly. The data sources, data preprocessing, applied statistical and Machine learning approaches will be studied in the next chapter. Furthermore, to address the visualization challenges mentioned in this chapter, we will go through our proposed visualization design in chapter 4.

Chapter 3

3 System Architecture and Data Processing

By having an overview of the literature, it was demonstrated in the previous chapters that since sensemaking is an open-ended activity and scientific information has complex inner structures and patterns, many existing VASes in this domain do not cover all the sub-activities and corresponding required information items during the process of sensemaking of scientific documents. Numerous systems were focused on the initial document exploration and filtering activities or representing bibliographic information, authors' information, and inter-document relationships of a corpus of scientific documents. These VASes can support researchers in developing an initial mental model about the documents they are interacting with. However, they pay less attention to tasks such as drilling into each scientific document, reading its textual content, and extracting key information items to address specific information needs, which play important roles in taking the sensemaking process into its next steps and adjusting researchers' internal representation.

In this chapter, we will examine the architecture of a VAS that integrates Machine Learning and Natural Language Processing techniques with visual approaches to support researchers to not only have an initial exploration through their corpus of scientific documents, but also drill into scientific documents, read their textual content in different modes (e.g., skimming, non-linear, or linear reading), discover key sections of each document, and find semantical relationships among different documents in rapidly. Furthermore, we will analyze the pipeline designed to process incoming scientific documents, the technologies used for its implementation, and the developed modules in detail.

3.1 Data Flow Design

Over the past few decades, most scientific documents have been stored in Portable Document Format (PDF) [71], and most academic databases provide PDF documents in response to researchers' queries. Therefore, we decided to design our data flow pipeline in a way that the PDF version of scientific documents counts as its initial external data source.

We have developed a RESTful web API[72] using Python programming language [73] and Flask framework[74] to process uploaded scientific PDF documents, perform preprocessing tasks, apply ML or NLP techniques, and provide endpoints to send required processed data items. Our VAS uses JavaScript Object Notation (JSON) representation for handling requests and responses. Figure 1 depicts an overview of all our VAS architecture, the stages, and modules within the pipeline, starting from the external data source and its corresponding parsing unit to all the provided endpoints.

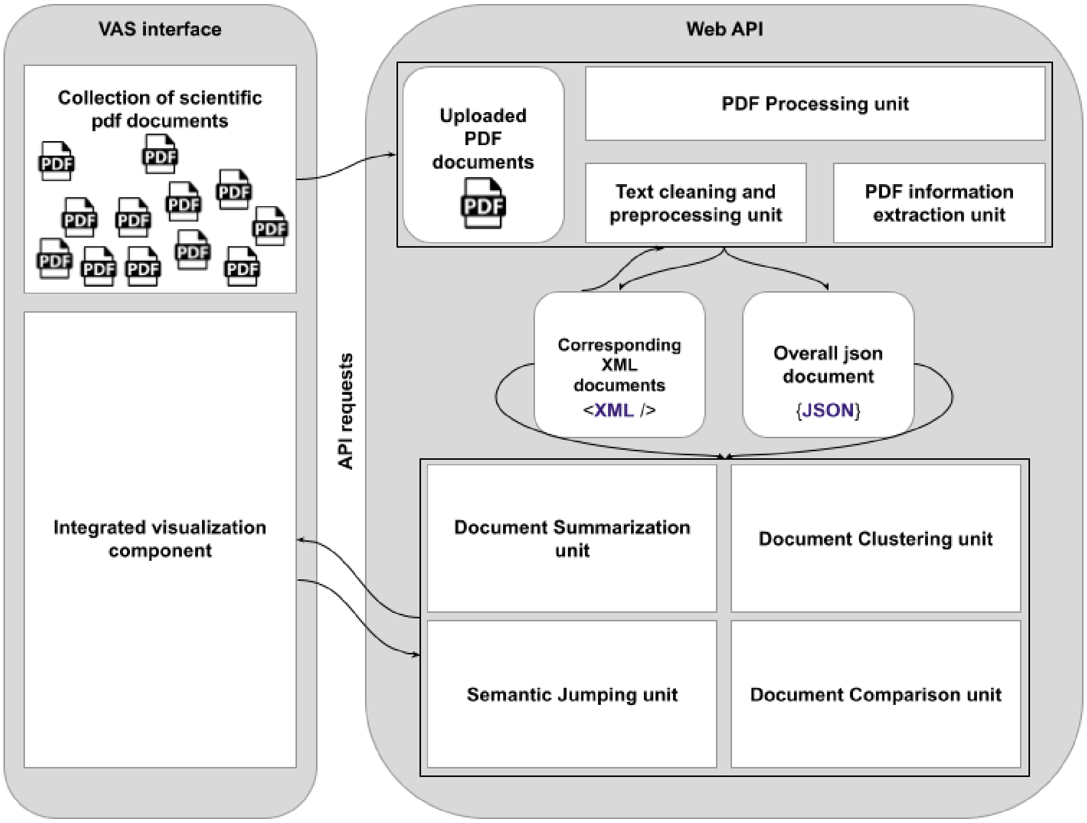

VAS interface
Collection of scientific pdf documents
PDF
Integrated visualization component
API requests
Web API
Uploaded PDF documents
PDF Processing unit
Text cleaning and preprocessing unit
PDF information extraction unit
Corresponding XML documents
<XML />
Overall json document
{JSON}
Document Summarization unit
Document Clustering unit
Semantic Jumping unit
Document Comparison unit

Figure 1 - overview of VAS structure, pipeline stages, and its sub-modules

3.1.1 PDF Processing unit

Through the form provided in the interface of our VAS, researchers can upload their desired collection of scientific PDF documents to examine further. Therefore, processing the content of the uploaded documents and extracting their contributing information is the first stage of our pipeline.

Due to the flexibility of PDF typesetting, scientific PDF documents come with numerous layouts in which the position of headings, footings, texts on side margins, tables, figures, and other components of a document are determined arbitrarily. Therefore, extracting the main textual content of a scientific document cleared from any noisy data (e.g., page headings or publisher's title) is a challenging task. Some studies have addressed this problem by adopting solutions such as rule-based methods, machine-learning models, or heuristics. To exemplify, pdftotext[3], PDFMiner[4], PDFBox[5], and xpdf[6] are some of the widely used tools to convert pdf documents into text or structured XML/HTML formats; however, they often fail to make a clear distinction between contributing and redundant texts in scientific PDF documents.

Among the existing tools, we decided to use Grobid[75], which uses machine learning approaches to extract, parse, and re-structure PDF documents (focusing on scientific PDF documents) into structured XML/TEI encoded documents. Grobid contains several models to analyze scientific documents' content, such as the "full-text" model that attempts to identify and structure appearing items (e.g., paragraphs, section titles, figures, tables, etc.) in the body text of a scientific document. To facilitate our web API with Grobid service, we used Grobid docker container[7] alongside its Python client library[8].

When researchers upload their desired collection of scientific PDF documents to the VAS, several API requests, as many as the number of uploaded documents, will be sent

[3] Available at: https://github.com/jalan/pdftotext

[4] Available at: https://github.com/euske/pdfminer

[5] Available at: https://pdfbox.apache.org/

[6] Available at: https://www.xpdfreader.com/

[7] Further information is available at: https://grobid.readthedocs.io/en/latest/Grobid-docker/

[8] Available at: https://github.com/kermitt2/grobid_client_python

to the web API containing each PDF document. Grobid service will generate an XML file containing the parsed and structured content of the document for each uploaded PDF file. These XML documents are used in other modules of the web API to process the textual content of the uploaded scientific documents.

3.1.2 Text Cleaning and Preprocessing unit

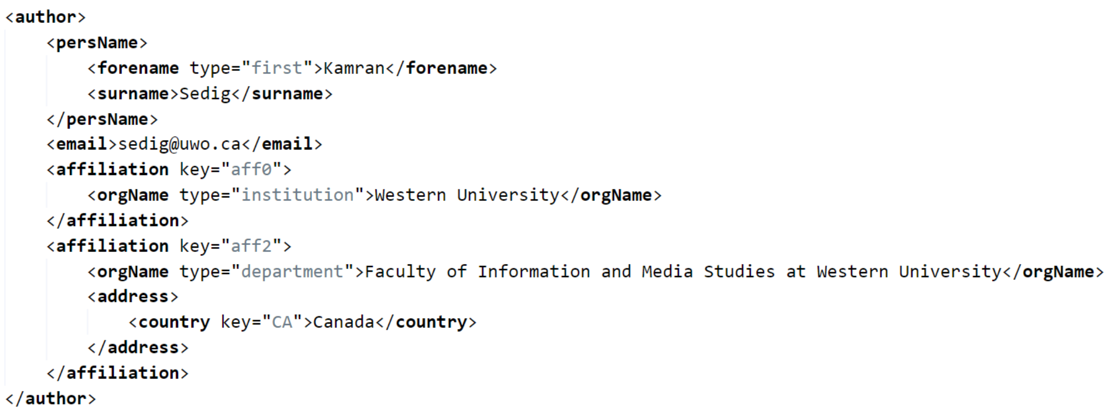
The preprocessing and cleaning of textual data of scientific documents is an essential step in our pipeline as it can guarantee the consistency and quality of analytics results. There is no specific endpoint for requesting text cleaning or preprocessing; instead, the module is used to prepare extracted textual data for other modules in the API, such as PDF information extraction or document summarization modules.

This module provides two major methods: 1) text cleaning and 2) word frequency normalization methods. The first method removes all the line breaks, special characters, single characters, reference indicators, and stop words[9] from the input text and returns it in lower case. It is important to mention that this method is only applied upon request and on specific sections of the textual content (e.g., paragraphs or titles) as formulas or some other sections of a scientific document may contain special/single characters that provide important information in the document. The second method tokenizes the input textual content into its constructing words, counts each word's appearance within the text, and normalizes the frequency of each word based on the frequency of the most frequent word afterward. This method is a statistical approach for extracting keywords of a scientific document and can be used in other modules such as document summarization.

[9] Stop words are a set of commonly used words in a language that carry out very little information and do not contribute to scientific documents semantically (e.g., “a”, “the”, “is”, “are”, etc.).

3.1.3 PDF Information Extraction unit

As mentioned in 3.1.1, the Grobid service generates an XML file[10] per each uploaded document containing structured extracted information of the original PDF file wrapped in specific XML tags. To cite an example, we can refer to Figure 2, which shows a segment of an XML file generated for the uploaded PDF version of [58] containing the extracted information of its first author structured using specific tags:



```
<author>
    <persName>
        <forename type="first">Kamran</forename>
        <surname>Sedig</surname>
    </persName>
    <email>sedig@uwo.ca</email>
    <affiliation key="aff0">
        <orgName type="institution">Western University</orgName>
    </affiliation>
    <affiliation key="aff2">
        <orgName type="department">Faculty of Information and Media Studies at Western University</orgName>
        <address>
            <country key="CA">Canada</country>
        </address>
    </affiliation>
</author>
```

Figure 2 - A segment of a sample XML file generated by Grobid

Provided the information in an original scientific PDF document, Grobid can recognize its title, authors, keywords, publishing data, publisher name, and other metadata. Therefore, our PDF information extraction module uses these generated XML files to provide requested information in JSON format and store it for further usage in other modules.

This module uses the Beautiful Soup Python library[11] for parsing XML files and pulling out desired information. It generates a JSON object containing the title, publishing date, authors, publisher name, keywords, abstract, and references of a scientific document

[10] Online demo is available at: https://cloud.science-miner.com/grobid/

[11] Further information and documentations are available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

alongside a unique ID based on the order of documents in the uploading process. It is important to mention that not all the methods in this module rely solely on the tags relevant to their functionality. As an example, the keywords extraction method not only extracts the keywords declared in the “keywords” tag in the XML file that is based on the keywords authors define at the beginning of their document, but also processes the body of the document and identifies the most frequent words and merges both sets to be more precise.

Other than the abovementioned information items, this module provides another method that processes the document's body in more depth. It recognizes the section titles, paragraphs, formulas, and lists in the body of a document. Furthermore, it parses each paragraph into its constructing sentences and calculates the score of each sentence based on the frequency of its words in the whole document. These sentence scores provide helpful information for document summarization or specific visual encodings that will be explained later.

As seen in Figure 1, there is an overall JSON document that holds the uploaded documents' general information. After parsing an XML file and extracting relevant information, this module adds the abovementioned generated JSON object to the array of existing objects in the overall document. In conclusion, our pipeline generates a JSON document file in response to the collection of uploaded scientific documents, which holds all their information except for the processed body of documents.

3.1.4 Document Clustering unit

Clustering is the process of separating a set of data objects into subsets (clusters) where data objects belonging to one cluster are similar to one another yet dissimilar to the object in other clusters[76]. Separating a collection of scientific documents into multiple clusters based on their textual content is an important step in discovering the relationships between different documents within a corpus. There are various clustering methods and algorithms such as hierarchal, partitioning relocation, density-based partitioning, grid-based, or constraint-based methods [77].

Experimenting with different clustering algorithms and reviewing the literature, we adopted the k-means algorithm for clustering the uploaded scientific documents. K-means[78] groups data objects into k clusters, relocates cluster centers, and re-assigns objects to clusters based on the minimum distance to cluster centroids iteratively. The document clustering method proposed in this module uses silhouette score alongside k-means algorithm provided by scikit-learn Python library[79] to find the optimal number of clusters and assign each scientific document to its relevant cluster.

Before applying the k-means algorithm on scientific documents, we concatenate each scientific document’s title and abstract together as these entities determine the semantic direction of documents with a high probability. In the next step, we apply a custom text cleaning method on the concatenated strings in which a custom set of scholarly stop words (e.g., “doi”, “et”, “al”, “figure”, “fig”, etc.) removal alongside word lemmatization are added to the basic text cleaning method to improve the clustering results. Vectorizing the preprocessed strings using sentence embeddings[80] and applying dimensionality reduction on the vectors are the following steps in this method.

The k-means algorithm is applied on the vectors using different values of k, and the corresponding silhouette scores are calculated to find the optimal number of clusters, and then documents are re-clustered using the optimal k. Afterwards, the most frequent unigrams, bigrams, and trigrams of the documents’ titles belonging to a cluster are calculated to find a label for the cluster. When clusters are determined and labeled, the cluster Id, label, and keywords are assigned to the JSON object of the scientific documents belonging to each cluster, and the overall JSON document gets updated in consequence.

3.1.5 Document summarization unit

By reviewing the existing similar tools in the literature, it can be concluded that document summarization is an essential component to support researchers in mitigating the issue of the high load of scientific documents and cover an adequate amount of information to remain up to date in their research domain. Automatic text summarization methods can be broadly classified into two major categories: 1) extractive and 2) abstractive text summarization. Extractive approaches crop out portions of a document and stitch them together to produce a condensed representation of the document[81]. Extractive summaries do not guarantee good narrative coherence; however, they can provide an approximate representation of the content of the text for relevancy judgement[82]. On the other hand, abstractive summarizations generate summaries from scratch and might rephrase or use words that were not in the original text[83]. Since abstractive summarizations require extensive natural language processing and heavy supervision and are limited to short documents[84], [85], we will focus on extractive methods to summarize the uploaded scientific documents in the pipeline.

Our document summarization unit adopts an extractive method that takes a single document and a size variable as inputs and generates the corresponding summary. This method parses the content of the input document into its constructing sentences and calculates the score of each sentence by the frequency of its constructing words in the whole document. According to the size variable, which is a fractional number ranging from 0 to 1 and determines the number of sentences of the output summary, the top-ranked sentences are selected and sorted based on their position in the original text (e.g., when the size variable is equal to 0.5, the method selects the top 50% of the sentences of the original text).

3.1.6 Semantic Jumping unit

As mentioned in the previous chapters, researchers do not read scientific documents in a linear fashion. They might focus on specific sections of a document more than others and jump to another section based on their information needs and research tasks; however, with the rapid growth of scientific information, it is challenging for researchers to discover their desired sections across one document or the whole corpus they are interacting with.

By computing the **semantic similarity** of different sections of scientific documents and based upon the idea of **hypertext links**, we can support researchers to carry out these activities automatically. Due to linguistic factors such as synonymy and polysomy, we cannot rely on the frequency or uniformity of words in different sections to discover their semantic closeness [86]. Therefore, we have adopted a modern method, Sentence-BERT (SBERT)[80], which is a modification of BERT[87] networks and manages to derive semantically meaningful sentence embeddings in a way where semantically similar sentences are close in the vector space[12]. SBERT can be used for semantic similarity search or clustering by performing a similarity measure like Manhattan / Euclidean distance on the derived sentence vectors.

