VISUALIZING KEYWORD CO-OCCURRENCE FROM TWO QUERIES TO AID EXPLORATORY SEARCH

by

Poppy Nicolette Riddle

Submitted in partial fulfilment of the requirements
for the degree of Master of Information

at

Dalhousie University
Halifax, Nova Scotia
April 2022

Dalhousie University is located in Mi'kma'ki,
the ancestral and unceded territory of the Mi'kmaq.
We are all Treaty people.

DEDICATION PAGE

This body of work represents a huge involvement and commitment by my spouse and daughter, who have given up playtime, weekends, and even moved across a continent. This thesis has been supported by copious baked goods, pots of lovely tea, patient listening, and a constant outpouring of love, support, and encouragement. This would not have happened without you, and while its not your cup of tea, I have found joy in my work!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study presents a method of keyword co-occurrences network visualization comparing two separate queries of a commercial database as a visual aid to mitigate an information seeker's uncertainty during the exploratory search process. The visualization compares keyword co-occurrence between two database queries illustrating how search terms have keyword exclusions that are unique to each and may inform the user of work might be missed or what might refine future searches. Results suggest that while some keywords are co-occurring and linked, there are exclusive keywords that may impact an information seeker's desired topic and may provide valuable information to support decision making in their exploratory search process. Interpretation of results using a design perspective suggests this visualization would present more benefits to the information seeker as an interactive visualization than a static one. The outcomes of this study may be implemented by an academic librarian with beginner level Python programming experience.

GLOSSARY

IR – Information retrieval includes "studies of peoples systems searching practices and typically task-oriented" (Case & Given, 2016, p. 15) research including the internet, online databases and library catalogs. It may include the types and forms of data being sought such as books, images, or tabular data. Relevance and context of the information seeker are among many complex internal and external factors that are studied within IR. Many information behavior models include information retrieval, but this is associated with the task, not the area of research.

Keyword – The author specified and indexed words or phrases with each peer-reviewed article, typically following the abstract. Verbene (2016) defines them as "short phrases that represent the content of a document or a document collection". Recognition of use of keywords is convenient and explainable to an information seeker as they can see where they are from and how they are derived from the document.

Keyword exclusions – These are the author defined keywords from each document that have not been found in one of the two searches. In this study, they are used to quantify dissimilarity with the other search.

Overlap and co-occurrence – Both terms refer to the occurrence of a keyword in both search results. They are used interchangeably throughout the literature, so it is a convenience to do the same here.

Search term – refers to the words typed into the WoS search engine to create results. This becomes a search string by the time a date range and restrictions on document types are applied. The use of search term in this study refers the words used in this process. Also called search query, it is a series of linked words along with Boolean operators, truncation, and other modifiers used to call information from a database (Smith & Wong, 2016).

Search result – refers to the list of documents received from the search in WoS and downloaded as a full record with bibliographic data.

WoS – Web of Science is a subscription-based set of databases for academic publications created by the Institute for Scientific Information (ISI) and currently owned by Clarivate.

ACKNOWLEDGEMENTS

Katy Börner's book, 10 years ago, created an indelible impression on me and it gently nudged me into this current stream, filling my life with lovely people.

Dr. Philippe Mongeon, my supervisor, who introduced me to the world of data, bibliometrics and demystified the whole of scientific structure. His continued confidence and support are inspiring.

Colin Conrad, who introduced me to a whole new world of coding and whose confidence in me and encouragement has been invaluable.

Alison Brown, whose aspirational good will and humour act as a beacon of how to lead.

Janet Music, whose patience, trust, and willingness to take time to connect has made for one of the most thrilling adventures I've been on!

Sam Taylor has shifted my perception of my abilities – I can do this.

SIM students with whom I've thoroughly enjoyed their company on this adventure and hope to continue to get to know them – thank you, Tomato-folks!

CHAPTER 1 INTRODUCTION

Introduction

1.1 The scenario

Imagine a student looking up the search term “social justice” as vast amounts of literature are returned. Expanding their query with “social injustice”, it is unclear what terms would help this student stay within their field of interest. Resolving this issue requires considerable investment in time, computational resources, and mental resources. Currently, it is not possible to see how your search terms relate to one another and how they are used within fields. If the student could see the relationships between associated keywords grouped by topics in a single, interactive image, it may enable them to choose terms within their intended topic for a more refined and targeted search.

1.1.1 Comprehensive statement

There is exponential growth of knowledge being produced and searching within it is messy due to its complex interconnections and terminological variations. This messiness is problematic for researchers and students as information seekers *to understand the whole* and to *understand how the parts relate*. In the academic context, students are working with little prior knowledge to make sense of the messiness, which in this context, is very challenging for *gaining subject knowledge* and *effective exploration of topics*. Current discovery systems in academic settings use *a list-based results page* resulting in a *lack of familiarity* of the breadth of a topic and *cognitive overload* resulting in *uncertainty* and anxiety during exploratory searching. Information visualization has proven potential to address uncertainty by revealing an overall view of a topic but has not yet made its way into the information retrieval systems we use. There

is a need to develop structural information visualizations which compare search strategies to better understand how topics relate with other topics and how search terms can exclude some topics. Information visualization may help information seekers reduce the time and cognitive resources it takes to gain knowledge of a topic and improve their confidence during exploratory search.

1.1.2 Breaking down the comprehensive statement

There continues to be an accelerated increase in scholarly production over the last 40 years, particularly with digital publishing. With increasingly vast amounts of information, information seekers, regardless of their expertise level, may "drown in this vast ocean of scholarly communication" (*Greeting Note for the 10th BIR Workshop*, 2020, 0:38) without the continual development of improved tools for exploration and IR. Shiffrin & Börner (2004) similarly use a sea-based metaphor to describe the increasing difficulty of "fishing this sea of desired information" (p. 5183).

Information seekers, such as students, using academic discovery systems engage with complex information spaces which can contribute to increased cognitive load (Demelo & Sedig, 2021; Dillon, 2000), increased time to attain comprehension due to overloading of short-term memory (Dillon, 2000), and risk of abandoning the search (White & Roth, 2009), or rejection of the technology (Dillon, 2000). Information spaces are naturally created by those trying to understand new information relationships, and physical space and mapping metaphors are convenient and well established (Demelo & Sedig, 2021; Dillon, 2000) to aid information navigation, communication, and tracking progress through some construct of order (Dillon, 2000), thus providing some shape for the information.

This messiness is problematic for academic information seekers *to understand the whole* and to *understand the parts of the whole*. In other words, how can they search effectively using terminology that is precise when they are unaware or unsure of the terminology within a field, or the right terminology to describe a phenomenon, or the conceptual limits of a term? Knowing the differences between terms and their conceptual frameworks allows one to communicate their goals precisely and to construct valid arguments. As an example of delineations of concepts, Bate's (2005) definition of information and knowledge evolved over years and numerous responses to objections. The scope of a term is important knowledge for an exploratory searcher to gain, especially understanding the limitations which shape conceptual frameworks (González-Valiente et al., 2021). Definitions also shift over time, or with context such as the linguistic variations in meaning between fields. In the scenario for this study, the exploratory searcher needs to understand the difference between "social justice" and "social injustice" to make decisions about what is relevant to pursue. Understanding the underlying ideological assumptions, cultural influences, or political interests from which a researcher or student may argue (Ayers et al., 2009) will be critical to them to formulate information seeking goals that yields relevant and useful literature.

In the academic context, many information seekers are working with little prior knowledge to make sense of that messiness, which in this context, is problematic for *gaining subject knowledge* and *effective exploration of topics*. I have used a student as a proxy for the novice information seeker, but this could equally apply to anyone performing an academic search based on their subject area knowledge and experience. So, novice information seeker, attempts to avert biased associations with student and instead attempts to relate this persona to anyone who needs to know more about how extensive a search term encapsulates a topic.

For an information seeker who is investigating an unknown subject area, they rather naturally engage in exploratory search – an open-ended, complex, dynamic search process for information that generally establishes topical boundaries within which to refine a more focused search. As part of an information seeking process leading to information retrieval (IR), exploratory search includes the behaviors of the information seeker informing, learning and critically considering fields, topics, search terms, authors, and sources that may be relevant.

Current discovery systems in academic settings are designed for precise retrieval and while used for exploratory search, require extensive time and cognitive resources to be successful at this task. Discovery systems used in libraries often use list-based results pages and "their strong dependency on precise searcher-generated queries" (He et al., 2019 citing di Sciascio, Sabol, & Veas, 2016). Discovery systems largely remain unchanged and the list-based results are difficult to remember or recover as topical interests drift during exploratory search, which has been found to be "cumbersome" (di Sciascio et al., 2016).

Drawing from cognitive load theory, (Sweller et al., 2011), *cognitive load* refers to the amount of working memory and cognitive capabilities utilized to learn new information (Agostinho et al., 2014). We can learn efficiently provided that information can be inputted by sensory memory, processed by working memory and then stored for later retrieval by long-term memory. However, as working memory is limited in its capacity and duration it can hold information, overloading can easily occur while being exposed to lots of new information that must be interpreted consciously. "[W]orking memory can hold between five and nine elements of novel unfamiliar information, or even less depending on the nature of processing (e.g., if some information must be contrasted or combined)" (Agostinho et al., 2014, p. 532). In the context of the information seeker, cognitive overload can easily occur when given long lists, when drifting

from one topic to another (Kangasrääsiö et al., 2016), or following one article after another (di Sciascio et al., 2016), in an attempt by the information seeker to understand the breadth of a topic. The way to work with the limited capacity and duration of working memory, is to help it link to patterns, or schemas, in long-term memory. The schema can be a pattern of moves, behaviours, wiring diagrams, or even prose (Agostinho et al., 2014). Think of schemas as the big picture.

Without an overall understanding or big picture, the information seeker experiences a *lack of familiarity* with the breadth of a subject area and thus restricting their ability to recognize and judge the relevancy of one topic or its keywords over others. This increases uncertainty (Kuhlthau, 1993, 1999a), and feelings of 'doubt, confusion, frustration and anxiety' (Kuhlthau, 2004), which may cause the exploratory information seeker to quit the search.

Information visualization is generally used for "scrutinizing the cognitive and theoretical aspects" (Kim et al., 2016) or the intellectual structure of data in contrast with data visualization in which "dimensionality reduction" (Kim et al., 2016) and quantitative representations are of more importance. Closely related to data visualization in its use of "computational techniques, algorithms and mathematics" (Kim et al., 2016), these terms are often used interchangeably. In the context of this study, information visualization will be used to communicate qualitative relationships rather than absolutes to illustrate the intellectual structure of information spaces. Information visualization may be able to address uncertainty by relieving cognitive overload, revealing an overall view of a topic, revealing patterns, enabling inferences, and improving mental models of the information space. While there are many studies and applications that have been developed, they are rarely integrated with academic IR interfaces as I will explore in Chapter 2.

Current IR systems use a faceted approach which allows the user to drill-down and achieve precision results. However, they are not so good at exploring the topics within unstructured data, such as comparing lists of keywords between documents (di Sciascio et al., 2016). The benefits of information visualization of unstructured data for topic exploration, may "foster analytical understanding of Boolean-type queries" since they do not provide any ranking or relevance score (di Sciascio et al., 2016) unlike current academic IR systems. "Well-designed interactive interfaces can effectively address information overload issues that may arise due to limited attention span and human capacity to absorb information at once" (di Sciascio et al., 2016).

Information visualization that provides an overall view of topics within a subject area may help information seekers emotionally during exploratory search, by reducing the time and cognitive resources it takes to gain knowledge of a topic and may improve their confidence. Studies have shown that improved knowledge of overall structure reduces cognitive load (di Sciascio et al., 2016; Munzner & Maguire, 2015), the knowledge about the breadth of topics aids in redefining broad terms (González-Valiente et al., 2021), that prior exposure improves awareness of search term origins (Hienert & Lusky, 2018), and that information visualization improves cognitive maps and improves learning tasks in complex information environments (Demelo & Sedig, 2021). If information seekers can experience reduced cognitive load and improved cognitive maps, then the cognitive state of uncertainty and its 'affected symptoms of anxiety and lack of confidence' (Kuhlthau, 1999, p. 401) may be reduced.

There is a need to develop structural information visualizations which compare search strategies to better understand how topics relate with other topics and how search terms can exclude some topics. Kuhlthau called this a 'zone of intervention' (Kuhlthau, 2004) as it represents an opportunity to address deficiencies in the current state of the art and due to ever changing

conditions. This study investigated a method to create a **2D distance-based map** comparing two search results to show the effects of **search term choice** during the **exploratory search process** and how each search result may include content excluded from the other. As an example, a search for "social justice" from a commercial database yields results including multiple fields or topics. Altering this search term with a narrower, broader, or related terms, such as "social injustice", produced different results. The intent of this bibliometric-enhanced visualization and its comparison with existing bibliometric techniques is to investigate a method of keyword co-occurrence visualization that can be further developed into a dynamic query application for use in academic libraries to address the emotional challenges identified by White & Roth (2009) of the information seeker during exploratory search.

1.2 Purpose of the study

The purpose of this study is to investigate how visualizing keyword similarity, co-occurrences, and exclusions in academic database search results may improve the exploratory search process. I suggest visualizing as a means of understanding the effects of changing search terms and their results by showing their co-occurrences and exclusions distributed across a spectrum. In short, it seeks to provide a comparative visualization of the structure of the two searches. The outcomes of this study are to provide a visualization method upon which to build a more robust application for usability testing with academic information seekers. It has been the intent to keep the methods and software used in this study to be reproducible, have low computational resource requirements, and may be implemented by an academic librarian with beginner level Python programming experience.

1.3 Research objectives

This study seeks to answer two research objectives. It will do this by providing a visualization comparing the keywords from documents retrieved from two search results. Topics will be identified by clusters and results will show the most frequent keywords from those clusters, in an attempt to reveal the topical breadth of the intellectual structure. Results will also compare the two search results using bibliographic coupling, which is a bibliometric technique used to identify communities of documents that have similarity due to their shared references. The research objectives can be summarized as an exploration to provide a visual representation of how one search may or may not exclude communities that the searcher should be aware of.

- RO1: To present an effective way of visualizing the effects of search term choice so that included or excluded keywords might be compared between the two search results.
- RO2: Compare the keyword-level analysis with a document-level analysis using a proven bibliometric method to determine if topical exclusivity exists between search terms.

Figure 1 illustrates the concept of comparing two search results and being able to identify the co-occurring keywords that both results share, and the exclusive keywords that are unique to each search.

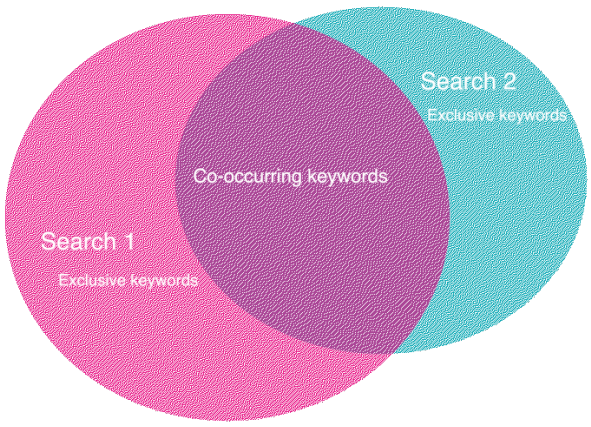

Figure 1: Venn Diagram illustrating the basic concept of keyword co-occurrence and exclusion.

1.3.1 Statement of the general theoretical and methodological approach

This method study explores the use of quantitative methods derived from bibliometrics to compare keyword distributions and clusters of co-occurrence to create a distance-based network map. Understanding how the visualized map may benefit the information seeker will be evaluated from a heuristic design perspective using design guidelines for interactive visualization (Card et al., 1999), data visualization (Munzner & Maguire, 2015), and graphic design (Kosslyn, 2006).

Referencing White & Roths's model of exploratory search (White & Roth, 2009), which establishes the emotional impact of uncertainty during the information seeking process, this study also depends on social constructivist theoretical perspectives (Kuhlthau, 2004) for situating the context of the information seeker (Agarwal, 2018) and their interaction with information visualization as mediations (Demelo & Sedig, 2021; Kuhlthau, 2004). I also acknowledge the privilege and power of map-makers, and the limitations of information maps as social constructions (De Bellis, 2009) for the user.

1.3.2 Research contributions

With this thesis, I seek to make the following contributions to future research through a heuristic analysis of a visualization that compares two search results. The first contribution investigates a means of visual representation comparing the author assigned keywords from two exploratory search results with clusters of co-occurrences. This artifact may provide two benefits to the information seeker. It may provide evidence to make choices that enable search completion (Kuhlthau et al., 2012), thus reducing the uncertainty during the information seeking process (Dervin, 1999; Kuhlthau, 1999a). It may also enable exploration of new subtopics or subfields

improving their information pathways (Greyson, 2019) and support creative (White & Roth, 2009) and serendipitous discovery (McCay-Peet & Toms, 2018).

The second contribution includes a Python-based notebook, available on Github (https://github.com/poppy-nicolette/search_term_comparison), that can be used for purposes beyond search term comparison, such as to analyze thesauri for evaluation of terminological similarity or overlap. Furthermore, the concept and its artifact of comparing two search results visually will be utilized in future work to develop a distributable interface to aid academic librarians working with researchers or scholars during exploratory search.

1.4 Originality of study

This work is similar to numerous prior studies and applications in its intent to address difficulties facing the novice information seeker yet presents a unique comparison of keywords from two search results to support *exploratory* search decisions. While exploratory search challenges for novice users have been addressed across a wide breadth of fields, including Human-Computer Interaction, Computer Science, Library & Information Sciences, Computer-Human Interfaces, and Psychology with launched applications, interactive displays, and speculative proposals (Federico et al., 2017), the digital domain is cruel and short-lived with the many no longer working or accessible. Although I cite significant older work, I'll review currently available, applications with similar objectives in Chapter 2.

1.5 Delimitations

For this study, a clarification between information seeking behavior models and information retrieval (IR) is needed. IR is defined loosely as finding materials from large collections to precisely satisfy a user's information need (Mayr & Scharnhorst, 2015). It is a hugely complex

area of study that bridges user studies, data models, and information systems (Mayr & Scharnhorst, 2015). Three significant difficulties arise in IR: vagueness of indexing terms, information overload, and term-based rankings that do not meet user needs (Mayr, 2016). Also related is bibliometric-enhanced information retrieval (BIR), which advocates for using tools for bibliometric analysis, such as footnotes, citations, authors, and journals as stratagems to improve the search process (Mayr, 2016) while at the same time blending bibliometrics, scientometrics and informetrics (Cronin & Sugimoto, 2014; Mayr & Scharnhorst, 2015). BIR researchers have identified current challenges, such as the need for more tools for exploring open information, possibly addressing the issue of inclusivity in databases by permitting further search and discovery based on emerging language (*Openness, Transparency, and Inclusivity in Science*, 2021) and resisting the standardization of language (Berube et al., 2018; *Openness, Transparency, and Inclusivity in Science*, 2021). Information seeking models, on the other hand, seek to understand and describe the process the user follows to satisfy an information need (Case & Given, 2016). Older studies typically wholistically framed the entire process from search to retrieval and more recent studies have investigated aspects within information seeking models for more clarity and definition.

While all these overlap and distinctions are grey, this study draws primarily from information seeking models for definition of the emotional impacts of searching, and the broad field of LIS and the more recent field of BIR for work on digital tools indented to address exploratory search challenges. As I was starting from a bibliometric perspective, I exhausted the LIS and BIR literature for studies that not only provided a base from which to ground my research, but may have theoretical and practical common ground. Naturally, Human-Computer Interaction fields, (HCI), Computer-Human Interfaces, (CHI), both within computer science, are also involved in

this problem space and while I have drawn from some of those fields, they have not been exhaustively explored. Federico et al. (2017) have created a systematic review of the state of the art of visualizing methods for scientific literature and patents, but as it is 5 years old now, another is badly needed.

Challenges extend beyond the emotional and cognitive impacts including language and cultural challenges where systems are designed for English speakers and Western interface design conventions. Issues of computational efficiencies, network connectivity, and the limitations of managed databases are outside the scope of this study's consideration. This visualization also privileges sight-normative persons and issues of accessibility have not yet been considered. Visualization in map form as well as tables are notoriously poorly considered for the needs of those using assistive technology and the outcomes of this study fail in that regard. These challenges will be confronted in Chapter 6 for future work.

1.6 Structure of the thesis

In Chapter 1, I have presented an imagined, but plausible scenario of a student performing an academic search to learn more about a topic. As an exploratory search process, the student, or information seeker as they'll be called throughout this text, is faced with messy information full of interconnections and variations in meaning. With little prior knowledge, the uncertainty of how to comprehend a new topic is time consuming and cognitively exhausting, heightening emotional states of anxiety and frustration. Search terms reveal list after list of returned relevant or citation-based results and are difficult to compare within current academic retrieval systems. However, with the mediation of information visualization that provides a visual representation of how search terms relate and simultaneously offers a way to explore topics within the subject

area, it may be possible to reduce time, cognitive load, and anxiety during the exploratory search process.

Chapter 2 will review and synthesize the relevant literature. I will explain the exploratory search process, how it fits within information seeking behavior models, and focus on the challenge of uncertainty inherent in exploratory search. The models introduced help us to see how exploratory search is experienced emotionally, with a roller coaster of uncertainty, anxiety, frustration, relief, and confidence. To understand how uncertainty can be mitigated, I will examine how the information seeker creates mental constructs and how information visualization can aid in their construction. In addition to the benefits of information visualization, I will introduce an evaluative framework for information visualization and guides for effective graphic design. I will then review appropriate bibliometric methods that have been used before to address exploratory search challenges with visualization and past studies and applications in pursuit of similar goals.

Chapter 3 will discuss the methods explored in this thesis including introducing the methods I've chosen for this study: keyword co-occurrence, the Leiden algorithm, and bibliographic coupling. I'll explain the sources of data, workflow and process of analysis, and implementation, including other software used, such as Gephi and VOSViewer. Where unique coding was required, I have tried to incorporate code that is computationally lightweight and is usable by a person with limited Python coding experience.

Chapter 4 will present the results including visualizations with supporting evidence using detail views, tables and graphs. This study is focused on the methods rather than evaluating the data as the objective of this study. However, to explain how the visualization works or compares to others, it will be necessary to talk about what narrative is emerging from the data.

Chapter 5 will discuss the results and interpretation and how findings support the mitigation of uncertainty. Primarily this will focus on how the visualization is working, but as a I introduced a scenario in Chapter 1, it will be necessary to explain how the visualization might be interpreted and utilized from the perspective of the student persona. I will reflect on some points from Chapter 2, such as what benefits are being achieved by this visualization.