Our semantic jump module includes two main methods that find semantically similar sentences in one document and semantically close documents, respectively. The first method takes one sentence, the document it belongs to, and a size variable. After vectorizing the input sentence and the parsed sentences of the input document, the cosine similarity of the pair of the input sentence and each sentence of the document is calculated to find the most similar sentences within the document. According to the input size variable, the top-ranked sentences and their position in the document will be returned as output. The second method functions in the same fashion; however, it vectorizes the abstracts of the documents stored in the overall JSON document to find similar documents to the concept carried by the input sentence.

3.1.7 Document Comparison unit

Discovering supporting evidence across scientific documents, tracing the results of similar experiments in a domain, and generally comparing scientific documents is an important activity during the sensemaking process for researchers. To support such activities, we have provided a module that can automate the comparison process to an extent. This module contains three major methods which take two documents as inputs and perform analytical processes on them.

The first method calculates the semantic similarity of two input documents by simply computing the cosine similarity of their vectorized abstracts of the documents and generating a fractional number between -1 to 1. On the other hand, the second method

[12] Further information and the corresponding Python framework is available at: https://www.sbert.net/

parses the body content of the input documents into their constructing sentences and ranks each sentence based on the appearance of a set of keywords which is a combination of the keywords of both documents. Therefore, similar sentences from the document will get higher scores than unique sentences. Finally, the third method provides semantic jumping functionality for both documents simultaneously. To explain further, it retrieves similar sentences alongside their position in each document to a concept given as the input; therefore, researchers can compare the documents' approach towards one concept at the same time.

In this chapter, we examined the Web API, implemented pipeline, and its sub-modules in depth. It is important to mention that various Machine Learning and NLP approaches other than the ones adopted in the abovementioned modules can be plugged into the modules without affecting the general structure of our API that shows the scalability of our VAS. In the next chapter, we will analyze the visualization component of the VAS in detail.

Chapter 4

4 Visualization and Interface Design

Based on the abstract design for data flow in the previous chapter, the interface of our VAS can access the processed information extracted from an uploaded collection of documents. Furthermore, the visualization component is reactive to the incoming data, updates dynamically, sends required requests to the Web API, and updates according to the received responses again in an interaction loop between representation and computing spaces. As discussed in 2.4, encoding large amounts of scientific information, using node-link diagrams as a common visualization technique, and numerous navigational interactions to drill into documents are some of the main limitations which prevent researchers from overcoming their cognitive challenges and carrying out the process of sensemaking properly. In this chapter, we will examine the integrated visualization component of our VAS that attempts to minimize the number of navigations and address the abovementioned issues to afford rapidity of the sensemaking process for researchers. In section 4.1, we will study the technologies used to implement the interface. In section 4.2, we will examine the visual approaches and the visual components of our VAS's interface. Finally, we will go through a usage and comparative scenario as a proof of concept in 4.3.

4.1 Integration of React JS and D3.JS

Data Driven Documents (D3)[88] is a JavaScript library that supports simple to very complex web-based interactive data visualizations. Unlike many other visualization libraries, D3 enables direct inspection and manipulation of *Document Object Model* (DOM[13]), allowing designers to map arbitrary data to DOM elements and design a wide variety of data-driven visualizations.

[13] DOM is the data representation of the objects which the structure and content of a document on the web is based upon, and enables programming languages to interact with a page to change its document

D3 also provides specific operators called “event handlers” that respond to user input, act on a selected set of elements, and potentially involve animated transitions to enable interactions in a visualization. In order to update visualizations according to data changes and keep them dynamic, basic visualizations use JavaScript *setInterval*[14] method to run their code in short time intervals. However, this method of handling data changes can be computationally intensive when the number of data and visual variables increases significantly. To address this problem, we have integrated D3.JS with React JS[15] library and one of its features, React Hooks.

React JS is a component-based JavaScript library created for building web-based user interfaces. In general, it uses Virtual DOM, which compares components’ previous states and updates only the changed items in the real DOM instead of updating all the components; therefore, it can significantly improve our VAS’s performance. Furthermore, it offers various extensions to support complete application architecture, such as Redux[16], a state container library that supports the consistency of stored data in the application. React Hooks are a set of functions that benefits from React state and life cycle features. One of these functions, *useEffect*[17], enables the application to run a specific body of code when a specific set of data variables changes. Thus, using React Hooks, we can develop dynamic complex visualization without running a large body of code in short time spans iteratively.

structure, style, and content (Further information is available at: https://developer.mozilla.org/en-US/docs/Web/API/Document_Object_Model/Introduction)

[14] Further information is available at: https://developer.mozilla.org/docs/Web/API/setInterval

[15] Further information is available at: https://reactjs.org/

[16] Further information is available at: https://redux.js.org/

[17] Further information is available at: https://reactjs.org/docs/hooks-effect.html

4.2 Integrated Visualization Component

Although visualization of scientific information mainly faces the challenge of a large number of scientific documents and information overload[28], to afford rapidity of information exploration and sensemaking process in general, we have decided to visualize all the processed scientific information in one visualization canvas. Performing numerous sub-tasks in one scrolling window to make sense of a collection of scientific documents might also be challenging and lead to potential loss of context[59]. Therefore, we adopted the fisheye view[89] visualization technique and interactive lenses to design a non-scrollable visualization canvas, enlarge specific regions of interest, suppress other regions, and maintain the global structure of the visual marks to avoid losing context.

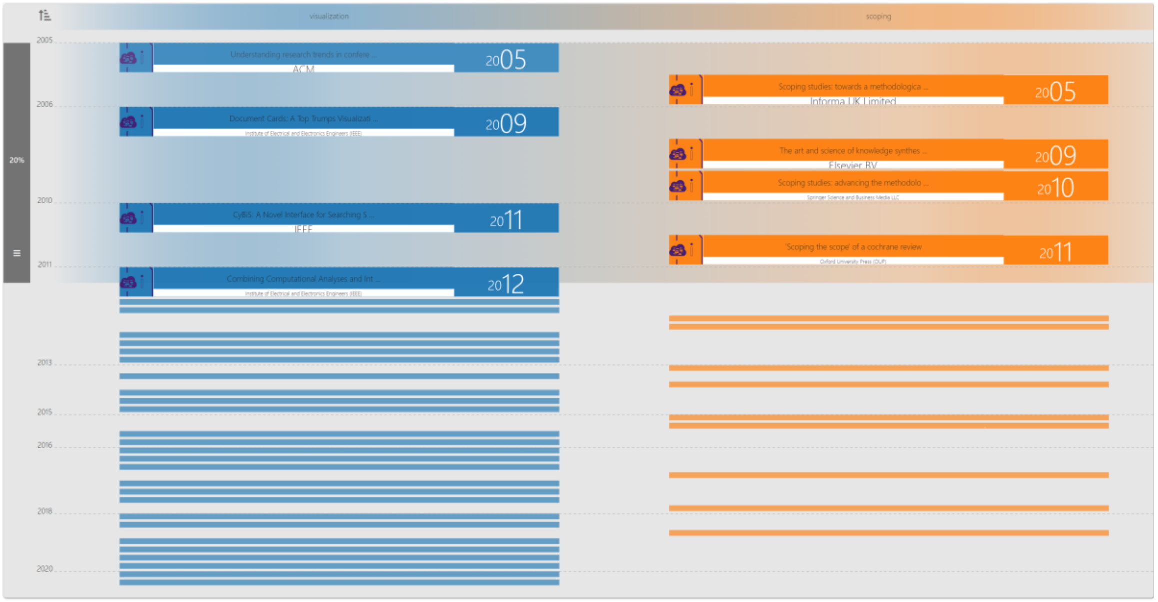

visualization
scoping
2005
2006
2010
2011
2013
2015
2016
2018
2020
20%
Understanding research trends in confere ...
ACM
2005
Document Cards: A Top Trumps Visualizati ...
Institute of Electrical and Electronics Engineers (IEEE)
2009
CyBiS: A Novel Interface for Searching S ...
IEEE
2011
Combining Computational Analyses and Int ...
Institute of Electrical and Electronics Engineers (IEEE)
2012
Scoping studies: towards a methodologica ...
Informa UK Limited
2005
The art and science of knowledge synthes ...
Elsevier BV
2009
Scoping studies: advancing the methodolo ...
Springer Science and Business Media LLC
2010
'Scoping the scope' of a cochrane review
Oxford University Press (OUP)
2011

Figure 3 - An overview of the visualization canvas

As represented in Figure 3, the uploaded documents are encoded to rectangular colored elements and positioned in different columns. It is studied that basic shapes can be perceived pre-attentively and should be used to convey the most important information [90]; therefore, we used rectangular glyphs (hereinafter referred to as “document rectangles”) to represent the position of each uploaded scientific document in the arranged canvas.

Furthermore, we have implemented a sliding lens that provides the fisheye view in our visualization. This lens uses the following formulas to dilate specific documents in the canvas:

$$\#\ of\ Enlarged\ Documents = \left(\frac{Sliding\ Window\ Height}{Canvas\ Height}\right)^2 \times \#\ of\ Documents$$

$$Document\ Rectangle\ Height = \frac{(Canvas\ Height - (\#\ of\ Documents \times margins^{18}))}{(2\ \times \#\ of\ Documents - \#\ of\ Enlarged\ Documents)}$$

$$Enlarged\ Document\ Rectangle\ Height = \left(\frac{Canvas\ Height}{Sliding\ Window\ Height}\right)^2 \times Document\ Rectangles\ Height$$

Researchers can use the controller bar on the left side of the sliding lens to drag it over all the canvas in vertical directions and change its height to cover more or fewer document rectangles in the lens. An indicator on the controller bar also displays the proportion of documents covered by the sliding lens.

Various arrangements and orderings of encoded information items can affect cognitive activities in different ways[91], and adjusting the information items in representation space can directly affect the ordering of information items in mental space[92]. Thus, we have devised our visualization with different document arrangements (e.g., based on publication dates, alphabetical order of document titles, or the number of out-link references mentioned in documents) that researchers can choose from. The canvas also contains several horizontal dashed axes as reference information to map document rectangles against their arrangement criteria values. However, as the height and position of document rectangles depend on the sliding lens, the positions of these axes float accordingly.

[18] Margins refers to the white space between two consecutive documents in one column

It is important to mention that document rectangles are positioned in different columns representing clusters of documents and using distinct colors to help researchers identify each cluster (see Figure 4). Studies show that the accuracy of perception different graphical attributes varies significantly[93], and some attributes are superior to the others in this sense; thus, we decided to use colors for encoding different clusters as they have proven to have serious effects on our perception and judgment of visual entities. The *cluster bar* at the top of the visualization canvas displays each cluster's label and enables researchers to toggle to the word clouds view by clicking on it (see Figure 5). Each word cloud can support researchers to get familiar with the context and the domain of the included documents in a rapid fashion.

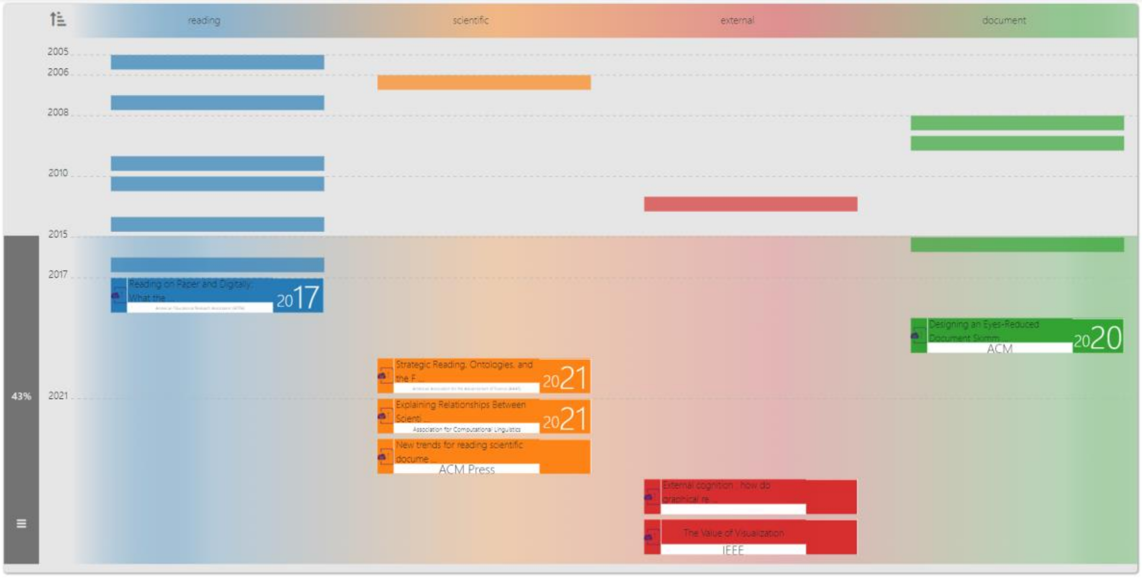

reading
scientific
external
document
2005
2006
2008
2010
2015
2017
2021
43%
Reading on Paper and Digitally: What the ...
2017
Designing an Eyes-Reduced Document Skimm...
ACM
2020
Strategic Reading, Ontologies, and the F...
2021
Explaining Relationships Between Scienti...
Association for Computational Linguistics
2021
New trends for reading scientific docume...
ACM Press
External cognition: how do graphical re...
The Value of Visualization
IEEE

Figure 4 - Encoding clustered documents with distinct colors and horizontal position