Chapter 6 will revisit the position of this research, summarize key findings, suggest practical applications, and suggest the importance of the study for its practical applications. I'll also discuss future research directions, of which there are many. As this study has been understood to be part of a larger drive to create useful things for library users, the future research directions identified are more for my own knowledge and skill development rather than gaps systematically identified in the academic literature.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

In this chapter, in four sections, I will explain exploratory search, how it fits within information seeking behavior models, and how information visualization may improve the exploratory search process. In the section on exploratory search, I will look at White & Roth's (2009) exploratory search model and define and discuss the challenges of information seekers experiencing uncertainty during the process. In the second section on information seeking behavior models, I position exploratory search relative to information seeking models, such as an updated White & Roth model by Hoeber & Shukla (2022) and Kuhlthau's Information Seeking Process (ISP) model (Kuhlthau, 2004) illustrating how uncertainty and other emotions add risk to the exploratory search process. I will discuss what uncertainty is and how it can be mitigated with an information seeker's construction of mental maps. The third section will look at how information visualization can help, and I will review literature guiding visual communication in best practices to create compelling and effective visualizations. In the last section, I'll review bibliometric methods appropriate for this study's research objectives and review prior studies and recent applications that enable comparing the keywords from two search results to identify exclusion and co-occurrence for the exploratory searcher. I'll wrap up this chapter acknowledging the limitations of visualizations as a mediation.

2.2 Exploratory search

2.2.1 Motivation and purpose of exploratory search

Exploratory search is defined by White & Roth (2009) as activities in which people need to learn about a topic to understand how to achieve their goal. They may even be unsure how they will achieve their goal, or not yet formulated a goal. As part of a basic need, motivations to search for

information may come from a desire to learn, a need to satisfy missing information, or a lack of understanding of the terminology within a subject (Marchionini, 2006). It may be applied to many situations from seeking social connections to targeting academic resources. It is usually related to complex information problems outside of the topical or subject areas of the information seeker's expertise, in a process that includes uncertainty and subsequent other emotional impacts. The lack of familiarity with a subject area, its subdisciplines and even the language that is unique and different between subject experts and novices (Zhang et al., 2008), adds to increase the knowledge gap for the information seeker and is the central challenge this study seeks to address.

2.2.2 White & Roth's model of Exploratory Search

Building upon earlier models of exploratory search by Marchionini (2006), White & Roth's (2009) expanded definition of exploratory search includes the information seeker beginning to search with or without a fully formed idea of their goal. Exploratory search is comprised of two sub-activities: *exploratory browsing* and *focused searching* (White & Roth, 2009). *Exploratory browsing* is intended to find the relevant knowledge, improve topic knowledge, and permit serendipitous discoveries. Within exploratory browsing, the process includes *discovery*, *learning* and *investigation*. *Discovery* is the process of encountering new, unknown information, *learning* is an internalization leading to understanding, and *investigation* is an analytical process in which one critically assesses the new information for relevancy.

As an information seeker becomes more confident and less uncertain, they may engage in the second part of exploratory search, *focused searching*. This is a more purposeful query of the information source, such as an academic database, with three activities: *query (re)formulation, result examination,* and *information extraction. Query (re)formulation* is a process of coming up

with a query, trying it, and changing it so that relevant information is provided. *Result examination* is the process of assessing the information for relevance, and *information extraction* is the process of selecting and saving or using the relevant information (White & Roth, 2009).

Users that are unfamiliar with their fields tend to bounce between the two sub-activities of *exploratory search* and *focused search* to establish familiarity prior to making decisions of information use, establishing their goal, or even abandoning their search. While the sub-activities involve smaller identifiable actions, the main point is that the exploratory search process involves both a browsing process to learn, and a focused process to test information for relevance. The exploratory process captures the intent to improve comprehension of a subject area by exploring a wide range of materials to enable one to recognize terms from a wide range of topics within the subject area. White & Roth define exploratory search, particularly the exploratory phases of *discovery, learning*, and *investigation*, as providing this breadth of exposure. This is in contrast with an iterative search process which tend to refine in a targeted area as shown in Figure 2.

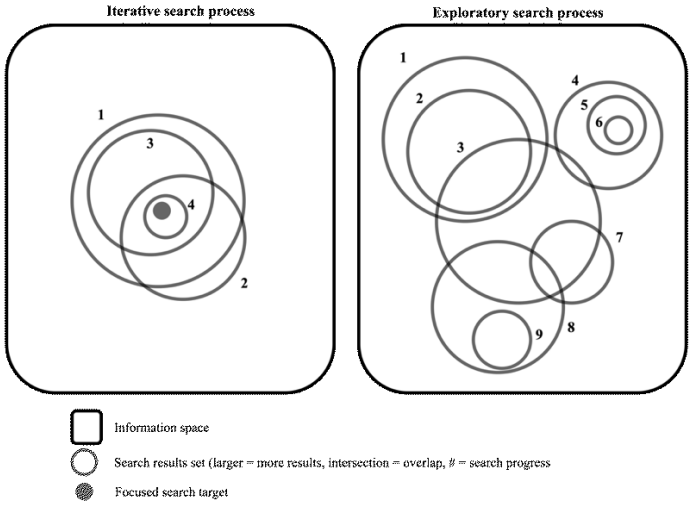

Figure 2: Model of exploratory search process compared with an iterative search process. From (White & Roth, 2009).

2.3 Information seeking behavior models

2.3.1 What is it?

Information seeking is a conscious effort to gather information in reaction to a need or an identified gap in one's knowledge. Information seeking behaviors can include conscious, unintentional, serendipitous, or even purposeful avoidance behaviors (Case & Given, 2016). As it usually focuses on person-centered research, information seeking behavior models are frameworks that explain how people deal with information seeking and usually account for contextual elements. Contextual elements can include one's cultural norms, geographical location, time-dependent factors, or in the case of this study, the emotional affects (Case & Given, 2016).

Exploratory searching fits within several information seeking behavior models such as the berrypicking model by Bates (1989), Dervin's sense-making models and concepts of gaps (Dervin, 1998), Scharnhorst's problem-solving framework within "unknown knowledge landscapes" (Scharnhorst, 2001), and Kuhlthau's model of information seeking process that specifically indicates the emotional impact on the seeker (Kuhlthau, 1991). White & Roth (2009) explain the intersection with many information seeking or search models in greater detail. I focus on Kuhlthau's model as most appropriate for this study due to its recognition of the emotional impact and the author's continued research to understand the significance of emotional uncertainty (Kuhlthau, 1991, 1999a, 1999b).

2.3.2 Explaining uncertainty

The White & Roth's (2009) model of exploratory search emphasizes uncertainty and how it impacts the information seeker throughout the process. Hoeber & Shukla (2022) take White & Roth's exploratory search model and present both exploratory search phases of *exploratory*

browsing and *focused searching* along with the concept of uncertainty rebounds, illustrating the non-linear nature of the exploratory search process and its emotional ups and downs as a counter to any interpretations that the process is linear sequence. The purpose of Hoeber & Shukla's integrated model is to show the importance of addressing the repeated experiences of uncertainty during the exploratory search process.

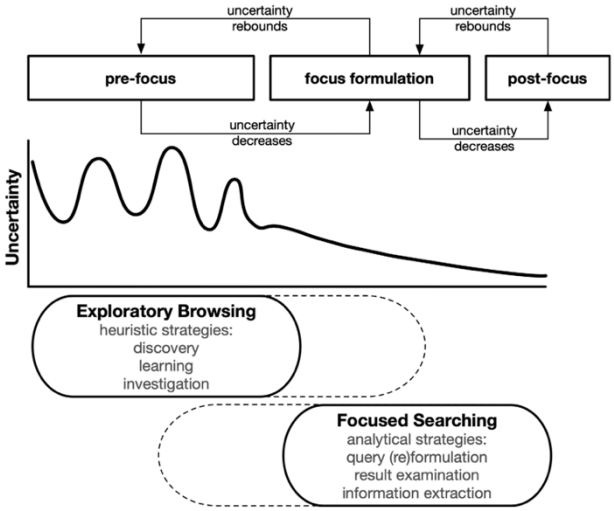

Figure 3: Model of the exploratory search process emphasizing the importance of uncertainty during the exploratory browsing phase, from Hoeber and Shukla (2022).

Kuhlthaus' model of the Information Search Process, or ISP (Figure 4), is important to defining uncertainty as it recognizes the significant emotional affects of the search process (Kuhlthau, 2004). The ISP contains 6 stages: *initiation, selection, exploration, formulation, collection*, and *presentation*. Each stage is associated with feelings commonly experienced by information seekers, regardless of their context as students, faculty, or professionals (Kuhlthau, 1999, 2004).

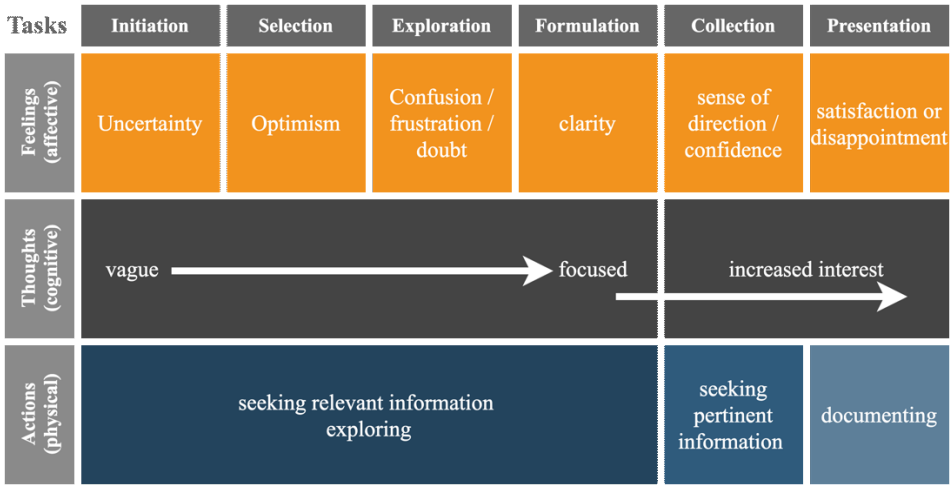

Figure 4: Kuhlthau's Information Search Process, (ISP) model. From (Kuhlthau, 2004).

The *initiation* stage represents the start of the information need, a metacognitive recognition that current knowledge may not be sufficient and there is a need for more. Uncertainty is very high at this stage.

The *selection* stage applies to topic selection for further inquiry. Studies by Kuhlthau (2004) found that criteria for selection included personal interests, external requirements, information availability, and time constraints.

In the *exploration* stage, those studied by Kuhlthau reported this stage as more heuristic, a following of the gut, in the search for information which can be "disorderly and confusing" (Kuhlthau, 1999a, 2004). Feelings of confusion, frustration, and doubt are common at this stage.

The *formulation* stage represents the culmination of increasing recognition of terms or topics and identification of relevant sources, which lead to setting or refining goals for the information seeking process. Although this is still a difficult stage involving thoughts that "spiral" (Kuhlthau, 2004, p. 83), evolve or emerge, emotions will lead to clarity indicating a shift from the negatively perceived emotions of the previous stage.

The *collection* stage involves retrieving documents that meet the relevance goals established previously and is limited by time, exhausting known resources, or some measure of having enough. Feelings of confidence increase at this stage as goals are being acted upon.

The *presentation* stage is the culmination of the process into applying the retrieved information into some form that satisfies the external or internal need for information, which, in the academic environment, is most frequently some written artefact. Generally, emotional states are of satisfaction and some intrinsic sense of reward, though feelings of boredom or disappointment can also occur.

The *formulation, collection*, and *presentation* stages represent an emotional transition from the negative feelings to more positively perceived ones of clarity, confidence, and satisfaction. The last stage can result in disappointment however, which may lead some information seekers to repeat the process.

The *exploration* stage is the height of emotional impact where the search is at risk and is where White & Roth explain the process more fully as their own model. Although White & Roth placed their exploratory search within many models, the overlap of White & Roth's model within Kuhlthau's ISP is easy to comprehend. White & Roth identified uncertainty as a key aspect that represents considerable risk to the exploratory search process. Kuhlthau also identifies it as a place of opportunity for librarians to seek solutions, or a "zone of intervention' (Kuhlthau, 2004, p. 128).

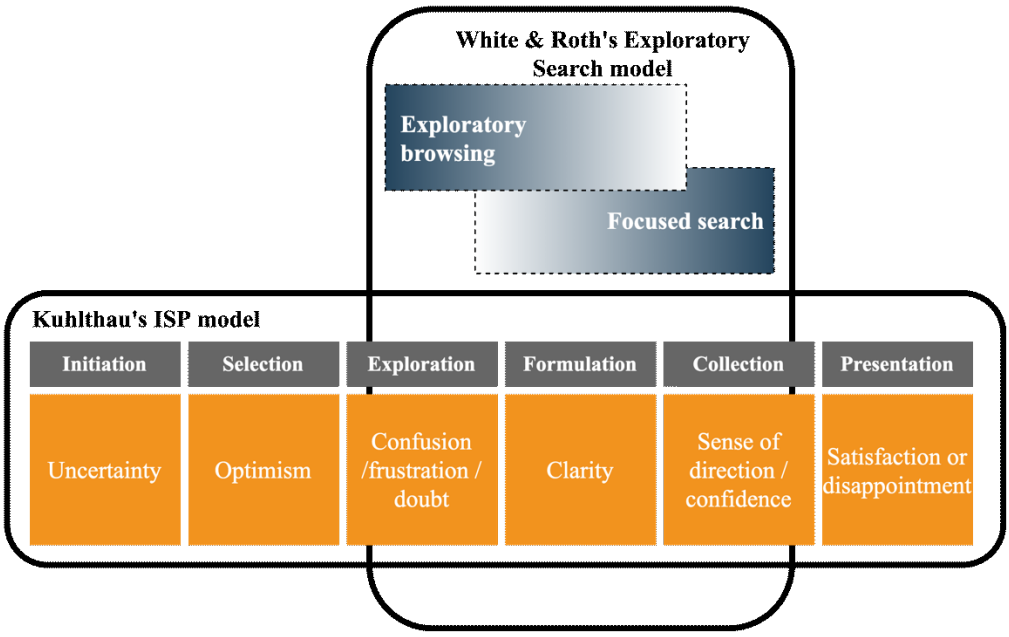

Figure 5: White & Roth's Exploratory Search model overlapping with Kuhlthau's ISP model.

2.3.3 How does one mitigate uncertainty?

Kuhlthau, White & Roth, and Hoeber & Shukla validate uncertainty as a real condition that deserves mediation. Uncertainty is established as an emotional state in both initiating and during the search process, due to a "lack of understanding, a gap in meaning, or a limited construct" (Kuhlthau, 2008, p. 33). It is the last contributing factor, that the outcome of this study seeks to address as a mediation. Moving from uncertainty and its affective emotional states can be facilitated by applying some of the six corollaries from Kuhlthau's borrowed, constructivist uncertainty principle (Kelly, 1963 cited in Kuhlthau, 2004). Of the six corollaries, process, formulation, redundancy, mood, prediction, and interest, the process corollary explains information seeking and its emotional affects as a dynamic process of meaning construction in which the information seeker creates meaning, makes associations, and interprets information, influenced by emotions and levels of expertise.

Other studies have provided insight into similar deconstructions as the process corollary within information seeking behaviors. Freksa (1999) identified a systematic method of how to create wayfinding systems. The concept of wayfinding systems, a process of making mental landmarks and creating a cognitive model, sometimes spatial, as a process of understanding complex information relationships. Wayfinding and waypoints, just as you might experience using a map and physical landmarks to get from Point A to Point B across town, may aid in helping information seekers bridge the gap between White & Roths' exploratory browsing and formulation stages or crossing the gap between Kuhlthau's exploratory and formulation stages, where we see improvements in emotional states and resolution to uncertainty.

Construction of meaning within the process corollary is further explained by the concept of cognitive maps. "When we encounter unfamiliar complex objects, we use our perception, intuition, and reasoning to form a mental model of their parts, relationships, and behaviors...When encounters present us complex objects that describe a space, like distance, position, or orientation, our cognitive processes form a specific type of mental model, the cognitive map" (Demelo & Sedig, 2021, p. 2).

The construction of some mental reference model is a critical piece of the puzzle to help aid the information seeker to relieve, shorten, or possibly reduce uncertainty and its resultant other emotions. The purpose of doing so, is to reduce the risk of the information seeker abandoning the search (White & Roth, 2009), reduce the task complexity (Demelo & Sedig, 2021; Dillon, 2000; Larsen & Ingwersen, 2005), and possibly reduce the time from *exploratory browsing* to *focused searching*.

So, we now understand how the various information seeking and exploratory search models are interrelated, how uncertainty is a common theme among them, and how uncertainty has been deconstructed to highlight the importance of helping the information seeker construct cognitive maps. We also understand the *exploratory browsing* stage as significant in that uncertainty is at its highest, compounded by the dynamic, spiralling, non-linear, heuristic nature of the information seeker as they attempt to cross the gap from *exploratory browsing*, to formulating some goal and move towards more confident states of *focused searching*. Now that we know the challenges, how they have been framed previously, and how we can deconstruct uncertainty with purpose to improve cognitive map formation, how can information visualization help?

To summarize exploratory search and its scenario, exploratory information seekers may be inexperienced with the fields in which their query lies, may be unsure about a process, or even unsure about their desired purpose for their search. Exploratory search is different from other models of searching actions as it can occur over multiple sessions, be open-ended, contextual, fluid and enabling not only decision-making for retrieval, but also topic orientation and learning subfields related to desired topics. Generally, a combination of browsing and focused searching, reducing disorientation (Dillon, 2000) and uncertainty provides a framework, or cognitive map for decision making (Demelo & Sedig, 2021; Larsen & Ingwersen, 2005) that can be communicated with other people for either individual or group benefits (White & Roth, 2009).

2.4 Information visualization as a tool

2.4.1 How is visualization helping?

The use of visualizations is appropriate as it can decrease cognitive burdens (He et al., 2019) and has been shown to increase engagement (Jiang & Fitzgerald, 2019), and close the gap between novice and experts regarding subject areas (Wu & Vakkari, 2018b). Information visualizations

benefit the exploratory searcher in six ways: (a)expanding working memory/cognitive resources (or relieving cognitive burden); (b)visualizing overall structure; (c)enabling patterns to be recognized; (d)permitting inferences of relationships through proximity; (e)improved cognitive maps and recognition of small changes; and (f)ideally, permitting the user to interact and change the data as expertise increases usually through parametric controls (Börner, 2010; Card et al., 1999). These are high-level goals and the visual embodiment itself should follow guidelines of visual communication that support the high-level goals (Kosslyn, 2006; Munzner & Maguire, 2015).

2.4.1.1 Expanding working memory/cognitive resources (or relieving cognitive burden)

Demelo & Sedig (2021) investigated how to improve cognitive map formation for users to visualize complex ontologies. As exploratory search involves learning about new topics or information spaces, the information seeker is likely faced with unstructured data from which to attempt to create some knowledge and recognition of the information space. As information visualization tools have been found to benefit task-learning, visualizations can also aid in cognitive map formations (Demelo & Sedig, 2021). Key to cognitive map formation is supporting a "staged process with repeated encounters" in which the user can utilize their sensory and cognitive systems to develop familiarity in the creation of an internal representation (Demelo & Sedig, 2021). They utilize the term *landmark* in the establishment of known areas related to unknown areas. This may seem familiar and synonymous with the previously discussed concepts of *waypoint* or *wayfinding*. While there is no set threshold for cognitive load, information visualizations can be tuned appropriately for the task limiting the amount of cognitive burden on the information seeker.

Visualizations are external representations, and they help to alleviate cognitive load by reducing short term memory burden (Munzner & Maguire, 2015). They have the added benefit of sometimes remaining static so that the user may return to them to reorient themselves during moments of disorientation or to support the repeated encounters needed during cognitive map formation.

2.4.1.2 Visualizing overall structure

Visualizing overall structure has been shown to be a benefit to the information seeker, particularly for exploratory search (He et al., 2019) by relieving the "browsing burden" (p. 5) caused by overloading the user's short term memory. Visualizing the overall structure of information makes improvements over list-based systems currently in place for academic literature search. Visualizations can help users construct an overview that can be helpful for complex tasks, leading to insights which drive the formulation of a high-quality query or "direct exploration of complex information needs" (White & Roth, 2009).

2.4.1.3 Enabling patterns to be recognized

Users of all types of information including both physical and digital information sources engage in looking for and establishing patterns and rely on these patterns for orientation and recognition of known and unknown areas (Dillon, 2000). For digital environments, spatial and semantic distance, the perceived difference between the meaning of two signifiers (Brooks, 1995), can be an effective way to aid the information seeker in establishing meaningful patterns that help create recognized areas that may lead to improvements in the creation of cognitive maps.

2.4.1.4 Permitting inferences of relationships through proximity

Semantic distance has long been known to be important for understanding relationships, or perceived similarity between documents, and how often terms co-occur (Brooks, 1995). Patterns are important and inferences are going to be made, so the importance of proximity may support the use of distance-based maps in which distance between visual nodes of information is a based on a meaningful value. The value of proximity has been validated in tools such as VOSViewer that utilizes distance-based maps (van Eck & Waltman, 2010) as well as in the development of the Lieden Algorithm (Traag et al., 2019) that is intended to improve upon communities, which supports enabling of patterns as well.

2.4.1.5 Improved cognitive maps and recognition of small changes

Improvement of cognitive maps is difficult to measure, as it is an internal construction. Studies from Demelo & Sedig (2021), Freksa (1999), and Dillon (2000) mentioned previously have all indicated factors that can improve cognitive map formation. Additionally, there have been previous studies using physical media, such as the creation of arts-based information maps of members of marginalized communities (Kitzie et al., 2021), the draw-and-write technique of Hartel et al. (2017), and the information world mapping methods of Greyson et al., (2020) which enable interpretation of information visually and spatially supplementing text-based analytic statements.

2.4.1.6 Parametric control to interact and control the data

It has long been established in the presentation of visual information that multiple views or an interactive means is necessary to satisfy user's needs to orient and comprehend the information provided. Overview, zoom, filter, and details-on-demand (Shneiderman, 1996) have been known to be essential to graphical user interfaces. Shneiderman also identified the technological ability

to have improved experiences with three additional tasks of relate, history, and abstract that benefit the user of such interfaces as they enable exploration even with gains in expertise.

Dillon (2000) also identifies that "knowledge-based differences", or expertise, needs to be considered and studied to create digital interfaces that offer cognitive compatibility. Dillon adds that individual differences in cognitive processes, such as memory or spatial ability, also needs to be accounted for in digital environments as these can affect the knowledge-base capabilities of users.