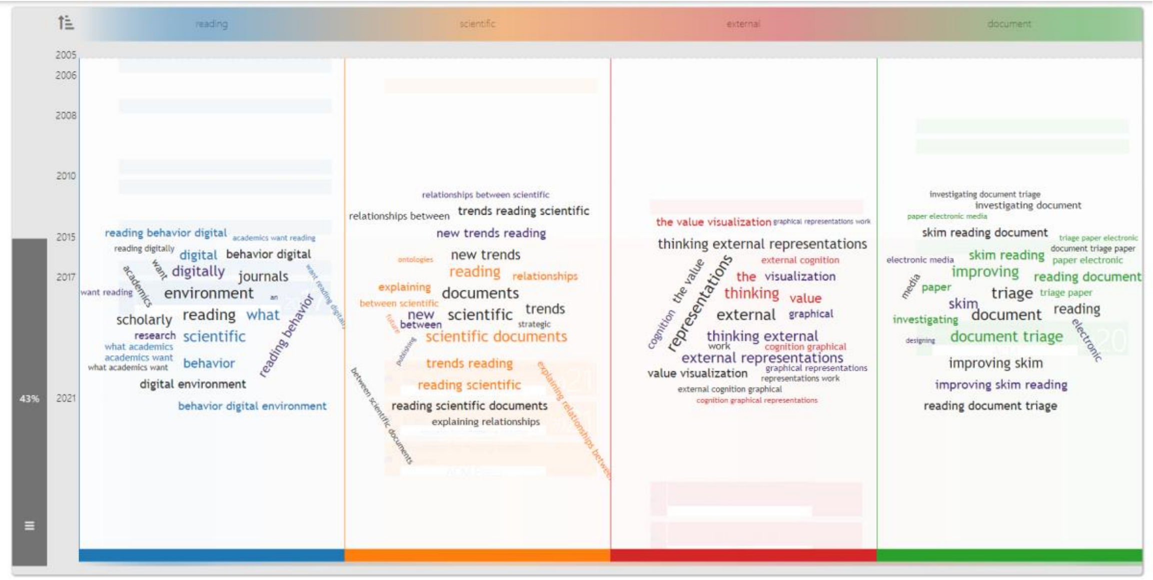


Figure 5 – Word cloud view of the document clusters

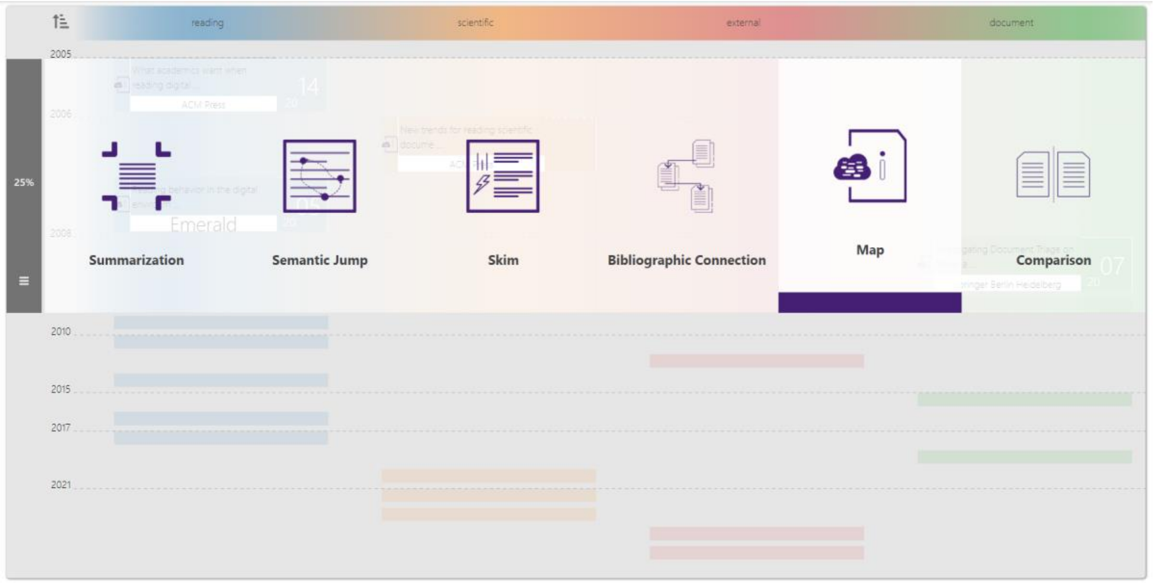


Figure 6 - List of proposed lenses

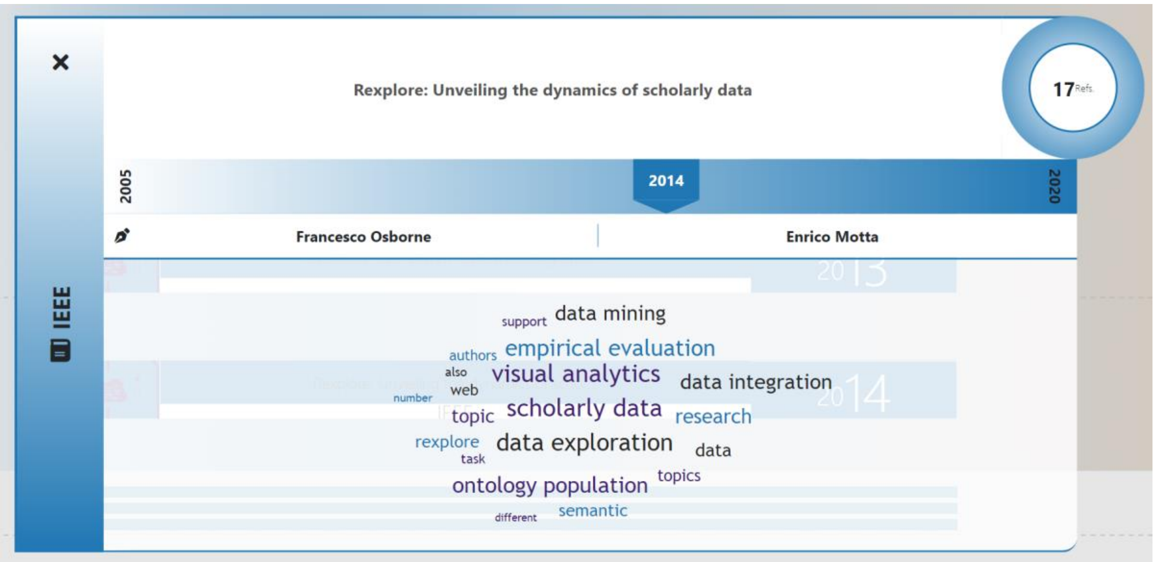
Due to the complexity and inner layers of scientific information, it is challenging to design visualizations in such a way as to afford perception by the human mind, and the effectiveness of such visualizations is a function of both data type and visualization goal[94]. Since our goal is to encode a large subset of scientific information items extracted from scientific documents and support researchers in making sense of the encoded data rapidly, we must mitigate researchers' perceptual and cognitive overload when interacting with the VAS. To address this problem, we have adopted an approach by which researchers can adjust the level of the interiority of the visualization and investigate latent information when required. Thanks to this feature, researchers can perceive the macrostructure of the encoded scientific documents, inquire about it, and drill into the latent information to answer the inquiries[95].

We used magic lenses to implement the abovementioned functionality. Magic lenses can be considered as visual filters that alter the presentation of visual elements to "reveal hidden information, enhance data of interest, or suppress distracting information"[96]. As mentioned before, interactive lenses can be considered as functions with two major stages: 1) the Selection stage, which captures what is to be affected by a lens, and 2) the Join stage, which joins the results obtained from the lens with the base visualization.

Although the sliding lens can act as a visual filter and reveal some inner information items of the focused scientific documents, the area of each document rectangle cannot afford enough space to encode required information items. Therefore, aside from selecting the sliding lens to enlarge desired documents, researchers can select a single document within the sliding lens, expand the lens, and drill into its resulting information. In conclusion, concerning the single document selection, it can be said that the visualization shows the lens selection and the generated lens results separately. Concerning the geometric properties of expanded lenses, we designed rectangular lenses compatible with the aspect ratio of the base visualization canvas and positioned them on top of the selected document rectangle (see Figure 7).

In this way, we managed to reduce the density and complexity of our visualization, allow researchers to focus on specific documents, and access the macrostructure of the corpus simultaneously. As represented in Figure 6, we have provided six lenses to cover different information items of scientific documents and properly support the rapid sensemaking process of researchers. In the rest of this section, we will analyze each of these lenses thoroughly.

4.2.1 Map Lens



Rexplore: Unveiling the dynamics of scholarly data

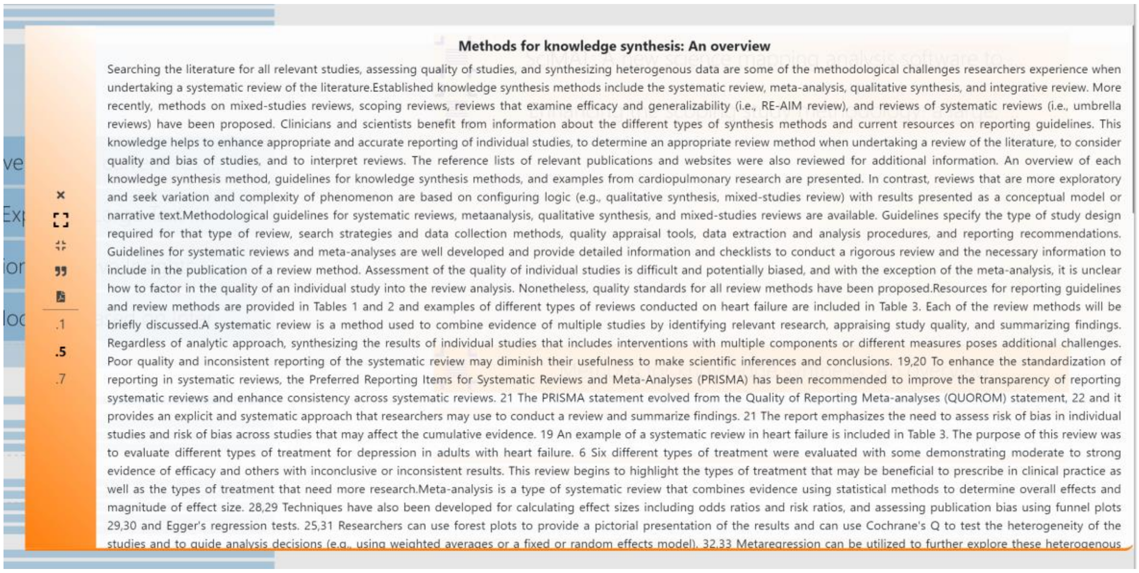
Francesco Osborne | Enrico Motta

Figure 7 - A sample expanded map lens snapshot

Map lens is the default lens of our VAS that attempts to provide an overview of the metadata about scientific documents. When the map lens is selected, the enlarged document rectangles covered by the sliding lens will display their corresponding scientific document's title, publication year, and publisher (see Figures 3 and 4). Researchers can click on their desired document rectangle to expand the map lens (see Figure 7) and access further information about the document. In addition to the title, publisher, and publication year, this lens provides the list of authors, keywords, and the number of out-link references mentioned in the document. To represent the publication year, we used a bar that encodes the range of publication years of the uploaded documents in the corpus and contains a base shape positioned relatively and displaying the publication year of the selected document. The same technique is used for representing the number of out link references where two centric circles are used to provide the relative perception of the number of references used in the selected document compared to the highest number of references mentioned in a document within the corpus.

4.2.2 Summarization Lens

The summarization lens aims to provide a condensed representation of a scientific document to support researchers in comprehending the content of that document rapidly. As mentioned in 3.1.5, a summarization module is provided in the Web API of our VAS that can be accessed by a specific endpoint. Therefore, when researchers select summarization lens from the lens menu (see Figure 6) and click on an enlarged document rectangle, the corresponding summarization lens expands and sends a request to the API alongside the selected document ID and the size of the summary (0.5 in the default state) and displays the returned response as the documents' summary.



Methods for knowledge synthesis: An overview

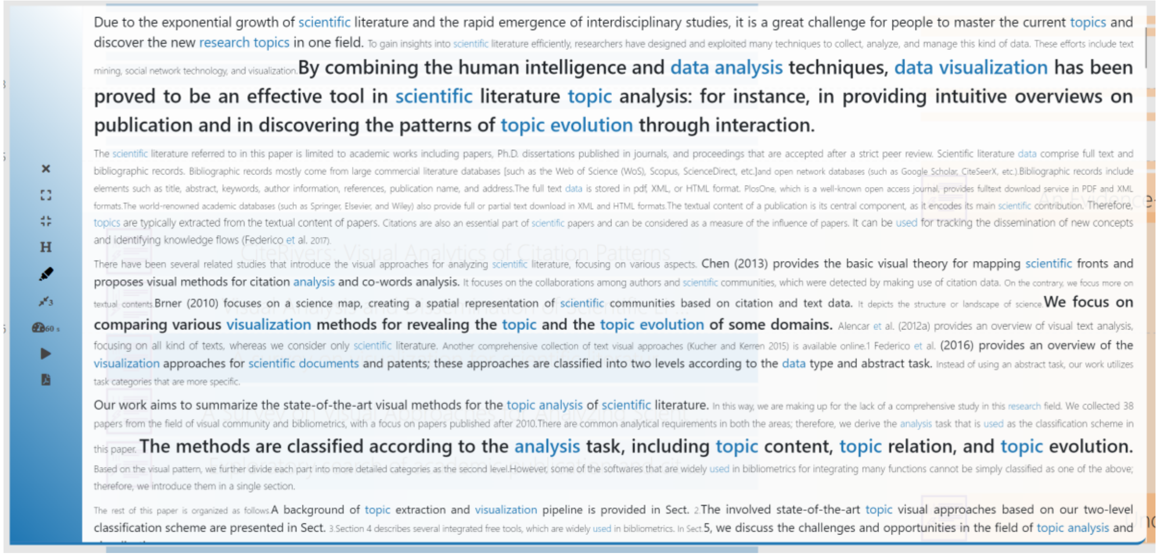
Searching the literature for all relevant studies, assessing quality of studies, and synthesizing heterogenous data are some of the methodological challenges researchers experience when undertaking a systematic review of the literature.Established knowledge synthesis methods include the systematic review, meta-analysis, qualitative synthesis, and integrative review. More recently, methods on mixed-studies reviews, scoping reviews, reviews that examine efficacy and generalizability (i.e., RE-AIM review), and reviews of systematic reviews (i.e., umbrella reviews) have been proposed. Clinicians and scientists benefit from information about the different types of synthesis methods and current resources on reporting guidelines. This knowledge helps to enhance appropriate and accurate reporting of individual studies, to determine an appropriate review method when undertaking a review of the literature, to consider quality and bias of studies, and to interpret reviews. The reference lists of relevant publications and websites were also reviewed for additional information. An overview of each knowledge synthesis method, guidelines for knowledge synthesis methods, and examples from cardiopulmonary research are presented. In contrast, reviews that are more exploratory and seek variation and complexity of phenomenon are based on configuring logic (e.g., qualitative synthesis, mixed-studies review) with results presented as a conceptual model or narrative text.Methodological guidelines for systematic reviews, metaanalysis, qualitative synthesis, and mixed-studies reviews are available. Guidelines specify the type of study design required for that type of review, search strategies and data collection methods, quality appraisal tools, data extraction and analysis procedures, and reporting recommendations. Guidelines for systematic reviews and meta-analyses are well developed and provide detailed information and checklists to conduct a rigorous review and the necessary information to include in the publication of a review method. Assessment of the quality of individual studies is difficult and potentially biased, and with the exception of the meta-analysis, it is unclear how to factor in the quality of an individual study into the review analysis. Nonetheless, quality standards for all review methods have been proposed.Resources for reporting guidelines and review methods are provided in Tables 1 and 2 and examples of different types of reviews conducted on heart failure are included in Table 3. Each of the review methods will be briefly discussed.A systematic review is a method used to combine evidence of multiple studies by identifying relevant research, appraising study quality, and summarizing findings. Regardless of analytic approach, synthesizing the results of individual studies that includes interventions with multiple components or different measures poses additional challenges. Poor quality and inconsistent reporting of the systematic review may diminish their usefulness to make scientific inferences and conclusions. 19,20 To enhance the standardization of reporting in systematic reviews, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) has been recommended to improve the transparency of reporting systematic reviews and enhance consistency across systematic reviews. 21 The PRISMA statement evolved from the Quality of Reporting Meta-analyses (QUOROM) statement, 22 and it provides an explicit and systematic approach that researchers may use to conduct a review and summarize findings. 21 The report emphasizes the need to assess risk of bias in individual studies and risk of bias across studies that may affect the cumulative evidence. 19 An example of a systematic review in heart failure is included in Table 3. The purpose of this review was to evaluate different types of treatment for depression in adults with heart failure. 6 Six different types of treatment were evaluated with some demonstrating moderate to strong evidence of efficacy and others with inconclusive or inconsistent results. This review begins to highlight the types of treatment that may be beneficial to prescribe in clinical practice as well as the types of treatment that need more research.Meta-analysis is a type of systematic review that combines evidence using statistical methods to determine overall effects and magnitude of effect size. 28,29 Techniques have also been developed for calculating effect sizes including odds ratios and risk ratios, and assessing publication bias using funnel plots 29,30 and Egger's regression tests. 25,31 Researchers can use forest plots to provide a pictorial presentation of the results and can use Cochrane's Q to test the heterogeneity of the studies and to guide analysis decisions (e.g., using weighted averages or a fixed or random effects model). 32,33 Metaregression can be utilized to further explore these heterogenous

Figure 8 - A sample expanded summarization lens snapshot

As represented in Figure 8, the summarization lens provides some options on the left side to enable researchers to change its representation and extract further information. It would afford the original abstract and the PDF view of the selected document if specific figures, tables, or other information items were required to enhance the rapid comprehension process. Furthermore, it affords different summarization size values (e.g., a summarization size of 0.1 extracts the most important 10% of the sentences of the selected document) so that researchers can balance the reading time and the depth of details for each document.

4.2.3 Skim Lens

Since skimming is one of the common activities that researchers carry out during the document triage and sensemaking process, we decided to provide a lens that supports rapid document scanning and skimming by integrating text analysis and visualization approaches. As mentioned in 3.1.3, the PDF information extraction module includes a specific function that parses a given document's content into its constructing sentences and assigns an importance score to each sentence based on the frequency of the words that appeared in it. Therefore, when researchers select the skim lens from the lens menu and click on an enlarged document rectangle, the corresponding skim lens expands and sends a request to the API alongside the selected document ID to receive its textual content (e.g., headings, paragraphs, lists, and formulas).



Due to the exponential growth of scientific literature and the rapid emergence of interdisciplinary studies, it is a great challenge for people to master the current topics and discover the new research topics in one field. To gain insights into scientific literature efficiently, researchers have designed and exploited many techniques to collect, analyze, and manage this kind of data. These efforts include text mining, social network technology, and visualization. By combining the human intelligence and data analysis techniques, data visualization has been proved to be an effective tool in scientific literature topic analysis: for instance, in providing intuitive overviews on publication and in discovering the patterns of topic evolution through interaction.

The scientific literature referred to in this paper is limited to academic works including papers, Ph.D. dissertations published in journals, and proceedings that are accepted after a strict peer review. Scientific literature data comprise full text and bibliographic records. Bibliographic records mostly come from large commercial literature databases [such as the Web of Science (WoS), Scopus, ScienceDirect, etc.]and open network databases (such as Google Scholar, CiteSeerX, etc.).Bibliographic records include elements such as title, abstract, keywords, author information, references, publication name, and address.The full text data is stored in pdf, XML, or HTML format. PlosOne, which is a well-known open access journal, provides fulltext download service in PDF and XML formats.The world-renowned academic databases (such as Springer, Elsevier, and Wiley) also provide full or partial text download in XML and HTML formats.The textual content of a publication is its central component, as it encodes its main scientific contribution. Therefore, topics are typically extracted from the textual content of papers. Citations are also an essential part of scientific papers and can be considered as a measure of the influence of papers. It can be used for tracking the dissemination of new concepts and identifying knowledge flows (Federico et al. 2017).