2.4.2 How do we apply these benefits?

Munzner and Maguire (2015) defines visualization as "allowing people to analyze data when they don't know exactly what questions they need to ask in advance", which echoes the conditions of the exploratory searcher. They identify four key components of visualization: the *interaction*, the *idiom*, *validity*, and *scalability*. The *interaction*, such as noted previously in Shneiderman (2009), permits exploration using other visual encodings. An example of this might be the color and shape of a node within a network, or the ability to zoom and filter. The *idiom* is the distinct approach and manipulation of visual elements to create an overall form, often recognizable as a visual metaphor. Examples of this are bar charts, line charts, or network maps. These require critical visual analysis. The *validity* is very difficult to measure but is usually performed with user studies in controlled experimental environments in which the visualization is part of an instrument. *Scalability* is concerned with three kinds of limitations including computational capacity, human perceptual and cognitive capacity, and display capacity. These four components should be analyzed continually throughout the visualization making process, although as noted, validity is very difficult, and these judgements are often heuristic or experiential. These components are not without precedent and build upon well established

research in data visualization of graphic density from Bertin, data-to-ink ratio from Tufte, limitations on human perception from Ware, and change blindness from Simons (Munzner & Maguire, 2015).

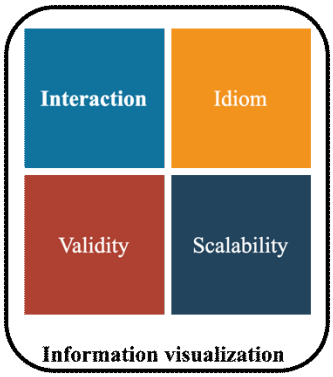

Figure 6: Munzner & Maguire's components of information visualization.

In addition to designing information visualization based on Munzner & Maguire's components, there are guidelines from graphic design practice that will likely improve engagement with an audience, directing attention to hierarchical importance, and promote understanding and memory. Kosslyn (2006) provides eight principles of effective graphic design to achieve these three goals of visual communication. To connect with your audience, the principle of relevance and the principle of appropriate knowledge should be applied to present not too much or too little information as well as using the right jargon, signs, and concepts that will be understandable. To direct the users' attention, the principles of salience, discriminality, and perceptual organization should be applied so that information is visually striking and used strategically, and differences in proportion must be evident. The principle of perceptual organization is more complex as it involves integrated versus separated dimensions to enable the reader to group things intuitively and to infer when things are similar or dissimilar. Visual communication can also promote understanding and meaning by applying three principles of compatibility, informative change,

and capacity of limitations. The principle of compatibility is providing understanding of cultural conventions, yet supporting basic assumptions, such as bigger equals more. The principle of informative change means that a reader can see change between states and capacity limitations are considered using visual elements to support short-term memory and processing limits. In other words, one shouldn't expect users to memorize or comprehend long lists of search results.

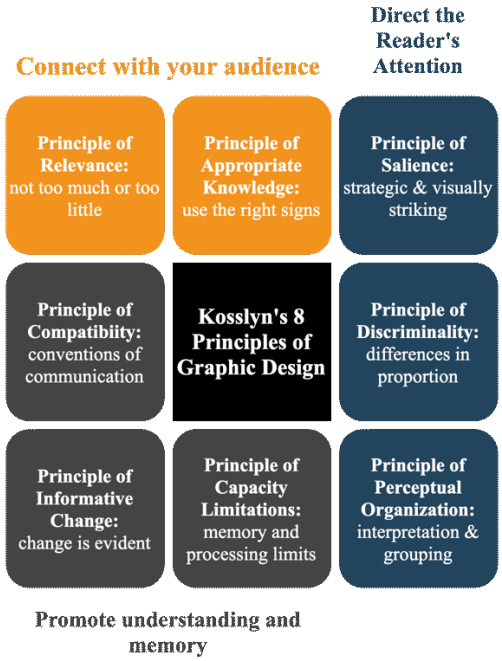

Figure 7: Kosslyn's eight principles of graphic design.

Many of these principles can be recognized in Munzner & Maguire's components and the six ways visualizations can help from Card et al. (1999) and Börner (2010). With these as evaluative frameworks, producing a visualization that fits within these boundaries is the goal.

To reduce the cognitive and emotional burden during the exploratory search process, I intend to utilize the benefits of visualization. I will create a visualization that should help expand working memory, provide an overall view of the exploratory search enabling patterns to be recognized,

enable inferences and improved cognitive processes, and provide some ability for exploration regardless of expertise.

2.5 Reviews of bibliometric approaches

In this section, I'll show previous examples of attempts to address exploratory search. As there are numerous bibliometric methods that can be used, I have narrowed down to two methods. I'll show how keyword co-occurrence methods and bibliographic coupling have been used previously to visualize search results for the intent of improving exploratory search and reduce the gap between novice and expert subject knowledge.

2.5.1 Keyword co-occurrence analysis for information search

Co-occurrence analysis, particularly visualized as a network, is widely used in bibliometrics for all aspects of investigating the intellectual structure of fields (Ding et al., 2001; Leydesdorff & Nerghes, 2017; Rip & Courtial, 1984) and works with a limited number of terms with a matrix representation of co-occurrence which can be statistically modelled. "Co-occurrence matrices … provide us with useful data for mapping and understanding the structures in the underlying document sets" (Leydesdorff & Vaughan, 2006, p. 1616). Co-occurrence matrices may be created from citations, words, authors, etc., and are commonly used in bibliometric and its related subfields. In this study, we are utilizing author assigned keywords to create the co-occurrence matrices. Keyword co-occurrences used in term maps may include any type of word and its combinations as phrases and is considered by González-Valiente et al. (2021) as "more optimal since it covers new topics, [and] includes the views of scientists" (p. 338).

Keyword co-occurrence is similar to co-word analysis, which was introduced by Callon, Courtial, Turner and Bauin (1983) and identifies co-occurrence of pairs of words derived from a

corpus and usually visualized as a distance-based map of nodes and edges with the strength of the link determined by the number of co-occurrences. The reading of this network is in identifying the clusters and their relations and not the individual words (González-Valiente et al., 2021).

Brooks (1995) examined textual factors influencing user's perception of bibliographic resources with semantic distance (relatedness of concepts in meaning) and term overlap (co-occurrences of the same term or phrase across a record or records) being two of the three important factors affecting judgement. While it might be considered for future work to include reference to or working within formal taxonomies, as Brooks did, semantic distance and term overlap are of most importance in this context, with an informal taxonomy being a construction developed by the exploratory seeker (Kuhlthau, 2004).

Keyword co-occurrence is easy to recognize from a user's perspective as it not only mimics search terms, yet also provides new terms for *discovery* and *learning*. Exposure to new words and phrases permits an exploratory searcher to move through learning and onto *investigation*, leading to reducing uncertainty. As one gain's confidence, the exploratory searcher tests new words or phrases that represent topics to support *analytical strategies*, or acts as waypoints during the *formulation/query(re)formulation* process of an end destination.

Waypoints require recurrence; the information seeker depends on their stability and needs to be able to return to them for orientation of themselves in these constructed cognitive models of digital spaces (Freksa, 1999). Keywords enable this stability as there are often recurrences across documents and their semantic meaning has some stability over time due to historical accumulation of documents.

2.5.1.1 Keyword clustering

Once co-occurrence has been identified, it is desirable to know if there are groups with similarities. In contrast with faceted approaches, where the goal is to narrow down along specified pathways within a formal structure, clustering provides the user with an overview of the structure and provides insights into possibly unknown communities calculated by some similarity measure (White & Roth, 2009).

Keywords used in co-occurrence models serve as a type of topic model and are free of the restrictions of a formal or predefined taxonomy (di Sciascio et al., 2016). "The model generation process does not infer any semantic information, instead it discovers patterns basing on term co-occurrence" (di Sciascio et al., 2016, p. 120). How these groupings occur depend on how clustering is defined.

While there are many community or similarity methods available, such as Latent Dirichlet Allocation, Ward Hierarchical clustering and the Louvain Community Detection algorithm, the Leiden algorithm (Traag et al., 2019) is appealing for its robustness and implementation in Python as well as inclusion in Gephi. The authors of the algorithm identified errors in detached communities common with the Louvain algorithm which may "yield communities that may be arbitrarily badly connected" (Traag et al., 2019, p. 1). To address this possibility, the Leiden algorithm "yields communities that are guaranteed to be connected" (Traag et al., 2019, p. 1). Utilizing this algorithm may help provide an improved visual accuracy for identifying clusters within the overall structures.

I acknowledge the limitation of individual keywords removed from their source in that the assumption that keywords can stand independent of their context is unrealistic (van Rijsbergen,

1980 as cited in Spink & Cole, 2006). Their semantic meaning and interpretation effects relevancy by the context of the information seeker. As an overall view, clustered with others due to co-occurrence, this may provide some level of topical inference.

2.5.1.2 Examples of past co-occurrence studies

Term co-occurrence analysis for visualizing dynamic queries has also been explored as an interface for digital libraries (Buzydlowski et al., 2002, p. 133) and it has the benefits of being an improvement over list form. Two of the visualization idioms (SOM or self-organizing maps and PFNET or Pathfinder networks) align well with the manually created mental maps of subject-area experts.

Visualizing keyword chaining for relevancy using an add-on for an existing search system, Athukorala et al. (2017) identified that using the traces, or recorded history, of keywords from webpages visited supports topic identification when analyzed as a historical trail. Liu et al. (2009) examined a method of comparing two search results and developed an algorithm to exploit XML data for differentiation between multiple web search results. However, the purpose of the study was to create an algorithm capable of differentiating between search results and did not result in a visualization, though this may be valuable for future research to build upon.

As a response to most search tools returning only a list of documents which prevent the user from understanding if "important papers have been missed or even whole subfields", Bascur, van Eck, and Waltman (2019) utilized a scatter/gather technique along with a unique algorithm for packed bubble chart to display the "structure of the search results" (p. 76). Its goals are similar to this study's goals and intended to be "understandable even to a user who has only a limited familiarity to a domain of interest" (Bascur et al., 2019, p. 76). While the outcomes of a study are

a proposal featuring a sequence of algorithms, it does align with this study's intent to provide a broad understanding and not for retrieval of a specific paper.

Coats (2020) created an open-source application, the Zipf Explorer Tool, that still functions as a web application for the comparison of keywords extracted from texts for the purpose of illustrating lexical diversity utilized for textual and discourse analysis. However, it does not provide search capabilities and provides a data visualization approach using graphs exhibiting Zipf's law. It also has not removed stop words, making exploring words that are important within the document time consuming to locate. Built using Python and open source, it could be valuable to build up on for future work.

IntentRadar, a feature of SciNet (Kangasrääsiö et al., 2014), is similarly aligned with this project's goals of aiding the novice information seeker during exploratory search. It interactively adjusts search results based on the user's refinement of keywords for relevance and arranges data visually using a radar-graph layout, where the center represents highest relevance, and the outside is new or future items for consideration. User research found that this system offers an improvement for exploratory search tasks when compared with traditional interfaces. However, it does not offer the ability to compare independent searches for similarity and is more of a drill-down or facet-based approach.

Hoeber & Shukla (2022) developed an interface for digital library search which utilized visualization of linked keywords as the main component. The visual linking of keywords from a search maintains the link with the source documents, making this a unique keyword visualization approach not seen in the other keyword visualizations. This provides a tool that can support both discovery and learning as well as re-finding during exploratory search (Capra et al., 2010). This

development is very exciting but is currently unavailable as its intended to be installed as a lightweight module within an academic digital library search interface using a “10 blue links style search engine results page” (Hoeber & Shukla, 2022, p. 5). It also lacks the ability to compare two search results simultaneously, a primary objective of this study.