There have been several related studies that introduce the visual approaches for analyzing scientific literature, focusing on various aspects. Chen (2013) provides the basic visual theory for mapping scientific fronts and proposes visual methods for citation analysis and co-words analysis. It focuses on the collaborations among authors and scientific communities, which were detected by making use of citation data. On the contrary, we focus more on textual contents.Brner (2010) focuses on a science map, creating a spatial representation of scientific communities based on citation and text data. It depicts the structure or landscape of science.We focus on comparing various visualization methods for revealing the topic and the topic evolution of some domains. Alencar et al. (2012a) provides an overview of visual text analysis, focusing on all kind of texts, whereas we consider only scientific literature. Another comprehensive collection of text visual approaches (Kucher and Kerren 2015) is available online.1 Federico et al. (2016) provides an overview of the visualization approaches for scientific documents and patents; these approaches are classified into two levels according to the data type and abstract task. Instead of using an abstract task, our work utilizes task categories that are more specific.

Our work aims to summarize the state-of-the-art visual methods for the topic analysis of scientific literature. In this way, we are making up for the lack of a comprehensive study in this research field. We collected 38 papers from the field of visual community and bibliometrics, with a focus on papers published after 2010.There are common analytical requirements in both the areas; therefore, we derive the analysis task that is used as the classification scheme in this paper. The methods are classified according to the analysis task, including topic content, topic relation, and topic evolution. Based on the visual pattern, we further divide each part into more detailed categories as the second level.However, some of the softwares that are widely used in bibliometrics for integrating many functions cannot be simply classified as one of the above; therefore, we introduce them in a single section.

The rest of this paper is organized as follows.A background of topic extraction and visualization pipeline is provided in Sect. 2The involved state-of-the-art topic visual approaches based on our two-level classification scheme are presented in Sect. 3Section 4 describes several integrated free tools, which are widely used in bibliometrics. In Sect.5, we discuss the challenges and opportunities in the field of topic analysis and

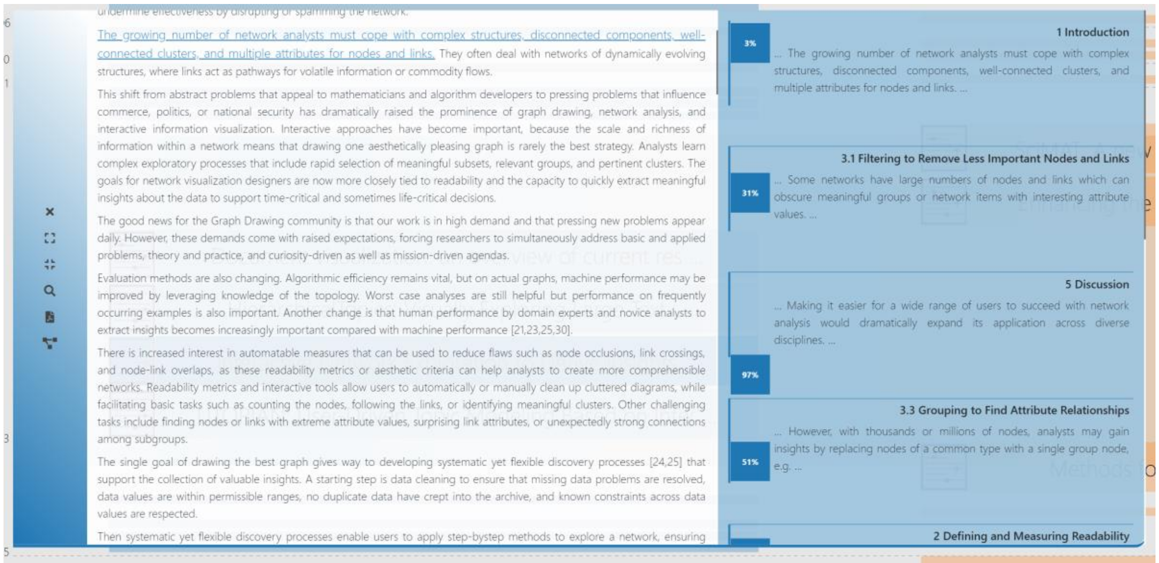
Figure 9 - A sample expanded skim lens snapshot

In order to enhance researchers' experience during the skimming process, we used pre-attentive visual attributes (e.g., size, hue, speed, direction) as they are difficult to ignore and unaffected by the high load of information[97], [98]. Therefore, we encoded each sentence with a specific font size according to the score calculated in the API, as size is considered highly pre-attentive and can be perceived rapidly[98]. We have considered a variable called *Compression Ratio*, which determines the font size ratio of the most important sentence to the least important one (equal to 2 by default). However, as size variations might interrupt readability for long passages of scientific documents[99], we afforded the compression ratio option (see Figure 9), enabling researchers to change the *Compression Ratio* value to 1.5, 2, or 3 relevant to their reading needs.

Furthermore, we encoded the keywords of the selected document in a different color (same as the color of the cluster that the selected document belongs to) to help researchers scan important sentences quicker (this functionality can be toggled in the options menu). In addition to the PDF view that covers any other information item required by a researcher, this lens affords auto skimming functionality where researchers can set the duration of the skimming process (e.g., 60, 120, 180 seconds) and let the content of the lens scroll automatically and in a top to bottom direction accordingly.

4.2.4 Semantic Jump Lens

This lens supports researchers in reading scientific documents in a non-linear fashion and rapidly accessing the sections that carry their information needs. As mentioned in 3.1.6, the semantic jump module provides semantically close sentences or documents given an input sentence or a search term and the selected document ID. Therefore, when researchers select the semantic jump lens from the lens menu and expand it, the textual content of the selected document will be displayed in the left pane of the lens (see Figure 10), and researchers can select a sentence to find its semantically similar sentences across the document. When the API discovers similar sentences and sends them back to the interface, they will be displayed on the right pane of the lens, and researchers can click on them to navigate to the corresponding section. Alongside each similar sentence, the section header and the position of that sentence within the selected document are displayed to help researchers choose which section they want to jump to.



undermine effectiveness by disrupting or spamming the network.

The growing number of network analysts must cope with complex structures, disconnected components, well-connected clusters, and multiple attributes for nodes and links. They often deal with networks of dynamically evolving structures, where links act as pathways for volatile information or commodity flows.

This shift from abstract problems that appeal to mathematicians and algorithm developers to pressing problems that influence commerce, politics, or national security has dramatically raised the prominence of graph drawing, network analysis, and interactive information visualization. Interactive approaches have become important, because the scale and richness of information within a network means that drawing one aesthetically pleasing graph is rarely the best strategy. Analysts learn complex exploratory processes that include rapid selection of meaningful subsets, relevant groups, and pertinent clusters. The goals for network visualization designers are now more closely tied to readability and the capacity to quickly extract meaningful insights about the data to support time-critical and sometimes life-critical decisions.

The good news for the Graph Drawing community is that our work is in high demand and that pressing new problems appear daily. However, these demands come with raised expectations, forcing researchers to simultaneously address basic and applied problems, theory and practice, and curiosity-driven as well as mission-driven agendas.

Evaluation methods are also changing. Algorithmic efficiency remains vital, but on actual graphs, machine performance may be improved by leveraging knowledge of the topology. Worst case analyses are still helpful but performance on frequently occurring examples is also important. Another change is that human performance by domain experts and novice analysts to extract insights becomes increasingly important compared with machine performance [21,23,25,30].

There is increased interest in automatable measures that can be used to reduce flaws such as node occlusions, link crossings, and node-link overlaps, as these readability metrics or aesthetic criteria can help analysts to create more comprehensible networks. Readability metrics and interactive tools allow users to automatically or manually clean up cluttered diagrams, while facilitating basic tasks such as counting the nodes, following the links, or identifying meaningful clusters. Other challenging tasks include finding nodes or links with extreme attribute values, surprising link attributes, or unexpectedly strong connections among subgroups.

The single goal of drawing the best graph gives way to developing systematic yet flexible discovery processes [24,25] that support the collection of valuable insights. A starting step is data cleaning to ensure that missing data problems are resolved, data values are within permissible ranges, no duplicate data have crept into the archive, and known constraints across data values are respected.

Then systematic yet flexible discovery processes enable users to apply step-bystep methods to explore a network, ensuring

3%

1 Introduction

... The growing number of network analysts must cope with complex structures, disconnected components, well-connected clusters, and multiple attributes for nodes and links. ...

31%

3.1 Filtering to Remove Less Important Nodes and Links

... Some networks have large numbers of nodes and links which can obscure meaningful groups or network items with interesting attribute values. ...

97%

5 Discussion

... Making it easier for a wide range of users to succeed with network analysis would dramatically expand its application across diverse disciplines. ...

51%

3.3 Grouping to Find Attribute Relationships

... However, with thousands or millions of nodes, analysts may gain insights by replacing nodes of a common type with a single group node, e.g. ...

2 Defining and Measuring Readability

Figure 10 - A sample semantic jump lens snapshot

The other options provided in this lens enable researchers to search for a specific concept instead of selecting an existing sentence in the document to retrieve semantically close sentences or document to the search term, toggle to the PDF view, and toggle to inter-document jump mode where similar other documents are retrieved instead of similar sentences within the document. It is also important to mention that the first items, both in similar sentences and documents list, allow researchers to navigate back to the original sentence or the document they were reading in the first place.

4.2.5 Bibliographic Connection Lens

Citation networks and bibliographic information of scientific documents have always been an important data source for researchers to follow research trends, discover similar works in one domain, and examine the relationships among a collection of scientific documents. Although our VAS mainly focuses on supporting researchers to make sense of the document's textual content rapidly, we designed a specific lens that uses a graph-based visualization technique to encode bibliographic connections among the uploaded documents.

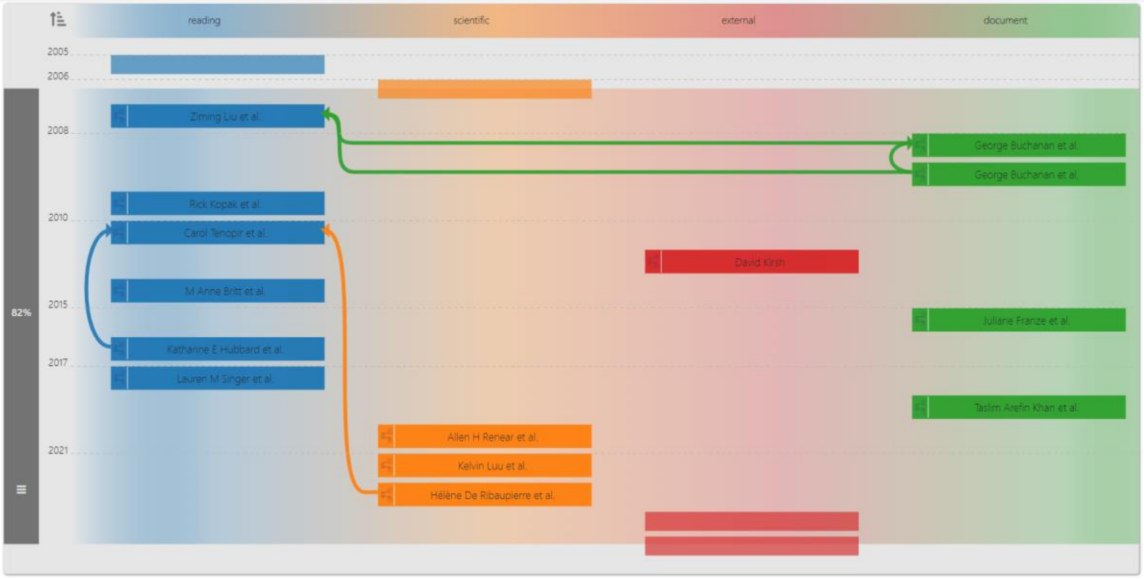


Figure 11 - A sample bibliographic connection lens snapshot

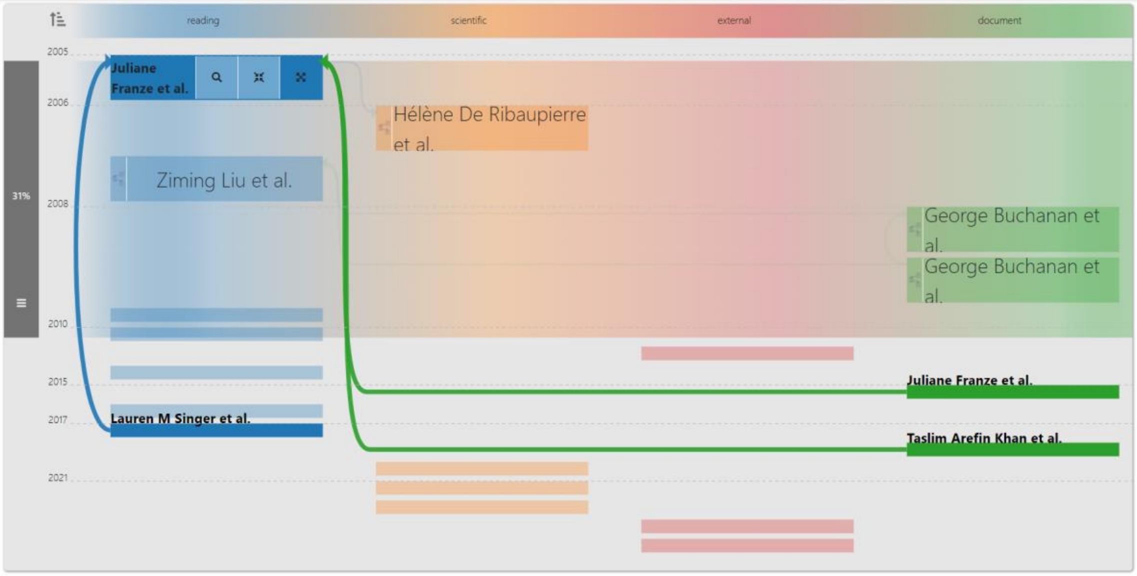


Figure 12 - Inspecting in-links and out-links of one document by bibliographic connection lens

By using the extracted titles and bibliographic references of the uploaded scientific documents, our bibliographic connection lens computes the in-links and out-links of each document and encodes the connections by using curved directional lines between related documents. Thanks to the unique vertical position of each document rectangle and specified margins between document cluster columns, no document rectangle could block a connection between two other documents in the canvas and interrupt researchers’ visual tasks to discover the connected documents (see Figure 11). However, the high number of connections can lead to the occlusion of curved connector lines and only allow researchers to locate the most and the least references documents. Therefore, as represented in Figure 12, we used the opacity visual attribute to fade extra connector lines when researchers click on one specific enlarged document rectangle to investigate its in-links or out-links and locate the connected documents.

4.2.6 Comparison Lens

As mentioned before, comparing scientific documents is a common activity among researchers to discover desired information items across similar studies during the sensemaking process. Therefore, we designed a comparison lens in our VAS, representing a similarity score, similar sentences, and similar concepts of the selected documents to support researchers in rapidly comparing their desired documents.

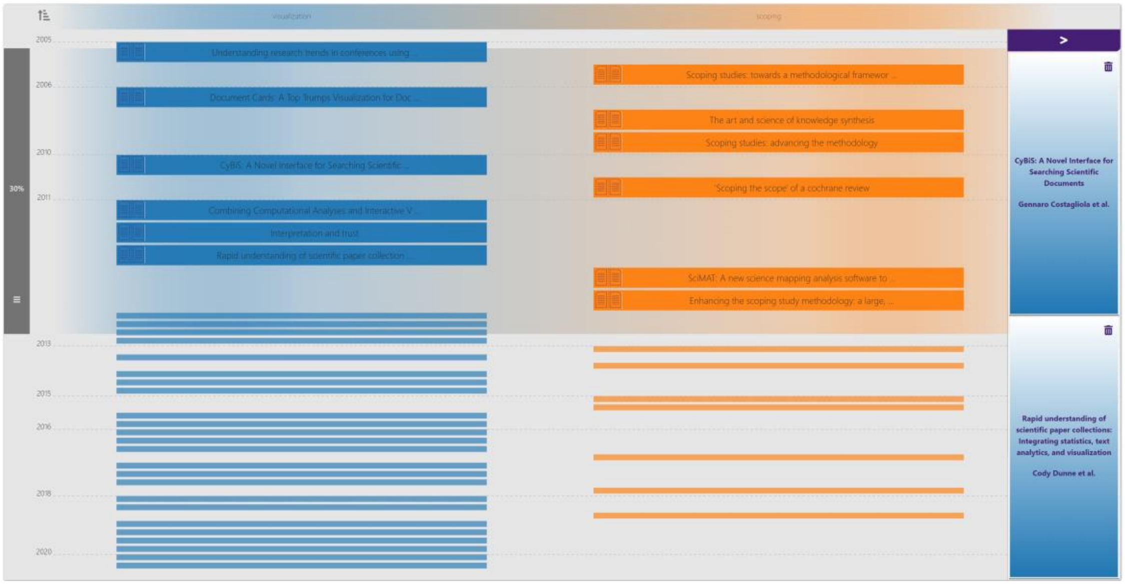
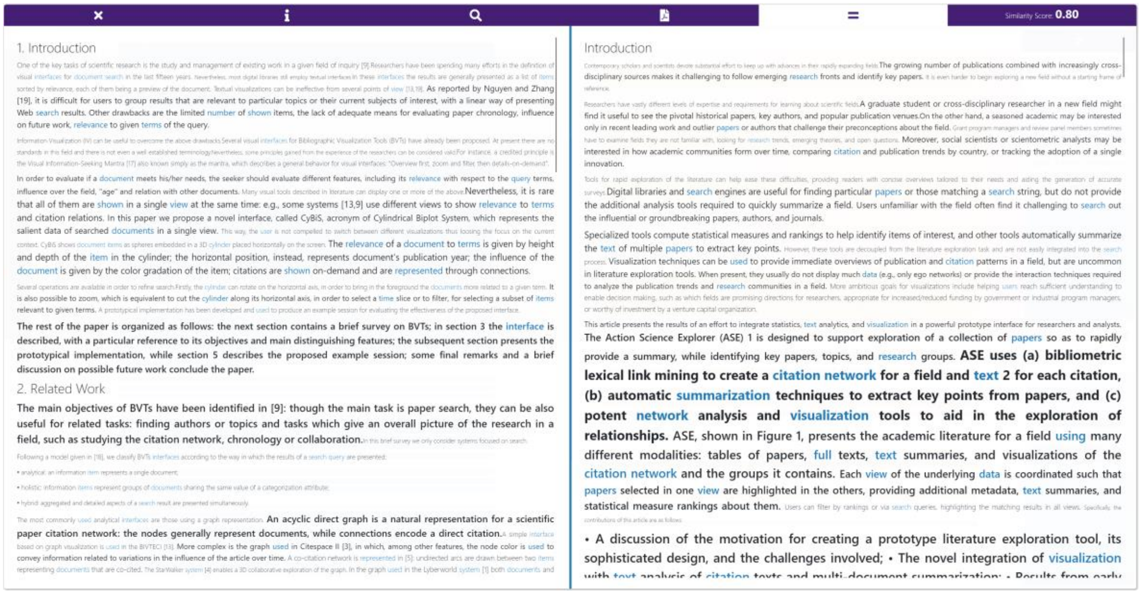

visualization
scoping
2005
2006
2010
2011
2013
2015
2016
2018
2020
30%
Understanding research trends in conferences using ...
Document Cards: A Top Trumps Visualization for Doc ...
CyBIS: A Novel Interface for Searching Scientific ...
Combining Computational Analyses and Interactive V ...
Interpretation and trust
Rapid understanding of scientific paper collection ...
Scoping studies: towards a methodological framewor ...
The art and science of knowledge synthesis
Scoping studies: advancing the methodology
'Scoping the scope' of a cochrane review
SciMAT: A new science mapping analysis software to ...
Enhancing the scoping study methodology: a large, ...
>
CyBiS: A Novel Interface for Searching Scientific Documents
Gennaro Costagliola et al.
Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization
Cody Dunne et al.

Figure 13 - selecting two documents to start comparison



1. Introduction

One of the key tasks of scientific research is the study and management of existing work in a given field of inquiry [9].Researchers have been spending many efforts in the definition of visual interfaces for document search in the last fifteen years. Nevertheless, most digital libraries still employ textual interfaces.In these interfaces the results are generally presented as a list of items sorted by relevance, each of them being a preview of the document. Textual visualizations can be ineffective from several points of view [13,19]. As reported by Nguyen and Zhang [19], it is difficult for users to group results that are relevant to particular topics or their current subjects of interest, with a linear way of presenting Web search results. Other drawbacks are the limited number of shown items, the lack of adequate means for evaluating paper chronology, influence on future work, relevance to given terms of the query.

Information Visualization (IV) can be useful to overcome the above drawbacks.Several visual interfaces for Bibliographic Visualization Tools (BVTs) have already been proposed. At present there are no standards in this field and there is not even a well established terminology.Nevertheless, some principles gained from the experience of the researchers can be considered valid.For instance, a credited principle is the Visual Information-Seeking Mantra [17] also known simply as the mantra, which describes a general behavior for visual interfaces: "Overview first, zoom and filter, then details-on-demand".

In order to evaluate if a document meets his/her needs, the seeker should evaluate different features, including its relevance with respect to the query terms, influence over the field, "age" and relation with other documents. Many visual tools described in literature can display one or more of the above.Nevertheless, it is rare that all of them are shown in a single view at the same time: e.g., some systems [13,9] use different views to show relevance to terms and citation relations. In this paper we propose a novel interface, called CyBiS, acronym of Cylindrical Biplot System, which represents the salient data of searched documents in a single view. This way, the user is not compelled to switch between different visualizations thus loosing the focus on the current context. CyBiS shows document items as spheres embedded in a 3D cylinder placed horizontally on the screen. The relevance of a document to terms is given by height and depth of the item in the cylinder; the horizontal position, instead, represents document's publication year; the influence of the document is given by the color gradation of the item; citations are shown on-demand and are represented through connections.

Several operations are available in order to refine search.Firstly, the cylinder can rotate on the horizontal axis, in order to bring in the foreground the documents more related to a given term. It is also possible to zoom, which is equivalent to cut the cylinder along its horizontal axis, in order to select a time slice or to filter, for selecting a subset of items relevant to given terms. A prototypical implementation has been developed and used to produce an example session for evaluating the effectiveness of the proposed interface.

The rest of the paper is organized as follows: the next section contains a brief survey on BVTs; in section 3 the interface is described, with a particular reference to its objectives and main distinguishing features; the subsequent section presents the prototypical implementation, while section 5 describes the proposed example session; some final remarks and a brief discussion on possible future work conclude the paper.

2. Related Work

The main objectives of BVTs have been identified in [9]: though the main task is paper search, they can be also useful for related tasks: finding authors or topics and tasks which give an overall picture of the research in a field, such as studying the citation network, chronology or collaboration. In this brief survey we only consider systems focused on search.

Following a model given in [18], we classify BVTs interfaces according to the way in which the results of a search query are presented:

- analytical: an information item represents a single document;
- holistic: information items represent groups of documents sharing the same value of a categorization attribute;
- hybrid: aggregated and detailed aspects of a search result are presented simultaneously.

The most commonly used analytical interfaces are those using a graph representation. An acyclic direct graph is a natural representation for a scientific paper citation network: the nodes generally represent documents, while connections encode a direct citation.A simple interface based on graph visualization is used in the BIVTECI [13]. More complex is the graph used in Citespace II [3], in which, among other features, the node color is used to convey information related to variations in the influence of the article over time. A co-citation network is represented in [5]: undirected arcs are drawn between two items representing documents that are co-cited. The StarWalker system [4] enables a 3D collaborative exploration of the graph. In the graph used in the Lyberworld system [1] both documents and

Introduction

Contemporary scholars and scientists devote substantial effort to keep up with advances in their rapidly expanding fields.The growing number of publications combined with increasingly cross-disciplinary sources makes it challenging to follow emerging research fronts and identify key papers. It is even harder to begin exploring a new field without a starting frame of reference.

Researchers have vastly different levels of expertise and requirements for learning about scientific fields.A graduate student or cross-disciplinary researcher in a new field might find it useful to see the pivotal historical papers, key authors, and popular publication venues.On the other hand, a seasoned academic may be interested only in recent leading work and outlier papers or authors that challenge their preconceptions about the field. Grant program managers and review panel members sometimes have to examine fields they are not familiar with, looking for research trends, emerging theories, and open questions. Moreover, social scientists or scientometric analysts may be interested in how academic communities form over time, comparing citation and publication trends by country, or tracking the adoption of a single innovation.

Tools for rapid exploration of the literature can help ease these difficulties, providing readers with concise overviews tailored to their needs and aiding the generation of accurate surveys.Digital libraries and search engines are useful for finding particular papers or those matching a search string, but do not provide the additional analysis tools required to quickly summarize a field. Users unfamiliar with the field often find it challenging to search out the influential or groundbreaking papers, authors, and journals.

Specialized tools compute statistical measures and rankings to help identify items of interest, and other tools automatically summarize the text of multiple papers to extract key points. However, these tools are decoupled from the literature exploration task and are not easily integrated into the search process. Visualization techniques can be used to provide immediate overviews of publication and citation patterns in a field, but are uncommon in literature exploration tools. When present, they usually do not display much data (e.g., only ego networks) or provide the interaction techniques required to analyze the publication trends and research communities in a field. More ambitious goals for visualizations include helping users reach sufficient understanding to enable decision making, such as which fields are promising directions for researchers, appropriate for increased/reduced funding by government or industrial program managers, or worthy of investment by a venture capital organization.

This article presents the results of an effort to integrate statistics, text analytics, and visualization in a powerful prototype interface for researchers and analysts. The Action Science Explorer (ASE) 1 is designed to support exploration of a collection of papers so as to rapidly provide a summary, while identifying key papers, topics, and research groups. **ASE uses (a) bibliometric lexical link mining to create a citation network for a field and text 2 for each citation, (b) automatic summarization techniques to extract key points from papers, and (c) potent network analysis and visualization tools to aid in the exploration of relationships.** ASE, shown in Figure 1, presents the academic literature for a field using many different modalities: tables of papers, full texts, text summaries, and visualizations of the citation network and the groups it contains. Each view of the underlying data is coordinated such that papers selected in one view are highlighted in the others, providing additional metadata, text summaries, and statistical measure rankings about them. Users can filter by rankings or via search queries, highlighting the matching results in all views. Specifically, the contributions of this article are as follows:

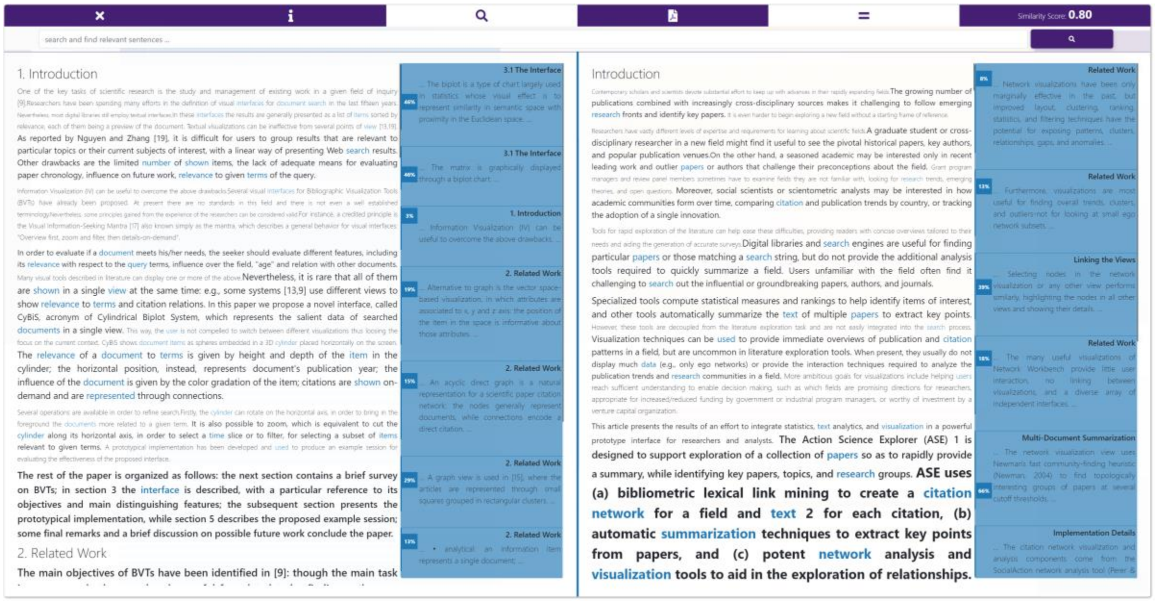
• A discussion of the motivation for creating a prototype literature exploration tool, its sophisticated design, and the challenges involved; • The novel integration of visualization with text analysis of citation texts and multi-document summarization; • Results from early

Figure 14 - A sample comparison lens snapshot

When researchers select the comparison lens from the lens menu, they can choose two documents from the canvas to start their comparison process (see Figure 13); and the comparison lens provides two adjacent panes to address the challenge of navigation from one document to the other and afford parallel reading for researchers. Unlike the abovementioned lenses, the comparison lens covers the whole canvas to represent the textual content of both documents properly.

As seen in Figure 14, similar to the skim lens, we encoded constructing sentences of each document with different font sizes. However, as mentioned in 3.1.7, the scores calculated for each sentence depend on the keywords of both documents. Thus, similar sentences would be displayed in bigger fonts than the other sentences to support researchers in discovering similarities of the selected documents quickly. Furthermore, the comparison lens provides a PDF view, and a map view in case researchers desire to compare some information items outside the textual content of the selected document.

Finally, aside from the similarity score of the two documents, which is displayed in the top right corner of the lens, the comparison lens also provides functionality like the semantic jump lens. Researchers can search for a specific concept and directly navigate to the sentences semantically similar to the search term in both documents, which helps them locate their desired sections of both documents rapidly (see Figure 15).



1. Introduction

One of the key tasks of scientific research is the study and management of existing work in a given field of inquiry [9].Researchers have been spending many efforts in the definition of visual interfaces for document search in the last fifteen years. Nevertheless, most digital libraries still employ textual interfaces.In these interfaces the results are generally presented as a list of items sorted by relevance, each of them being a preview of the document. Textual visualizations can be ineffective from several points of view [13,19]. As reported by Nguyen and Zhang [19], it is difficult for users to group results that are relevant to particular topics or their current subjects of interest, with a linear way of presenting Web search results. Other drawbacks are the limited number of shown items, the lack of adequate means for evaluating paper chronology, influence on future work, relevance to given terms of the query.

Information Visualization (IV) can be useful to overcome the above drawbacks.Several visual interfaces for Bibliographic Visualization Tools (BVTs) have already been proposed. At present there are no standards in this field and there is not even a well established terminology.Nevertheless, some principles gained from the experience of the researchers can be considered valid.For instance, a credited principle is the Visual Information-Seeking Mantra [17] also known simply as the mantra, which describes a general behavior for visual interfaces: "Overview first, zoom and filter, then details-on-demand".

In order to evaluate if a document meets his/her needs, the seeker should evaluate different features, including its relevance with respect to the query terms, influence over the field, "age" and relation with other documents. Many visual tools described in literature can display one or more of the above.Nevertheless, it is rare that all of them are shown in a single view at the same time: e.g., some systems [13,9] use different views to show relevance to terms and citation relations. In this paper we propose a novel interface, called CyBiS, acronym of Cylindrical Biplot System, which represents the salient data of searched documents in a single view. This way, the user is not compelled to switch between different visualizations thus loosing the focus on the current context. CyBiS shows document items as spheres embedded in a 3D cylinder placed horizontally on the screen. The relevance of a document to terms is given by height and depth of the item in the cylinder; the horizontal position, instead, represents document's publication year; the influence of the document is given by the color gradation of the item; citations are shown on-demand and are represented through connections.

Several operations are available in order to refine search.Firstly, the cylinder can rotate on the horizontal axis, in order to bring in the foreground the documents more related to a given term. It is also possible to zoom, which is equivalent to cut the cylinder along its horizontal axis, in order to select a time slice or to filter, for selecting a subset of items relevant to given terms. A prototypical implementation has been developed and used to produce an example session for evaluating the effectiveness of the proposed interface.

The rest of the paper is organized as follows: the next section contains a brief survey on BVTs; in section 3 the interface is described, with a particular reference to its objectives and main distinguishing features; the subsequent section presents the prototypical implementation, while section 5 describes the proposed example session; some final remarks and a brief discussion on possible future work conclude the paper.

2. Related Work

The main objectives of BVTs have been identified in [9]: though the main task

3.1 The Interface
... The biplot is a type of chart largely used in statistics whose visual effect is to represent similarity in semantic space with proximity in the Euclidean space. ...

3.1 The Interface
... The matrix is graphically displayed through a biplot chart. ...

1. Introduction
... Information Visualization (IV) can be useful to overcome the above drawbacks. ...

2. Related Work
... Alternative to graph is the vector space-based visualization, in which attributes are associated to x, y and z axis: the position of the item in the space is informative about those attributes. ...

2. Related Work
... An acyclic direct graph is a natural representation for a scientific paper citation network: the nodes generally represent documents, while connections encode a direct citation. ...

2. Related Work
... A graph view is used in [15], where the articles are represented through small squares grouped in rectangular clusters. ...

2. Related Work
... • analytical: an information item represents a single document; ...

Introduction

Contemporary scholars and scientists devote substantial effort to keep up with advances in their rapidly expanding fields.The growing number of publications combined with increasingly cross-disciplinary sources makes it challenging to follow emerging research fronts and identify key papers. It is even harder to begin exploring a new field without a starting frame of reference.