VOSViewer (van Eck & Waltman, 2010) is well known as a bibliometric tool for interpreting, analyzing, and visualizing bibliometric maps. It not only can perform textual analysis such as keyword co-occurrence, but also bibliographic coupling. It includes community detection using Louvain and Leiden algorithms among others and optimizes the visualization for best fit and readability. It has an easy learning curve and is freely available as open-source software. For the academic librarian, it is an indispensable tool not only for creating visualizations but as a performance standard by which to measure your own development. It processes terms for co-occurrence, not based on author assigned keywords, but from noun-phrases from part-of-speech tagging using a natural language processing library (van Eck & Waltman, 2011). Since it applies a filter to only use adjective-noun and noun phrases determined for relevance by their own technique, this then excludes phrases that include verbs or adverbs. It does not enable comparison of two search results datasets unless the user is aware of overlay mapping techniques within VOSViewer and the data processing required (González-Valiente et al., 2021). González-Valiente et al. (2021) provide results of comparing two search results to understand the overlap between two queries of multiple search terms from WoS. The data was prepared by automatic extraction of noun-phrases using VOSViewer, but also further manipulated by creation of a thesaurus in which place names, months, institutions, and “generalist phrases” were excluded. While the intent of their study is to understand the terminological similarities between two specific terms, in this case Information Management and Knowledge Management, it is different

from this study's objective to provide a new means of comparing author assigned keyword overlap and exclusion.

2.5.1.3 Conclusion to keyword co-occurrence analysis

Keyword co-occurrence analysis was chosen for its ease of linking keyword cluster to recognizable topics and the ease and speed of processing and visualizing data of varying size. I have identified that there are no applications or means of creating visualizations that permit one to visually compare the co-occurring author assigned keywords from two independent search results, effectively giving the viewer an overview of excluded and overlapping keywords found in the documents. As groupings of keywords can be used to identify topics, using keyword co-occurrence analysis as the basis for a visualization is appropriate for the purpose of providing an overview of topics to aid in a structural understanding and promote construction of an exploratory searcher's cognitive maps. But keyword co-occurrence isn't the only method that has been used to visualize the topical structure of a collection of documents. In the next section, I'll look at how bibliographic coupling has been used.

2.5.2 Bibliographic coupling

Originally used by Kessler (1963), bibliographic coupling is a citation-based method that is the "sharing of one or more references by two documents" (Small, 1973). As the number of common references increases, so does its strength. It has been used to determine the degree of topic similarity between two papers. Unlike co-citation analysis, which is a dynamic measure, it is a static measure. In other words, the two authors can't go back to add or subtract references to their paper, changing the strength. It has been identified by Klavans & Boyack (2017) as superior to other means of citation-based measures, such as direct citation and co-citation and was preferred for its accuracy of discipline partitions over other methods, except the bibliographic

coupling-based citation-text hybrid approach (Klavans & Boyack, 2017). Compared with textual similarity approaches, it has been shown that "document-based taxonomies provide a more accurate representation of disciplines than do journal-based taxonomies" (Klavans & Boyack, 2017, p. 2). Direct citation, "where articles are linked if one references another, only considers links from within the set" (Klavans & Boyack, 2017, p. 20) and is only recommended when very long-time windows are used, which may be of concern for searches with time limited ranges. Bibliographic coupling has the advantage over the other two as being able to cluster the newer papers at the cost of the older papers in contrast with co-citation analysis. Based on limitations and capabilities of each process – direct, co-citation analysis, and bibliographic coupling – bibliographic coupling presents benefits for studies where there is a limited timeframe and papers tend to be newer than older.

2.5.2.1 Prior studies using bibliographic coupling for visualization

The recently discussed VOSViewer can provide also provide distance-based maps of bibliographic coupling. Visualizing communities based on the number of common references between two documents, it provides a document level view of clusters of highly related articles. But, as with the previous limitation for keyword co-occurrence, visually encoding the nodes representing those documents with their search source was not possible at this time.

There are many web-based, or downloadable applications for exploring information visually using a bibliometric or citation-based approach, such as Connected Papers (*Connected Papers*, n.d.), Paperscape (*Paperscape*, n.d.) Scite (*Scite*, n.d.), Citation Gecko (*Citation Gecko*, n.d.), and CitNetExplorer (*CitNetExplorer*, n.d.). However, the existing information visualization web applications do not permit the comparison of one search result to another, but they currently work and are available to exploratory searchers.

Litmaps (*Litmaps*, n.d.) stands out from the previously mentioned applications as it does have the capability to compare two search results. Using seed papers to create a network map, you can compare two sets against each other. While this is still focused at the document level, it may provide the exploratory searcher with a structural sense of a topic area. This might be a valuable place to explore and validate in future work, to see if the comparison of two searches in Litmaps provides the user with an improved cognitive map.

He et al. (2019) developed PaperPoles, a visual analytics system that assists searchers for relevancy based on citation links of papers "that are known to be relevant". It provides a search interface and permits exploration of the results by topic and relevance. The downside to this approach is that it requires two seed papers to start to effectively guide exploratory search (He et al., 2019), which may not be of benefit for the novice, particularly if they have not formulated a goal.

Radial Sets (Alsallakh et al., 2013) comes closest to the intent of this project with its focus on showing different types of overlap between two sets of data for the purpose of pattern identification in data. Although no longer existing, its use of a radial layout, and frequency-based representations is relevant for validity of this study's visual *idiom*. It may serve as a validation in future research as it addresses some issues of scalability by combining visual idioms, such as bubble packing and radial charts.

2.5.3 Limitations to visualizations

Of course, limitations should be recognized for these maps – they are constructions built with tools oriented from particular and contextual perspectives, burdened with their history, situatedness within English-speaking Western societies, just as much as they are bounded by

their technological complexity (Day, 2014). De Bellis (2009 p. 142) comments that "the design and construction of maps has been deconstructed by historians…[the maps] betrays its social context of production and its latent identity of technology over power". Building network maps, while intended for the user to explore the structure or infer insights about spatial relationships, may include assumptions about relationships. To provide an example from bibliographic coupling analysis, authors create perceived or self-constructed dependencies on previous work and the bibliographic coupling not only reveals those dependencies, but also conceals all the other influences behind this simple reduction (De Bellis, 2009). Additionally, information seekers, may also not be visually literate when it comes to displaying information graphically, and may find difficulties with mappings or graphs compared with intermediate or expert users (Wu & Vakkari, 2018a). The graphic design principles (Kosslyn, 2006) and the key components of visualization (Munzner & Maguire, 2015) are an attempt to provide a guidance framework to mitigate some of these issues during development.

Many of these studies and applications are tantalizingly close to meeting the objectives of this study, but individually they fall short of not being able to compare the results from two search terms, or they offer a means of comparison, but use citation-based measures to do so, such as with Litmaps. The terminological similarities study by González-Valiente et al. (2021) could have been used as a procedural basis to achieve this study's research objectives, but, from my perspective, it has the barrier of not being able to utilize the author's defined keywords. Collectively, they represent a vast breadth of methods for answering my research objectives, but do not address the need to use write an open-source code that would be interpretable and usable by an academic librarian with some familiarity with Python.

2.6 Conclusion

This chapter contains four sections in which I provided an overview of exploratory search, information seeking models and uncertainty, information visualization, and appropriate methods identified. In Section 1, I explained exploratory search and highlighed how White & Roth's (2009) definition of the process acknowledges uncertainty as presenting challenges to the information seeker. Section 2 showed how exploratory search fits within information seeking behavior models, such as Hoeber & Shukla's modified White & Roth model (2022), with emphasis on Kuhlthau's ISP model (Kuhlthau, 1991, 1999a, 1999b, 2004; Kuhlthau et al., 2012) for the purpose of understanding how uncertainty affects emotions that contribute negatively to the experience. I further explained how uncertainty can be mitigated through facilitating the construction of waypoints (Freksa, 1999) and cognitive map formation (Demelo & Sedig, 2021) through the use of visualizations. Segueing into Section 3, I examined how information visualization can help mitigate uncertainty as a static image in this current study, but also when in a development framework for future work as a dynamic interface. The benefits of information visualization were explained in six points with five relevant for static visualizations: relieving cognitive burden, visualizing overall structure, pattern recognition, inferring proximity relationships, and improved cognitive maps (Börner, 2010; Card et al., 1999). I also reviewed an evaluative framework from Munzner & Maguire (2015) that is useful in the development of visualization. I then introduced Kosslyn's (2006) eight principles of graphic design that are useful for creating engaging and effective visual communication. These may be useful in the future for more thorough analysis comparing the effectiveness of visualization idioms. The fourth section covered the history and usage of two methods, keyword co-occurrence analysis including an appropriate clustering algorithm, and bibliographic coupling. As the Leiden algorithm must have something to work on, it is applied to keyword co-occurrence as an

outcome, though these are complementary, but separate methods. This section also examined other studies and current applications using keyword-based or citation-based methods, where visualization is the main interaction component for the exploratory searcher. I concluded this chapter acknowledging the limitations of visualizations as a mediation and identifying how the prior studies and applications would not have allowed me to meet my research objectives in a way that might be deployable by an academic librarian.

To summarize, the assumptions are many: that information seekers experience uncertainty and emotions during the exploratory search process that can impair or threaten success; that uncertainty is well studied, but a normal part of information seeking but can be mediated; that uncertainty can be mediated by helping information seekers construct cognitive maps that give them an overall view that helps them towards formulation; that visualization of the overall structure of information is an appropriate mediation; that information visualization has been used before and that two methods, keyword co-occurrence with clustering and bibliographic coupling can be used for visualizing the structure of topics to compare two search results. These assumptions underlie the conclusion that a unique information visualization is needed to compare the author defined keywords from two independent search results so that an information seeker may understand how they are related by exclusions and overlapping terms. In Chapter 3, I will discuss the methods of creating this unique information visualization.

CHAPTER 3 METHODS

3.1 Introduction

3.1.1 General methodological approach and model evaluation

This chapter will explain the two methods being used to generate outcomes to answer the two research objectives. The first research objective, (RO1), to present an effective way of visualizing the effects of search term choice so that included or excluded keywords might be compared between the two search results, will be achieved by utilizing keyword co-occurrence with Leiden algorithm clustering driving the priorities for the layout algorithm. It was anticipated this method will produce a visualization of the overall relationship between two search terms with regards to keywords found from their respective search results. The effectiveness of the visualization will be analyzed heuristically by comparing the outcomes against benefits of information visualization (Börner, 2010; Card et al., 1999), components of information visualization (Munzner & Maguire, 2015), and relevant graphic design principles (Kosslyn, 2006). The second research objective (RO2) to compare the keyword-level analysis with a document-level analysis using a proven bibliometric method to determine if topical exclusivity exists between search terms, will be answered using bibliographic coupling in a network map with results coded by which search result they originated. RO1 will be explored by creating a keyword map using keyword co-occurrence supported with keyword clustering maps, and RO2 will be investigated by creating a publication map using bibliographic coupling and comparing it with the keyword co-occurrence maps. The keyword co-occurrence analysis was combined with clustering and layout algorithms to present a visualization of exclusive and shared terms in each search result. Keywords are extracted from the Web of Science (WoS) datasets and minimally manipulated in Python to preserve the author intent.

Keyword co-occurrence and clusters from the Leiden algorithm were used to address RO1 with various visualizations in Gephi explored before a suitable idiom was found that enabled clear communication of the shared and exclusive keywords of two independent searches in WoS. The Leiden algorithm compared the keywords within the overall body of keywords at the document level and found communities based on occurrence that suggest similarity. It is expected this will create topically wider communities based on the co-occurrence with relatively low term frequency counts, when compared with other applications of the Leiden algorithm on keyword-based clusters from full texts.