Researchers have vastly different levels of expertise and requirements for learning about scientific fields.A graduate student or cross-disciplinary researcher in a new field might find it useful to see the pivotal historical papers, key authors, and popular publication venues.On the other hand, a seasoned academic may be interested only in recent leading work and outlier papers or authors that challenge their preconceptions about the field. Grant program managers and review panel members sometimes have to examine fields they are not familiar with, looking for research trends, emerging theories, and open questions. Moreover, social scientists or scientometric analysts may be interested in how academic communities form over time, comparing citation and publication trends by country, or tracking the adoption of a single innovation.

Tools for rapid exploration of the literature can help ease these difficulties, providing readers with concise overviews tailored to their needs and aiding the generation of accurate surveys.Digital libraries and search engines are useful for finding particular papers or those matching a search string, but do not provide the additional analysis tools required to quickly summarize a field. Users unfamiliar with the field often find it challenging to search out the influential or groundbreaking papers, authors, and journals.

Specialized tools compute statistical measures and rankings to help identify items of interest, and other tools automatically summarize the text of multiple papers to extract key points. However, these tools are decoupled from the literature exploration task and are not easily integrated into the search process. Visualization techniques can be used to provide immediate overviews of publication and citation patterns in a field, but are uncommon in literature exploration tools. When present, they usually do not display much data (e.g., only ego networks) or provide the interaction techniques required to analyze the publication trends and research communities in a field. More ambitious goals for visualizations include helping users reach sufficient understanding to enable decision making, such as which fields are promising directions for researchers, appropriate for increased/reduced funding by government or industrial program managers, or worthy of investment by a venture capital organization.

This article presents the results of an effort to integrate statistics, text analytics, and visualization in a powerful prototype interface for researchers and analysts. The Action Science Explorer (ASE) 1 is designed to support exploration of a collection of papers so as to rapidly provide a summary, while identifying key papers, topics, and research groups. ASE uses (a) bibliometric lexical link mining to create a citation network for a field and text 2 for each citation, (b) automatic summarization techniques to extract key points from papers, and (c) potent network analysis and visualization tools to aid in the exploration of relationships.

Related Work
... Network visualizations have been only marginally effective in the past, but improved layout, clustering, ranking, statistics, and filtering techniques have the potential for exposing patterns, clusters, relationships, gaps, and anomalies. ...

Related Work
... Furthermore, visualizations are most useful for finding overall trends, clusters, and outliers-not for looking at small ego network subsets. ...

Linking the Views
... Selecting nodes in the network visualization or any other view performs similarly, highlighting the nodes in all other views and showing their details. ...

Related Work
... The many useful visualizations of Network Workbench provide little user interaction, no linking between visualizations, and a diverse array of independent interfaces. ...

Multi-Document Summarization
... The network visualization view uses Newman's fast community-finding heuristic (Newman, 2004) to find topologically interesting groups of papers at several cutoff thresholds. ...

Implementation Details
... The citation network visualization and analysis components come from the SocialAction network analysis tool (Perer &

Figure 15 - semantic jump functionality in comparison lens

In conclusion, using interactive lenses helped us design an integrated visualization to represent scientific information from a collection of documents and reduce the navigational interactions with VAS to afford the rapidity of the sensemaking process. It is also important to mention that the visualization techniques used in this VAS allow us to implement other lenses and cover more scientific information items without adding extra windows or navigational interactions. In the next section, we will examine our VAS in a scenario and compare its functionality with similar tools in the literature.

4.3 A usage and comparative scenario

In this section, we describe a scenario to demonstrate how our VAS can assist researchers in handling their research tasks involving a high number of scientific documents, and we also assess the strengths and shortcomings of our VAS compared to the related existing tools. For the purpose of this scenario, we introduce a hypothetical junior researcher, John, who is new to the field of Conversational AI and aims to write a report on the evolution of NLP-based chatbots over the past years.

In the first step, John aims to collect an adequate number of publications related to his research task; therefore, he runs a search on IEEE Xplore digital library[19] for conference and journal articles containing “chatbot” in their document title and related to natural language processing topic which returns 112 scientific document and their corresponding PDF files. In the next step, he must address the raised research questions by reviewing the collected scientific documents to carry out his research task. Questions such as the following possibly arise when conducting such a study:

- What are the existing definitions proposed for chatbots?
- What are the typologies, taxonomies, or domains for chatbots?
- What are the strategies and NLP techniques used to design different chatbots?
- Who are the notable researchers in this domain, and how different their perspectives towards chatbots are?
- What are the limitations of the existing chatbots?

To answer these questions, John cannot rely solely on the citation networks, extracted topics, or document clusters; instead, he should spot the key documents, drill into their textual content, and extract the information items related to his research question. Therefore, John starts his thorough study by uploading the collection of scientific documents onto the VAS. After uploading 112 scientific PDF documents, the VAS running on a Windows machine with a Core i7-6700HQ CPU and 16GB RAM took 5 minutes and 48 seconds to process all the documents and run a clustering algorithm on 88 documents which were properly parsed. In the next step, the canvas depicted in Figure 16 shows John an overview of the uploaded documents arranged by their publication date and clustered into two major groups.

[19] https://ieeexplore.ieee.org/

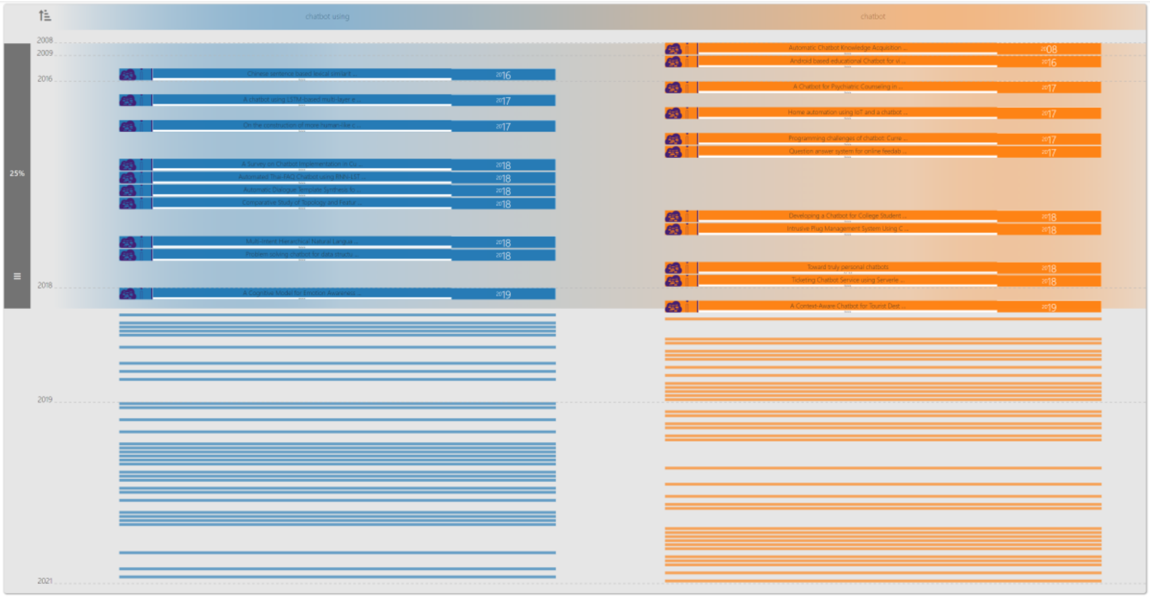

chatbot using
chatbot
2008
2009
2016
25%
2018
2019
2021
Chinese sentence based lexical similarit ... 2016
A chatbot using LSTM-based multi-layer e ... 2017
On the construction of more human-like c ... 2017
A Survey on Chatbot Implementation in Cu ... 2018
Automated Thai-FAQ Chatbot using RNN-LST ... 2018
Automatic Dialogue Template Synthesis fo ... 2018
Comparative Study of Topology and Featur ... 2018
Multi-Intent Hierarchical Natural Langua ... 2018
Problem solving chatbot for data structu ... 2018
A Cognitive Model for Emotion Awareness ... 2019
Automatic Chatbot Knowledge Acquisition ... 2008
Android based educational Chatbot for vi ... 2016
A Chatbot for Psychiatric Counseling in ... 2017
Home automation using IoT and a chatbot ... 2017
Programming challenges of chatbot: Curre ... 2017
Question answer system for online feedab ... 2017
Developing a Chatbot for College Student ... 2018
Intrusive Plug Management System Using C ... 2018
Toward truly personal chatbots 2018
Ticketing Chatbot Service using Serverle ... 2018
A Context-Aware Chatbot for Tourist Dest ... 2019

Figure 16 - Initial state of the canvas after uploading John's collected documents

Before any interaction with the VAS, John can discover that a higher number of documents were published between 2019 and 2021 than the period of 2008-2016, which shows the increase of attention towards this domain in the past years. Concerning similar tools, Paper Forager[62], which provides a histogram filter displaying the number of publications in each year, can also support John to infer such hidden patterns of the corpus in one view; however, the list view technique used by Jigsaw[61] and ASE[59] might require additional scrolling to provide such information.

As all the uploaded documents are approximately similar with regards to their content, the clustering module has only generated a minimum number of clusters: 1)“chatbot using” that contains most publications concerning the implementation of a chatbot in a specific language or an area; and 2) “chatbot” that contains mainly publications about the design considerations, challenges, methodologies, analysis, and reviews of chatbots; however, there can be publications from other categories in another cluster due to the similarity of all publications. John can modify the height of the sliding lens to scan document rectangles in more detail and drag the sliding lens from top to the bottom of the canvas to have a quick overview of the title, publication year, and cluster of each document.

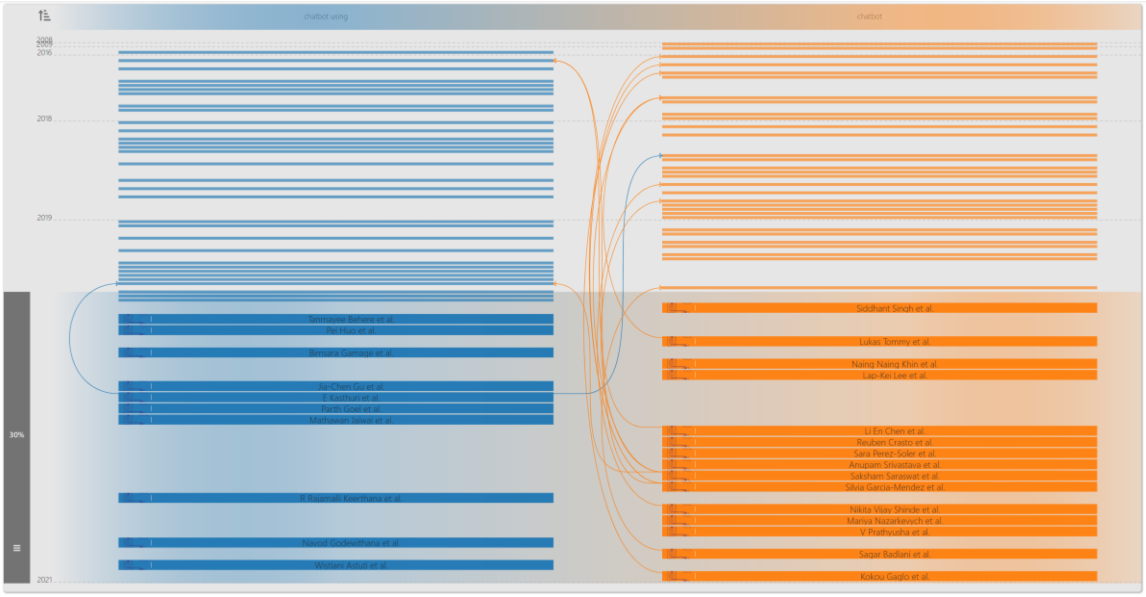


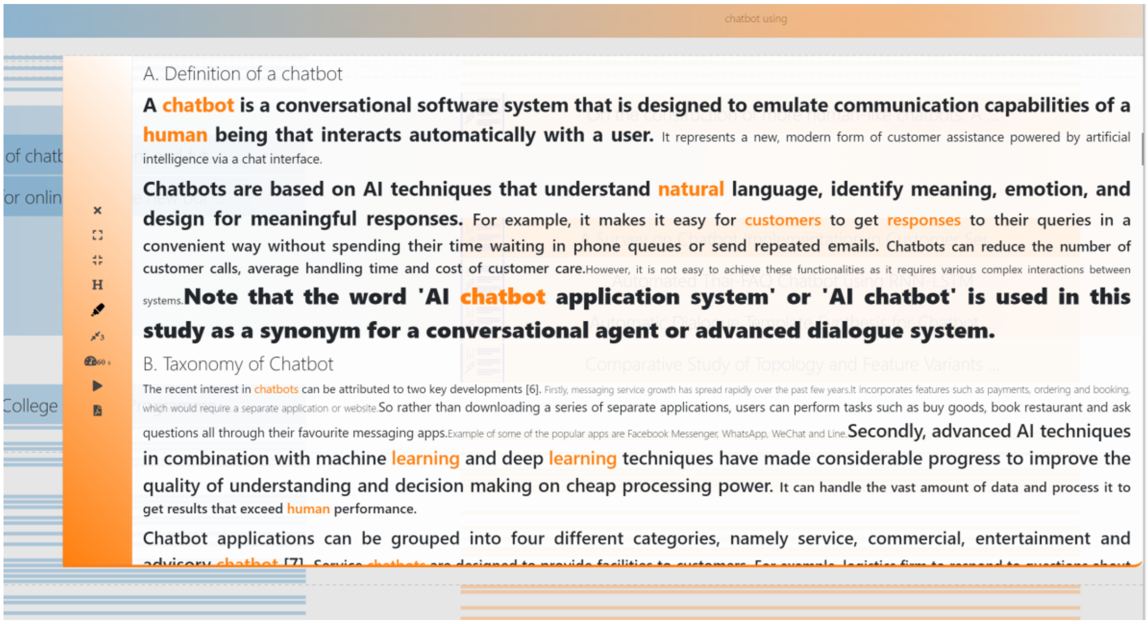
Figure 17 - Overview of the canvas when bibliographic connection lens is selected

There are several ways to identify key scientific documents and authors. One common approach is to rely on the bibliographic connections between the documents and the number of citations each document or its corresponding authors received. To support this approach, ASE enables researchers to rank document *nodes* based on their received citations and filter top desired documents for further analysis. One other approach is to find the most occurring topics and find the related documents and authors within the corpus. To exemplify, Jigsaw enables researchers to re-order its list views based on *frequency-of-occurrence*, connect the entities of different list views, and filter desired entities from each list.

The bibliographic connection lens provided in our VAS supports researchers to discover citation-based connections between scientific documents and locate documents with a higher density of connections around them as the most citing or cited documents (see Figure 17). However, in a research task like John’s, where approximately similar concepts are covered in all the documents and most of the documents are published in a recent short period, relying on the abovementioned approaches would not properly help John locate key documents and authors. Rather, he should look for documents that provide important information, such as survey reviews of existing chatbots, the novelty of implementation, and usage in different areas (e.g., education, healthcare, or customer service). Furthermore, as John’s primary research task is to examine the evolution of chatbot systems, he must analyze documents published in different years to trace the approaches adopted in different years.

By ordering documents based on their publication date and scanning the document rectangles using the map and bibliographic connection lenses, John perceived that the majority of the documents are concerned with implementing a chatbot system related to education, health care, human resource management, tourism and visiting guidance, recommendation system, and customer service. Furthermore, in case he encountered some documents whose titles were not informative enough, he could use the summarization lens to read the abstract section and condensed representation of the body of those documents to discover their research concerns and goals. Meanwhile, John also managed to locate several comparative surveys and review studies that could help him gain an overall insight into chatbots' definitions, taxonomies, and design challenges; therefore, he selected the skim lens from the lens menu drill into the content of these review studies. It is also important to mention that our VAS affords to switch lenses via the mouse scroll wheel to accelerate the process of lens selection and avoid any potential distraction from focusing on a specific document rectangle.

For the sake of conciseness of this scenario, we will describe the skimming process of only one of the survey documents that John aimed to examine, A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks[100]. Using the skim lens, John can have a quick view over the headings of this document by *headings view* option of the lens and discover that this document provides definition, taxonomy, analysis, and comparison of chatbots.



A. Definition of a chatbot

A chatbot is a conversational software system that is designed to emulate communication capabilities of a human being that interacts automatically with a user. It represents a new, modern form of customer assistance powered by artificial intelligence via a chat interface.

Chatbots are based on AI techniques that understand natural language, identify meaning, emotion, and design for meaningful responses. For example, it makes it easy for customers to get responses to their queries in a convenient way without spending their time waiting in phone queues or send repeated emails. Chatbots can reduce the number of customer calls, average handling time and cost of customer care.However, it is not easy to achieve these functionalities as it requires various complex interactions between systems.**Note that the word 'AI chatbot application system' or 'AI chatbot' is used in this study as a synonym for a conversational agent or advanced dialogue system.**

B. Taxonomy of Chatbot

The recent interest in chatbots can be attributed to two key developments [6]. Firstly, messaging service growth has spread rapidly over the past few years.It incorporates features such as payments, ordering and booking, which would require a separate application or website.So rather than downloading a series of separate applications, users can perform tasks such as buy goods, book restaurant and ask questions all through their favourite messaging apps.Example of some of the popular apps are Facebook Messenger, WhatsApp, WeChat and Line.Secondly, advanced AI techniques in combination with machine learning and deep learning techniques have made considerable progress to improve the quality of understanding and decision making on cheap processing power. It can handle the vast amount of data and process it to get results that exceed human performance.

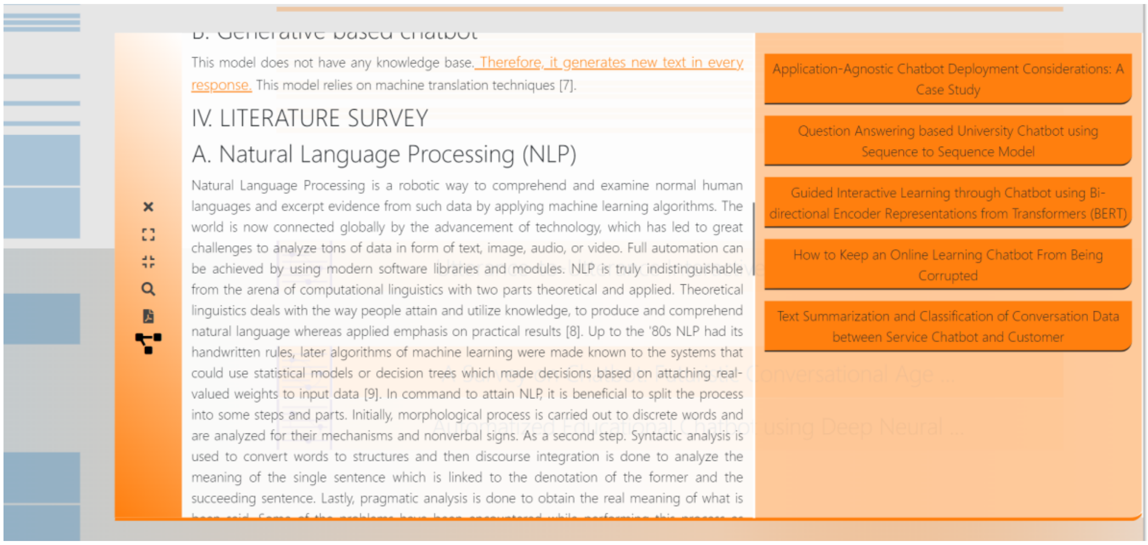
Chatbot applications can be grouped into four different categories, namely service, commercial, entertainment and advisory chatbot [7]. Service chatbots are designed to provide facilities to customers. For example, logistics firm to respond to questions about

Figure 18 - A snapshot of the skim lens representing the textual content of [100]

Therefore, as represented in Figure 18, when John scrolls through the document and gets to the Definition section, the skim lens encodes the important sentences with bigger font size to support John focus on the main definition, functionalities, and terminologies of chatbots rather than the provided examples in the section. In this way, John can skim survey documents effectively, gain an insight into several chatbots and their implementation and usability details, and update his internal representation of the document collection he is examining rapidly.

Concerning the skim lens, although our VAS manages to discover key sentences and gains researchers' attention by encoding techniques in an automated fashion; however, this lens can be enhanced by automatic auditory skimming and automatic navigation to referred figures or tables in the text similar to the application designed by [37].

After skimming the first set of key documents and getting familiar with fundamental concepts of chatbots, John can use the semantic jump lens to locate the documents related to each design strategy, methodology, or use case he has examined in the survey documents. To cite an example, when John gets familiar with generative based chatbots, chatbots without any knowledge base that generate new text in every response, in [101] and wants to find semantically similar documents in the corpus for further analysis or supporting evidence, he can select the *inter-document jump* option from the semantic jump lens and select on the sentence: “Therefore, it generates new text in every response” or any other similar sentence in that section.



B. Generative based chatbot

This model does not have any knowledge base. Therefore, it generates new text in every response. This model relies on machine translation techniques [7].

IV. LITERATURE SURVEY

A. Natural Language Processing (NLP)

Natural Language Processing is a robotic way to comprehend and examine normal human languages and excerpt evidence from such data by applying machine learning algorithms. The world is now connected globally by the advancement of technology, which has led to great challenges to analyze tons of data in form of text, image, audio, or video. Full automation can be achieved by using modern software libraries and modules. NLP is truly indistinguishable from the arena of computational linguistics with two parts theoretical and applied. Theoretical linguistics deals with the way people attain and utilize knowledge, to produce and comprehend natural language whereas applied emphasis on practical results [8]. Up to the '80s NLP had its handwritten rules, later algorithms of machine learning were made known to the systems that could use statistical models or decision trees which made decisions based on attaching real-valued weights to input data [9]. In command to attain NLP, it is beneficial to split the process into some steps and parts. Initially, morphological process is carried out to discrete words and are analyzed for their mechanisms and nonverbal signs. As a second step. Syntactic analysis is used to convert words to structures and then discourse integration is done to analyze the meaning of the single sentence which is linked to the denotation of the former and the succeeding sentence. Lastly, pragmatic analysis is done to obtain the real meaning of what is

Application-Agnostic Chatbot Deployment Considerations: A Case Study

Question Answering based University Chatbot using Sequence to Sequence Model

Guided Interactive Learning through Chatbot using Bi-directional Encoder Representations from Transformers (BERT)

How to Keep an Online Learning Chatbot From Being Corrupted

Text Summarization and Classification of Conversation Data between Service Chatbot and Customer

Figure 19 - A snapshot of the semantic jump lens while performing an inter-document jump from [101] to relevant documents based on a specific sentence

As represented in Figure 19, the semantic lens has retrieved five closest documents to the selected sentence, and by clicking on each of them, the current lens closes, and the lens corresponding to the selected document would open. As an example, when John clicks on the second document, Question Answering based University Chatbot using Sequence to Sequence Model[102], its lens would open, and he can drill into its textual content to discover the Seq2Seq and RNN encoder-decoder models used in the implementation of this chatbot to generate answers based on the input questions.

Concerning the semantical connection of documents, it is also important to mention that the *Document View* proposed by Jigsaw[61] enables researchers to select similar documents based on specific topic entities and drill into their textual content to trace concepts across different documents. Furthermore, ParallelTopic[44] also enables researchers to select documents based on a specific topic and drill into their details. However, the semantic jump lens provided in our VAS enables researchers to trace a custom concept rather than a topic extracted from the document across one or more documents based on the sentence embeddings.

After finding related documents to each design methodology, usage area, or implementation technique and limitation, John can pick pairs of documents to compare and address his research questions with a strong knowledge basis. However, as mentioned before, sensemaking is an iterative process; therefore, John might add some other scientific documents to his collection and examine the updated set of documents in the VAS.

In conclusion, this scenario shows that the non-scrollable canvas and its integration with interactive lenses can effectively support researchers to gain an overview over all the documents, discover bibliographic connections, locate key documents, drill into textual content of documents, trace specific concepts across the corpus to discover semantic connections, and rapidly compare desired documents during the sensemaking process. However, our VAS comes with several limitations as well that might undermine its advantages. The need for large screen displays for more effectiveness, lack of ability to annotate and create custom groups of scientific documents, and the inability to display document figures and tables in the lenses are some of the shortcomings of our VAS which can be addressed in the future.

Chapter 5

5 Conclusion

In this thesis, we examined ontological attributes of scientific documents and elaborated on potential sub-activities researchers might carry out during the sensemaking process of scientific documents. We also investigated the role of external representations and particularly interactive visual analytics systems in supporting researchers and enhancing their cognitive power to make sense of a collection of scientific documents.

However, due to the complexity and high load of information items extracted from scientific documents, many VASes try to cover a limited subset of information items (e.g., bibliographic information) and support researchers in basic sub-activities (e.g., initial search and exploration of scientific documents). In addition, the visualization techniques (e.g., node-link graphs, scatter plots, coordinated windows, etc.) used in the existing VASes to encode scientific information require researchers to have additional interactions with the tools to focus on a specific document.

The abovementioned challenges were our main motivations to design our VAS based on an innovative visualization that can support researchers in drilling into different information items of their desired scientific documents and remain aware of the other documents in the corpus. As mentioned before, the interactivity of a VAS can directly affect its efficiency in supporting its users; therefore, we aimed to provide optimized interactions both within the internal components of the system as well as the interactions related to the researchers and the system.

As shown in the usage scenario, displaying all the documents in one view equipped with a fisheye view sliding lens would support researchers to discover specific patterns within the corpus with minimum interactions. Furthermore, using interactive expanded lenses enabled researchers to access latent information layers of scientific documents when required. In this way, not only we managed to avoid high density in our visualization, but also, we developed a mechanism by which we could extend our visualization to encode more information items extracted from scientific documents. Therefore, by integrating machine learning and NLP techniques, Grobid PDF processing system, and visualization techniques, we managed to focus on the textual contents of scientific documents.

5.1 Discussion

Although rigorous and real-world evaluations of our system would be beneficial, the thorough examination and comparison conducted in 4.3 provided a basis to gain insight into the effectiveness, advantages, and shortcomings of our VAS. Aside from the holistic visualization canvas of our VAS, which mitigates the initial learning curve for researchers, it also supports researchers in multiple layers of abstraction. The VAS allows researchers to unwrap detailed information of each scientific document, address their broad set of initial questions and develop new questions throughout their investigation process, which is compatible with the iterative nature of the sensemaking process.

Since programming languages, frameworks, and technologies used in developing a VAS can directly affect the overall quality of its interactions on both internal and external levels, and meanwhile, the architecture of the system must be designed in such a level of abstraction to support the extensibility of the system, we designed a modular API using Python language to be able to plug the state-of-the-art machine learning algorithms and techniques into a module without affecting the overall structure of the API. Furthermore, the component-based design of lenses enables us to develop new analytical modules and insert their respective resulting data into the visualization by designing new lens components. Last but not least, implementing our visualization on a web-based platform using D3.js facilitated easier accessibility to the system.

Although our VAS has managed to address many challenges and shortcomings of previous similar tools; however, it comes with several limitations which could deteriorate its usability. Other than the need for large screen displays to be more beneficial and the restriction of using only PDF documents, our VAS has some other limitations which, we will examine in the next section.

5.2 Future Work

Firstly, although the textual content of scientific documents contains their main contributions and helps researchers extract their required information items, figures and tables of these documents might also be key sources of information. Due to the particular challenges related to the extraction of figures and tables, such as widely differing spacing conventions of scientific documents, avoiding false positives and still extracting various captions, and extracting vector graphics [103] which required thorough independent research, we decided to rely only on the textual content and provide *PDF view* option in our lenses in cases of requirement. However, navigating back and forth between PDF view and the main view of lenses could result in disconnection between researchers' mental process and lens representation. Therefore, efficient extraction of figures and tables of scientific documents and encoding them directly in the lenses could be a future enhancement of our VAS.

Secondly, to support the iterative nature of the sensemaking process more effectively, we could afford dynamic insert or removal of scientific documents. Furthermore, the columnar representation of clusters in our visualization could be utilized to enable researchers to create new clusters and group their desired set of documents in a separate cluster. In addition to grouping documents, we can add annotation-based interactions to our VAS to support researchers in organizing their discoveries throughout their investigation process.

References

[1] V. Sorge, D. Ahmetovic, C. Bernareggi, and J. Gardner, “Scientific Documents,” in *Web Accessibility*, Y. Yesilada and S. Harper, Eds. London: Springer London, 2019, pp. 397–415. doi: 10.1007/978-1-4471-7440-0_22.

[2] A. Öchsner, *Introduction to Scientific Publishing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-38646-6.

[3] M. Gusenbauer, “Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases,” *Scientometrics*, vol. 118, no. 1, pp. 177–214, Jan. 2019, doi: 10.1007/s11192-018-2958-5.

[4] “Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references.” https://asistdl-onlinelibrary-wiley-com.proxy1.lib.uwo.ca/doi/epdf/10.1002/asi.23329 (accessed Oct. 04, 2021).

[5] G. Marchionini, “Exploratory search: from finding to understanding,” *Commun. ACM*, vol. 49, no. 4, pp. 41–46, Apr. 2006, doi: 10.1145/1121949.1121979.

[6] “A Brief History of Research Synthesis - Iain Chalmers, Larry V. Hedges, Harris Cooper, 2002.” https://journals.sagepub.com/doi/abs/10.1177/0163278702025001003?casa_token=N7fjC2n9yJIAAAAA:ght2lt56K-vqZMMWyemHHhaCZR2UIC0YnkhBwQWY98PWrUk62mXaXn05VNRVvHOMoR_AuwxBUyO1_g (accessed Aug. 17, 2021).

[7] H. L. Colquhoun *et al.*, “Scoping reviews: time for clarity in definition, methods, and reporting,” *Journal of Clinical Epidemiology*, vol. 67, no. 12, pp. 1291–1294, Dec. 2014, doi: 10.1016/j.jclinepi.2014.03.013.

[8] H. Arksey and L. O’Malley, “Scoping studies: towards a methodological framework,” *International Journal of Social Research Methodology*, vol. 8, no. 1, pp. 19–32, Feb. 2005, doi: 10.1080/1364557032000119616.

[9] P. Pirolli and D. M. Russell, “Introduction to this Special Issue on Sensemaking,” *Human–Computer Interaction*, vol. 26, no. 1–2, pp. 1–8, Mar. 2011, doi: 10.1080/07370024.2011.556557.

[10] G. W. Furnas and D. M. Russell, “Making Sense of Sensemaking,” p. 2, 2005.

[11] D. M. Russell, M. J. Stefii, P. Pirolli, S. K. Card, and C. Hill, “The Cost Structure of Sensemaking,” p. 8.

[12] O. Buchel and K. Sedig, “Making Sense of Document Collections with Map-Based Visualizations,” p. 237, 2012.

[13] F. Osborne, E. Motta, and P. Mulholland, “Exploring Scholarly Data with Rexplore,” in *Advanced Information Systems Engineering*, vol. 7908, C. Salinesi, M. C. Norrie, and Ó. Pastor, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 460–477. doi: 10.1007/978-3-642-41335-3_29.

[14] F. Loizides and G. Buchanan, “An Empirical Study of User Navigation during Document Triage,” in *Research and Advanced Technology for Digital Libraries*, vol. 5714, M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, and G. Tsakonas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 138–149. doi: 10.1007/978-3-642-04346-8_15.

[15] G. Buchanan and F. Loizides, “Investigating Document Triage on Paper and Electronic Media,” in *Research and Advanced Technology for Digital Libraries*, Berlin, Heidelberg, 2007, pp. 416–427. doi: 10.1007/978-3-540-74851-9_35.

[16] H. de Ribaupierre and G. Falquet, “New trends for reading scientific documents,” in *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing - BooksOnline ’11*, Glasgow, Scotland, UK, 2011, p. 19. doi: 10.1145/2064058.2064064.

[17] F. Heimerl, Q. Han, S. Koch, and T. Ertl, “CiteRivers: Visual Analytics of Citation Patterns,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 190–199, Jan. 2016, doi: 10.1109/TVCG.2015.2467621.

[18] W. Tanner, E. Akbas, and M. Hasan, “Paper Recommendation Based on Citation Relation,” in *2019 IEEE International Conference on Big Data (Big Data)*, Dec. 2019, pp. 3053–3059. doi: 10.1109/BigData47090.2019.9006200.

[19] P. Pirolli and S. Card, “The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis,” Jan. 2005.

[20] D. Kirsh, “Interaction, External Representation and Sense Making,” Jan. 2009.

[21] Y. Qu and G. W. Furnas, “Sources of structure in sensemaking,” in *CHI ’05 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2005, pp. 1989–1992. doi: 10.1145/1056808.1057074.

[22] E. Hutchins, *Cognition in the Wild*. MIT Press, 1995.

[23] J. J. van Wijk, “The value of visualization,” in *VIS 05. IEEE Visualization, 2005.*, Oct. 2005, pp. 79–86. doi: 10.1109/VISUAL.2005.1532781.

[24] V. Kaptelinin and B. A. Nardi, *Acting with Technology: Activity Theory and Interaction Design*. MIT Press, 2006.

[25] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, “Visual Analytics: Definition, Process, and Challenges,” in *Information Visualization*, vol. 4950, A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 154–175. doi: 10.1007/978-3-540-70956-5_7.

[26] B. Fortuna, M. Grobelnik, and D. Mladenić, “Visualization of text document corpus,” *Informatica*, pp. 497–502, 2005.