Bibliographic coupling was used to identify topical clusters at the document level and identify what proportion of documents may or may not be excluded from a search result. Even though each search may have excluded keywords that are clustered together, does this also apply to documents? Or are the documents immune from as high of a rate of exclusion due to their use of five or more keywords, thus mitigating the effect? Bibliographic coupling is also performed as a comparison for the topics/communities found in the Keyword Map using a known bibliometric method typically used for topical clustering as a means of answering RO2. Although direct citation has been shown to be more accurate, bibliographic coupling has been utilized due to its lack of needs for minimum reference counts (Klavans & Boyack, 2017) and use for revealing the relative subject areas of authors.

Both methods were chosen for two reasons: their simplicity and ease of use in a Python environment in which I can understand and control what is happening to the data and the outcomes. While there are more complicated methods, particularly evident in the BIR literature using machine learning and trained data sets, or keyword extraction from full text scanning, they are also very resource intensive, having larger implications in their implementation, *validity,*

scalability, and deployment. As the goal of this study and its ongoing trajectory is to address the needs of exploratory search, computationally heavy applications such as the current application Open Knowledge Maps (*Open Knowledge Maps*, n.d.), may not appeal due to the considerable lag and concerns for overinvesting in exploratory search directions that may be abandoned (White & Roth, 2009). An additional reason for creating my own coded solution instead of using an existing application, such as the process outlined by González-Valiente et al. (2021) in VOSViewer, is the use of Author keywords in the WoS full record. These provide better representation of the document's content compared with the Keywords Plus data from WoS (Zhang et al., 2016). Keywords Plus, algorithmically generated from the tittles of the article's references, can provide a broader descriptive view which may be a good addition for use with mapping knowledge structure (Zhang et al., 2016). As I wanted to preserve authorial intent, I chose to use only the author keywords.

This study used existing methods of visualization layout algorithms found in Gephi (Bastian et al., 2009), an open-source software for network visualization. Gephi focuses on network graphs with many options for layouts/visual idioms, and *scalability* is always an issue in visualization. There are three kinds of limitations, *computations capacity, human perceptual cognitive capacity*, and *display capacity* (Munzner & Maguire, 2015). Munzner & Maguire cite significant research identifying limitations for graphic density (Bertin, 1967) data-to-ink ratio (Tufte 1983), human perception (Ware, 2013) and change blindness (Simons, 2000). As such, visualizations created in Gephi were kept as simple as possible with consideration of the above points.

To decide which network mapping layout was best suited for this task, I utilized, heuristically, Munzner & Maguire's 3-part What-Why-How analysis framework for developing a visualization: what equates to the data, why equates to the task by the user, and how equates to

the idiom (Munzner & Maguire, 2015) as a means of identifying what type of visual idiom would be appropriate. In this study, the data is keywords, the purpose is to see the connections where they are found in documents together and to represent the quantity of those connections, and the idiom needs to show a range between two extreme values. As I'll referenced in Chapter 2, previous studies have used visual idioms of spider, radial, and linear graphs as well as network maps. A network map as an idiom makes the most sense as it is perfect for showing relationships between elements representing nodes. As my own mapping skills were dependent upon Gephi's network mapping capabilities, selection of visual idioms was limited to the layout algorithms within Gephi. I found that the Dual Circle layout provided a replicable representation and would support the limited range of values for positioning of the nodes.

With information from past studies informing my choices, my methods of analysis chosen to meet the research objectives, and a method in Gephi to reliably create a network map that is appropriate for the data and meets the needs of the first research objective, I moved forward with the following workflow as shown in Figure 8.

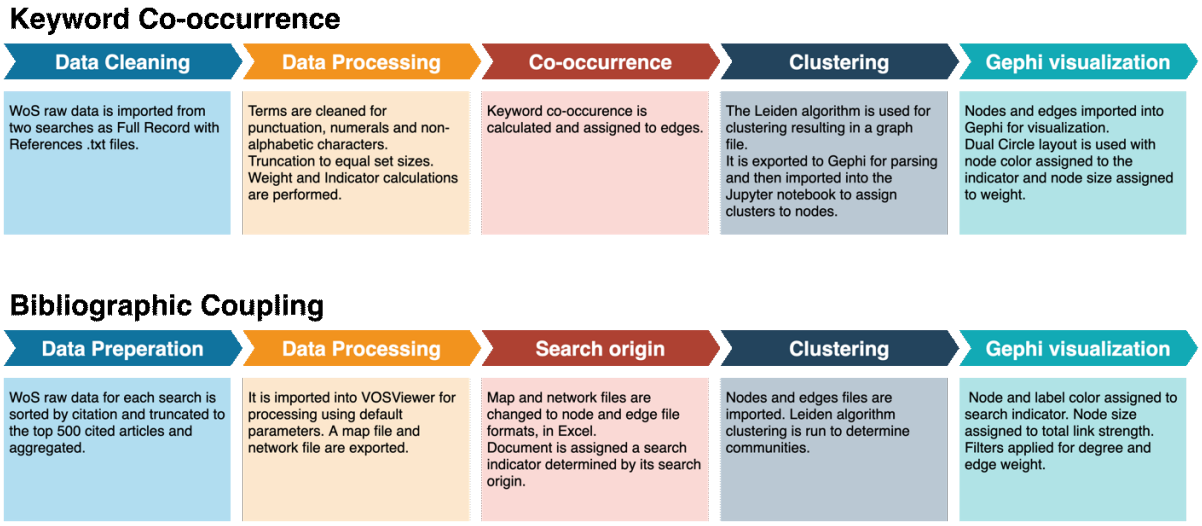

Figure 8: Workflow for answering research objectives.

3.2 The data and the treatment of the data

3.2.1 Data sources

The data was limited to a commercial academic database of peer-reviewed articles, though similar analysis or need for knowledge could encompass other means of scholarly communication including conference proceedings and presentations, reviews, letters, news, blogs, textbooks, whitepapers, reports, visual and audio resources, or other data sources.

As the data source is peer-reviewed articles from a commercial database, there are two sources of keywords and this study will only use the author defined keywords, provided by the author(s) when the manuscript is submitted for review to a journal. WoS has keywords created computationally, but these were excluded from the study. I also recognize there are other methods to gather important terms from a document's abstract or main text using TF-IDF. This was considered but kept outside the scope of this study to focus on the task of creating an effective visualization.

This study used data from the Web of Science (WoS) with search terms TOPIC='social justice' and TOPIC='social injustice', limited to 2010-2020 and only retrieved peer-reviewed articles. Downloads were Full Record with Citation References in .txt format. Acknowledgements of WoS's shortcoming, particularly for humanities and social sciences, have been recognized (Mongeon & Paul-Hus, 2016), but this database was chosen based on its ease of use and cleanliness of data in contrast with databases such as Google Scholar which were not designed for bibliometric research (Mongeon & Paul-Hus, 2016; Orduna-Malea et al., 2017).

Table 1. Summary of documents downloaded from Web of Science on January 20, 2022.

	Social Justice search	Social injustice search	Total
Articles returned from initial search	8,164	430	8,594
Articles with author keywords	6,736	359	7.095
Total number of keywords	36,809	986	37,795

3.2.2 Data processing

3.2.2.1 Keyword co-occurrence data processing

Data processing was performed in a Jupyter Notebook in Python using the Pandas and string libraries for all cleaning and frequencies. Documents with no author keywords were dropped (Table 1). Keyword phrases were maintained in the cleaning process with punctuation and non-alphabetic characters removed. Keywords words were not lemmatized to maintain the semantic quality of author-assigned keywords and to improve readability in detail-on-demand views.

Two measures for nodes were created: *weight*, which is the value counts for all terms of both searches creating a list of unique terms with normalized frequency, and *indicator*, which is the sum of percentages of a keyword from one search appearing in the other search. Prior to calculating weight, the dataset was limited to the top 500 keywords of each search by normalized term frequency to address of differing corpus size as the "social injustice" set was 1/18 the size of the "social justice" corpus of keywords. Both were concatenated into a new dataframe from which *weight* was calculated by adding the two values. *Indicator* is calculated by dividing each keyword normalized frequency by the *weight*. The results for one, in this case the "social injustice" keywords were inverted by multiplying by -1. So, all "social injustice" keywords have a negative *indicator* value from -1 to 0 and all "social justice" terms have a positive *indicator*

value of 0 to +1. Keywords that are only found in one set or the other would only have a value of -1 or +1. This data process results in a unique list of keywords with both a *weight* (normalized frequency of aggregated keywords) and *indicator* (a value that represents how shared a keyword is between the two sets.)

First, the percentage of search 1 of search 2 was found by the following equation:

$$P_{search\ 1\ in\ search\ 2} = \left(\frac{kf_{search\ 1}}{(kf_{search\ 1} + kf_{search\ 2})}\right)$$

Second, the percentage of search 2 in search 1 was found by the following:

$$P_{search\ 2\ in\ search\ 1} = \left(\frac{kf_{search\ 2}}{(kf_{search\ 1} + kf_{search\ 2})}\right) \times -1$$

To create the *indicator* value, both percentage were added. The result is a value to suggest how often a keyword appears in two searches.

$$Indicator = P_{search\ 1\ in\ search\ 2} + P_{search\ 2\ in\ search\ 1}$$

The cleaned nodes file has 851 unique keywords from 500 of the most frequent keywords from each search, meaning there were 149 words duplicated between the two searches for this dataset.

The edges file was created by creating a co-occurrence matrix from the nodes file and the indexed list of keywords of each document from the entire corpus. For each keyword, a count is registered for every document the keyword is found in the author keywords (column 'DE' in the WoS data). For example, the keyword "social injustice" was found in three documents (so has an edge weight of 3) with the keyword "discrimination", but was was found in only 1 document with the keyword "feminism" resulting in an edge weight of 1 between "social injustice" and

"feminism". The edges file contained 19,264 rows and includes the source, target, and weight columns to be imported in Gephi for visualization.

3.2.2.2 Keyword Clusters

The Leiden algorithm from the leidenalg Python library (Traag, 2018; Traag et al., 2019) was applied to the co-occurrences using the Modularity Vertex Partition with its default values. As the leidenalg library depends upon igraph (Nepusz et al., 2003), another Python library, a graph file was exported. Gephi was used to parse the graph file to extract the cluster information, which was imported back into the Python notebook and assigned to the nodes file. The nodes file exported contains the id, keyword, weight, indicator, and cluster.

3.2.2.3 Bibliographic Coupling data processing

Data processing for RQ2 for bibliographic coupling was performed by importing the WoS datasets directly into Excel. Both sets were sorted by total citation count, (Z9), and the top 500 "social justice" and all of the "social injustice" (n=431) were selected. The two selections were aggregated into one file and exported. The aggregated data was imported into VOSViewer (van Eck & Waltman, 2010) and analyzing the data at the document level using default settings. The network and map files were exported and opened in Excel. The DOI was used to search the original WoS export files to determine the search origin of each record, resulting in a new column for 'search' which includes either a 1 or a -1. This was saved as a nodes file in .csv format. The network file from VOSViewer was also imported into Excel, a header inserted, and saved as an edges file in .csv format to be read into Gephi for visual analysis.

3.3 Data analysis & visualization

3.3.1 Keyword co-occurrence analysis

Nodes and edges files from the Jupyter notebook were imported into Gephi. Node color was assigned to the indicator value, (ranging from -1 to +1) and node size by weight (normalized frequency). Edges were left as directed and weighted by the edge weight (included in the edges file from the Jupyter notebook). The color scale below also represents results distributed across the entire node count.