[27] K. Sedig, “Interactive Mathematical Visualisations: Frameworks, Tools and Studies,” in *Trends in Interactive Visualization*, R. Liere, T. Adriaansen, and E. Zudilova-Seinstra, Eds. London: Springer London, 2009, pp. 343–363. doi: 10.1007/978-1-84800-269-2_16.

[28] H. Ji and W. Gan, “Data Visualization for Making Sense of Scientific Literature,” in *2020 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*, Jan. 2020, pp. 870–873. doi: 10.1109/ICITBS49701.2020.00191.

[29] B.-C. Björk, M. Laakso, P. Welling, and P. Paetau, “Anatomy of green open access,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 2, pp. 237–250, 2014, doi: 10.1002/asi.22963.

[30] C. Tenopir, D. W. King, S. Edwards, and L. Wu, “Electronic journals and changes in scholarly article seeking and reading patterns,” *Aslib Proceedings*, vol. 61, no. 1, pp. 5–32, Jan. 2009, doi: 10.1108/00012530910932267.

[31] J. Franze, K. Marriott, and M. Wybrow, “What academics want when reading digitally,” in *Proceedings of the 2014 ACM symposium on Document engineering*, New York, NY, USA, Sep. 2014, pp. 199–202. doi: 10.1145/2644866.2644894.

[32] K. E. Hubbard and S. D. Dunbar, “Perceptions of scientific research literature and strategies for reading papers depend on academic career stage,” *PLOS ONE*, vol. 12, no. 12, p. e0189753, Dec. 2017, doi: 10.1371/journal.pone.0189753.

[33] T. Hillesund, “Digital Reading Spaces: How Expert Readers handle Books, the Web and Electronic Paper,” *First Monday*, 2010, doi: 10.5210/FM.V15I4.2762.

[34] Z. Liu, “Reading behavior in the digital environment: Changes in reading behavior over the past ten years,” *Journal of Documentation*, vol. 61, no. 6, pp. 700–712, Jan. 2005, doi: 10.1108/00220410510632040.

[35] R. Kopak and C. Chiang, “An interactive reading environment for online scholarly journals: The Open Journal Systems reading tools,” *OCLC Systems & Services: International digital library perspectives*, vol. 25, no. 2, pp. 114–124, Jan. 2009, doi: 10.1108/10650750910961910.

[36] B. Schilit, G. Golovchinsky, and M. Price, “Beyond Paper: Supporting Active Reading with Free Form Digital Ink Annotations,” *Conference on Human Factors in Computing Systems - Proceedings*, Aug. 1999, doi: 10.1145/274644.274680.

[37] T. A. Khan, D. Yoon, and J. McGrenere, “Designing an Eyes-Reduced Document Skimming App for Situational Impairments,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2020, pp. 1–14. doi: 10.1145/3313831.3376641.

[38] K. Sedig and P. Parsons, “Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework,” *Synthesis Lectures on Visualization*, vol. 4, no. 1, pp. 1–185, Apr. 2016, doi: 10.2200/S00685ED1V01Y201512VIS005.

[39] A. Sinha *et al.*, “An Overview of Microsoft Academic Service (MAS) and Applications,” in *Proceedings of the 24th International Conference on World Wide Web*, New York, NY, USA, May 2015, pp. 243–246. doi: 10.1145/2740908.2742839.

[40] J. Liu, T. Tang, W. Wang, B. Xu, X. Kong, and F. Xia, “A Survey of Scholarly Data Visualization,” *IEEE Access*, vol. 6, pp. 19205–19221, 2018, doi: 10.1109/ACCESS.2018.2815030.

[41] C. Zhang, Z. Li, and J. Zhang, “A survey on visualization for scientific literature topics,” *J Vis*, vol. 21, no. 2, pp. 321–335, Apr. 2018, doi: 10.1007/s12650-017-0462-2.

[42] P. Federico, F. Heimerl, S. Koch, and S. Miksch, “A Survey on Visual Approaches for Analyzing Scientific Literature and Patents,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 9, pp. 2179–2198, Sep. 2017, doi: 10.1109/TVCG.2016.2610422.

[43] G. Costagliola and V. Fuccella, “CyBiS: A Novel Interface for Searching Scientific Documents,” in *2011 15th International Conference on Information Visualisation*, Jul. 2011, pp. 276–281. doi: 10.1109/IV.2011.95.

[44] W. Dou, X. Wang, R. Chang, and W. Ribarsky, “ParallelTopics: A probabilistic approach to exploring document collections,” in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2011, pp. 231–240. doi: 10.1109/VAST.2011.6102461.

[45] X. Jiang and J. Zhang, “A text visualization method for cross-domain research topic mining,” *J Vis*, vol. 19, no. 3, pp. 561–576, Aug. 2016, doi: 10.1007/s12650-015-0323-9.

[46] D. Fried and S. G. Kobourov, “Maps of Computer Science,” in *2014 IEEE Pacific Visualization Symposium*, Mar. 2014, pp. 113–120. doi: 10.1109/PacificVis.2014.47.

[47] “TileBars: Visualization of Term Distribution Information in Full Text Information Access.”

[48] L. Nowell, R. France, D. Hix, L. Heath, and E. Fox, “Visualizing Search Results: Some Alternatives to Query-Document Similarity,” Jan. 1996, vol. 19, pp. 67–75. doi: 10.1145/243199.243214.

[49] S. Havre, E. Hetzler, K. Perrine, E. Jurrus, and N. Miller, “Interactive visualization of multiple query results,” in *IEEE Symposium on Information Visualization, 2001. INFOVIS 2001.*, Oct. 2001, pp. 105–112. doi: 10.1109/INFVIS.2001.963287.

[50] J.-K. Chou and C.-K. Yang, “PaperVis: Literature Review Made Easy,” *Computer Graphics Forum*, vol. 30, no. 3, pp. 721–730, 2011, doi: 10.1111/j.1467-8659.2011.01921.x.

[51] H. Hauser, “Generalizing Focus+Context Visualization,” in *Scientific Visualization: The Visual Extraction of Knowledge from Data*, G.-P. Bonneau, T. Ertl, and G. M. Nielson, Eds. Berlin/Heidelberg: Springer-Verlag, 2006, pp. 305–327. doi: 10.1007/3-540-30790-7_18.

[52] J. C. Roberts, “State of the Art: Coordinated amp; Multiple Views in Exploratory Visualization,” in *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, Jul. 2007, pp. 61–71. doi: 10.1109/CMV.2007.20.

[53] C. Tominski, S. Gladisch, U. Kister, R. Dachselt, and H. Schumann, “Interactive Lenses for Visualization: An Extended Survey,” *Computer Graphics Forum*, vol. 36, no. 6, pp. 173–200, 2017, doi: 10.1111/cgf.12871.

[54] K. Kucher and A. Kerren, “Text visualization techniques: Taxonomy, visual survey, and community insights,” in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, Apr. 2015, pp. 117–121. doi: 10.1109/PACIFICVIS.2015.7156366.

[55] K. Sedig, P. Parsons, and A. Babanski, “Towards a Characterization of Interactivity in Visual Analytics,” vol. 3, no. 1, p. 17, 2012.

[56] P. Parsons and K. Sedig, “Adjustable properties of visual representations: Improving the quality of human-information interaction,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 3, pp. 455–482, 2014, doi: 10.1002/asi.23002.

[57] P. Parsons, K. Sedig, A. Didandeh, and A. Khosravi, “Interactivity in Visual Analytics: Use of Conceptual Frameworks to Support Human-Centered Design of a Decision-Support Tool,” in *2015 48th Hawaii International Conference on System Sciences*, Jan. 2015, pp. 1138–1147. doi: 10.1109/HICSS.2015.138.

[58] K. Sedig and P. Parsons, “Interaction Design for Complex Cognitive Activities with Visual Representations: A Pattern-Based Approach,” *AIS Transactions on Human-Computer Interaction*, vol. 5, no. 2, pp. 84–133, Jun. 2013.

[59] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr, “Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization,” *Journal of the American Society for Information Science and Technology*, vol. 63, no. 12, pp. 2351–2369, 2012, doi: 10.1002/asi.22652.

[60] J. Stasko, C. Görg, and Z. Liu, “Jigsaw: Supporting Investigative Analysis through Interactive Visualization,” *Information Visualization*, vol. 7, no. 2, pp. 118–132, Jun. 2008, doi: 10.1057/palgrave.ivs.9500180.

[61] C. Görg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. Stasko, “Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw,” *IEEE Trans Vis Comput Graph*, vol. 19, no. 10, pp. 1646–1663, Oct. 2013, doi: 10.1109/tvcg.2012.324.

[62] J. Matejka, T. Grossman, and G. Fitzmaurice, “Paper Forager: Supporting the Rapid Exploration of Research Document Collections,” presented at the Graphics Interface 2021, Apr. 2021. Accessed: Oct. 16, 2021. [Online]. Available: https://openreview.net/forum?id=QhN4tUZd8r

[63] P. Riehmann, M. Potthast, B. Stein, and B. Froehlich, “Visual Assessment of Alleged Plagiarism Cases,” *Computer Graphics Forum*, vol. 34, no. 3, pp. 61–70, 2015, doi: 10.1111/cgf.12618.

[64] S. Wu, Y. Zhao, F. Parvinzamir, N. Th. Ersotelos, H. Wei, and F. Dong, “Literature Explorer: effective retrieval of scientific documents through nonparametric thematic topic detection,” *Vis Comput*, vol. 36, no. 7, pp. 1337–1354, Jul. 2020, doi: 10.1007/s00371-019-01721-7.

[65] S. Koch, H. Bosch, M. Giereth, and T. Ertl, “Iterative integration of visual insights during patent search and analysis,” in *2009 IEEE Symposium on Visual Analytics Science and Technology*, Oct. 2009, pp. 203–210. doi: 10.1109/VAST.2009.5333564.

[66] C. Chen, “CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 359–377, 2006, doi: 10.1002/asi.20317.

[67] J. Choo, C. Lee, C. K. Reddy, and H. Park, “UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1992–2001, Dec. 2013, doi: 10.1109/TVCG.2013.212.

[68] N. J. van Eck and L. Waltman, “Software survey: VOSviewer, a computer program for bibliometric mapping,” *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010, doi: 10.1007/s11192-009-0146-3.

[69] J. Matejka, T. Grossman, and G. Fitzmaurice, “Citeology: visualizing paper genealogy,” in *CHI ’12 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, May 2012, pp. 181–190. doi: 10.1145/2212776.2212796.

[70] N. J. van Eck and L. Waltman, “CitNetExplorer: A new software tool for analyzing and visualizing citation networks,” *Journal of Informetrics*, vol. 8, no. 4, pp. 802–823, Oct. 2014, doi: 10.1016/j.joi.2014.07.006.

[71] P. Li, X. Jiang, and H. Shatkay, “Figure and caption extraction from biomedical documents,” *Bioinformatics*, vol. 35, no. 21, pp. 4381–4388, Nov. 2019, doi: 10.1093/bioinformatics/btz228.

[72] S. M. Sohan, C. Anslow, and F. Maurer, “A Case Study of Web API Evolution,” in *2015 IEEE World Congress on Services*, Jun. 2015, pp. 245–252. doi: 10.1109/SERVICES.2015.43.

[73] “Python Release Python 3.9.7,” *Python.org*. https://www.python.org/downloads/release/python-397/ (accessed Oct. 18, 2021).

[74] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*. O’Reilly Media, Inc., 2018.

[75] *GROBID*. 2021. [Online]. Available: https://github.com/kermitt2/grobid

[76] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[77] P. Berkhin, “A Survey of Clustering Data Mining Techniques,” in *Grouping Multidimensional Data: Recent Advances in Clustering*, J. Kogan, C. Nicholas, and M. Teboulle, Eds. Berlin, Heidelberg: Springer, 2006, pp. 25–71. doi: 10.1007/3-540-28349-8_2.

[78] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979, doi: 10.2307/2346830.

[79] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[80] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *arXiv:1908.10084 [cs]*, Aug. 2019, Accessed: Oct. 20, 2021. [Online]. Available: http://arxiv.org/abs/1908.10084

[81] A. M. Rush, S. Chopra, and J. Weston, “A Neural Attention Model for Abstractive Sentence Summarization,” *arXiv:1509.00685 [cs]*, Sep. 2015, Accessed: Oct. 21, 2021. [Online]. Available: http://arxiv.org/abs/1509.00685

[82] J. L. Neto, A. A. Freitas, and C. A. A. Kaestner, “Automatic Text Summarization Using a Machine Learning Approach,” in *Advances in Artificial Intelligence*, Berlin, Heidelberg, 2002, pp. 205–215. doi: 10.1007/3-540-36127-8_20.

[83] S. Chopra, M. Auli, and A. M. Rush, “Abstractive Sentence Summarization with Attentive Recurrent Neural Networks,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, Jun. 2016, pp. 93–98. doi: 10.18653/v1/N16-1012.

[84] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: a survey,” *Artif Intell Rev*, vol. 47, no. 1, pp. 1–66, Jan. 2017, doi: 10.1007/s10462-016-9475-9.

[85] S. Erera *et al.*, “A Summarization System for Scientific Documents,” *arXiv:1908.11152 [cs]*, Aug. 2019, Accessed: Oct. 21, 2021. [Online]. Available: http://arxiv.org/abs/1908.11152

[86] S. J. Green, “Building hypertext links by computing semantic similarity,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 5, pp. 713–730, Sep. 1999, doi: 10.1109/69.806932.

[87] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv:1810.04805 [cs]*, May 2019, Accessed: Oct. 22, 2021. [Online]. Available: http://arxiv.org/abs/1810.04805

[88] M. Bostock, V. Ogievetsky, and J. Heer, “D^3 Data-Driven Documents,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011, doi: 10.1109/TVCG.2011.185.

[89] G. W. Furnas, “Generalized fisheye views,” *SIGCHI Bull.*, vol. 17, no. 4, pp. 16–23, Apr. 1986, doi: 10.1145/22339.22342.

[90] R. Borgo *et al.*, “Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications,” May 2013, p. %pages_from%-%pages_to%. doi: 10.2312/conf/EG2013/stars/039-063.

[91] W. Peng, M. O. Ward, and E. A. Rundensteiner, “Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering,” in *IEEE Symposium on Information Visualization*, Oct. 2004, pp. 89–96. doi: 10.1109/INFVIS.2004.15.

[92] D. Kirsh, “Complementary Strategies: Why We Use Our Hands When We Think,” *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, no. T, pp. 161–175, 1995.

[93] W. S. Cleveland and R. McGill, “Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods,” *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984, doi: 10.2307/2288400.

[94] S. R. Dos Santos, “A framework for the visualization of multidimensional and multivariate data,” phd, University of Leeds, 2004. Accessed: Oct. 24, 2021. [Online]. Available: https://etheses.whiterose.ac.uk/1316/

[95] S. G. Eick, “Visual discovery and analysis,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 44–58, Jan. 2000, doi: 10.1109/2945.841120.

[96] E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose, “Toolglass and Magic Lenses: The See-Through Interface,” p. 8.

[97] R. M. Shiffrin and W. Schneider, “Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory,” *Psychological Review*, vol. 84, no. 2, pp. 127–190, 1977, doi: 10.1037/0033-295X.84.2.127.

[98] R. Brath and E. Banissi, “Using Typography to Expand the Design Space of Data Visualization,” *She Ji: The Journal of Design, Economics, and Innovation*, vol. 2, no. 1, pp. 59–87, Mar. 2016, doi: 10.1016/j.sheji.2016.05.003.

[99] T. Sanocki and M. C. Dyson, “Letter processing and font information during reading: Beyond distinctiveness, where vision meets design,” *Atten Percept Psychophys*, vol. 74, no. 1, pp. 132–145, Jan. 2012, doi: 10.3758/s13414-011-0220-9.

[100] M. Nuruzzaman and O. K. Hussain, “A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks,” in *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, Oct. 2018, pp. 54–61. doi: 10.1109/ICEBE.2018.00019.

[101] P. Goel and A. Ganatra, “A Survey on Chatbot: Futuristic Conversational Agent for User Interaction,” in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, May 2021, pp. 736–740. doi: 10.1109/ICSPC51351.2021.9451763.

[102] N. N. Khin and K. M. Soe, “Question Answering based University Chatbot using Sequence to Sequence Model,” in *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Nov. 2020, pp. 55–59. doi: 10.1109/O-COCOSDA50338.2020.9295021.

[103] C. Clark and S. Divvala, “PDFFigures 2.0: Mining figures from research papers,” in *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, Jun. 2016, pp. 143–152.

Curriculum Vitae

Name:	Amirreza Haghverdiloo Barzegar
Post-secondary Education and Degrees:	Kharazmi University Tehran, Iran 2015-2019 B.Sc.
Related Work Experience	Graduate Teaching Assistant University of Western Ontario 2020-2021
	Research Assistant University of Western Ontario 2020-2021
Awards	WGRS scholarship for 2020-2021, University of Western Ontario