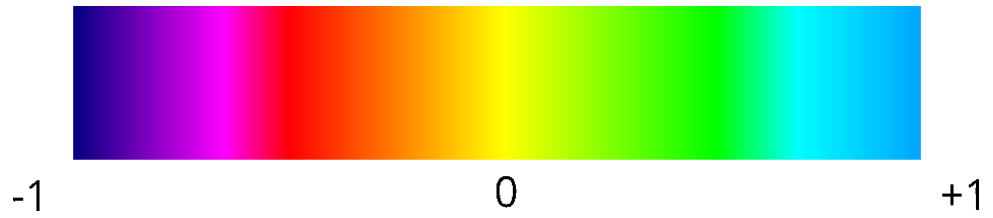

Figure 9. Color map for nodes with cyan representing the +1.0 indicator value for “social justice” and dark blue representing the -1.0 indicator value for “social injustice”. Nodes with color in between the two extremes indicate a co-occurrence in both search results.

I utilized the Dual Circle layout algorithm with nodes sorted by the indicator attribute as first priority and then by cluster as second priority. Layout was adjusted using compression and label adjust functions for readability. An edge degree filter was applied to reduce edges to just those over 2, to improve readability. Node labels were colored similarly to the parent node with label size attributed to weight. A PNG file was exported and annotated in a vector-based graphics program to add in map labels, as shown in the Results section below as the Keyword Map, and the tables were created in Excel using the nodes and edges files.

3.3.2 Keyword clustering

Keyword clusters were identified visually using Gephi. While the Leiden algorithm found 22 communities when run in the Jupyter notebook, many of these from cluster 9 to 22 having very

few keywords. Clusters 9 through 22 were aggregated into one cluster to create a cluster size comparable to the others. As the contents of this aggregated cluster number 10 were all exclusive to one search result, ("social injustice"), and were mapped together spatially, the decision to reduce these down to one cluster seemed logical. To create a specific keyword clustering visualization, the same Dual Circle layout was maintained from the previous step, but node color was assigned to cluster number from the nodes file. A PNG image file was exported and annotated in the same manner as the keyword map to create the keyword clusters map. Bar graphs and tables of keyword clusters were created in Excel and imported into the final report.

3.3.3 Bibliographic coupling visualizations

Both prepared nodes and edges files were imported into Gephi for visualization. Node color was assigned to the search indicator (a binary value of 1 or -1), and node size assigned to weight. Node label size was assigned to weight and label color aligned with search indicator. Force Atlas 2 was used as a layout algorithm, though other algorithms showed similar results. Two visualizations were created with filters applied. The publication map (Figure 16) contains no filters and visualizes the aggregated data of the top cited papers of each search. The Giant components view of the publication map applies filters for degree and edge weight so that the giant component of the highly related papers can be seen. PNG files were exported from Gephi and annotated in a vector-based graphics program for use in this report.

3.4 Conclusion

In this section, I have described the methods used to answer both research questions using two analyses. To answer primary research objective (RO1) I used keyword co-occurrence analysis and measures of weight, (normalized keyword frequency in the corpus) and indicator, (a value of -1 to +1 of how much a term is found in one search or the other, with exclusive terms having the

extreme values of -1 or +1) to create a keyword map and keyword clustering results. The second research objective (RO2) attempts to validate the topics revealed between each search term by using bibliographic coupling to create a publication map, which reveals intellectual structure by finding common references between documents, and thus topics that authors of both documents are addressing. The intent with comparing these two methods is to show the potential of a new way of illustrating search term effects by grounding it with a well-established bibliometric tool. The keyword map and publication map, along with their supporting views will be analyzed heuristically by comparing the outcomes against benefits of information visualization (Börner, 2010; Card et al., 1999), components of information visualization (Munzner & Maguire, 2015), and relevant graphic design principles (Kosslyn, 2006) discussed in Chapter 2. In the next chapter, I will explain the visualization outputs of the keyword map, its supporting keyword cluster maps and tables, and the publication map.

CHAPTER 4 RESULTS

4.1 Keyword map

In this section, I will cover results in three sections: the keyword map, keyword clusters, and the publication map. The keyword map shows the comparison of keywords from two different search results, enabling identification of the proportion of exclusive terms to each. The keyword clusters section examines the clusters identified using the Leiden algorithm, which may reveal topics due to their co-occurrence. Lastly, the publication map shows how each search is represented at the document level using bibliographic coupling and if there are topical exclusion areas.

4.1.1 Structural view

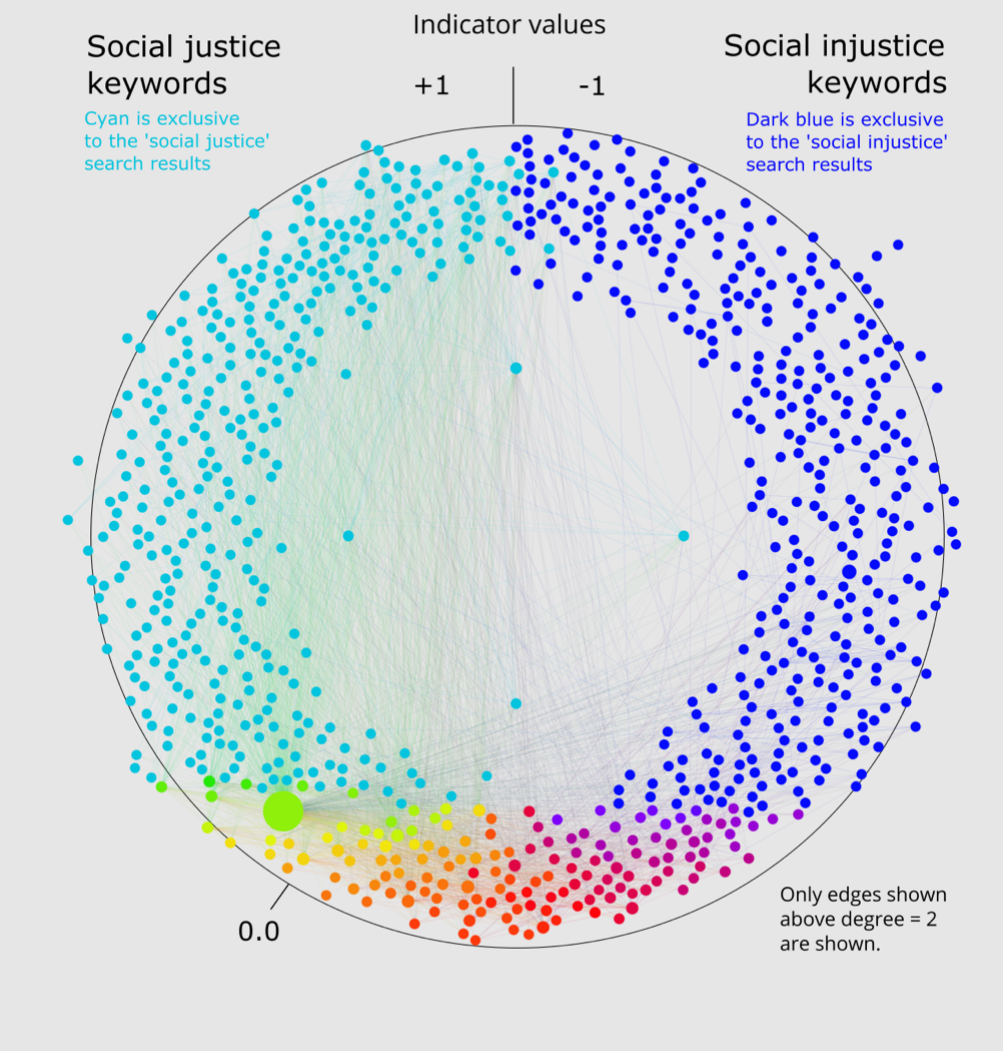

Figure 10. Keyword map showing the co-occurrence of keywords. Each node represents an author keyword/phrase and size representing its frequency within its respective search.

Figure 10 illustrates the intent behind this study – to show the differences between search terms by indicating keywords common between the two, but also exclusive to each other. The figure shows keyword distribution by the indicator value of the top 1000 most frequent keywords from documents of each search with node color designating exclusion or co-occurrence. Cyan represents one end of the spectrum as the "social justice" search (*indicator value* = +1), and dark blue represents the other end of the spectrum as the "social injustice" search (*indicator value* = -1), both starting at the top and then progressively co-occurring keywords area arranged from

each limit. The midpoint in the values is defined by the '0.0' which was estimated by sampling the nodes in Gephi. Edges represent the co-occurrences between keywords and node size represents its weight or normalized frequency across the aggregate data. Nodes that are in cyan or magenta are exclusive and do not occur in the search results of the other.

Even though a spectrum has been used, the transition from the extreme values, such as +1, or -1, is abrupt. For example, all the cyan nodes in "social justice" have a value of +1 and all the dark blue nodes in "social injustice" have an indicator value of -1.

As the visualization shows the top most 500 keywords from each search, this is a small proportion of the "social justice" keyword set compared with the "social injustice" set. There are considerably more non-linked keywords on the "social injustice" side, which means that these words do not have more than one connection, since there is a filter applied to reduce the edges to those below 2. An edge weight filter was applied to reduce visual clutter.

Table 2 shows the top 10 terms by weight for keywords exclusive to the two searches and the most common terms equally shared between both search result sets. This data was extracted from the Python notebook for keyword co-occurrence analysis. As there were no keywords with an indicator of 0.00 in the truncated dataset, a range of +/-0.02 was utilized to select these 10 from the 1000 keywords in the analysis.

Table 2. The top 10 by weight (normalized count for each respective search) for the extreme ends of the spectrum for “social justice”, “social injustice”, and those that equally occur between both search results.

Exclusively “social justice” keywords	**Equally shared Keywords**	**Exclusively “social injustice” keywords**
ethics	displacement	social injustice
critical pedagogy	society	consumption
advocacy	gentrification	poetry
social movements	human trafficking	injustice
empowerment	right to the city	global education
equality	diversity	kant
power	politics	nigeria
qualitative research	teacher education	social sustainability
multiculturalism	activism	communication for solidarity
sustainable development	higher education	social protest

4.1.2 Detailed view of keyword map with node labels

Figure 11 is a detail view of the top portion of Figure 10 with added node labels to identify exclusive keywords. Nodes and edges have been removed for clarity. Here the effects of search term choice reveal what is missed by one search or the other. Keywords from Table 2 can be found in Figure 11, for example, ‘multicuturalism’ can be found in the list of exclusive keywords and in the keywords shown in cyan along the top half of the border. It is important to keep in mind that this is a truncated list of the most frequently occurring keywords (n=851) and of the excluded terms and that many more may be seen if the visualization contained all 37,795 keywords.

Figure 11. Detail or zoomed-in view of the keyword map with node labels. Cyan represents exclusive keywords to the "social justice" results, and dark blue represents exclusive keywords from the "social injustice" set.

Of the exclusive keywords, 41% affect what the information seeker would miss in choosing one term over the other for this limited dataset based on 500 keywords from each search. This equates to 819 for "social injustice" and 15,182 for "social justice" if the same proportion is applied to the entire search results dataset.

4.2 Keyword clusters

4.2.1 Distribution of keywords across clusters?

Examining the "social justice" exclusive terms suggests they are spread across the clusters and do not seem isolated, though clusters do have aspects of exclusivity. Figure 12 shows that the

exclusive “social justice” keywords are distributed across clusters and not confined to just one cluster, though they clearly dominate clusters 1 and 2. Similarly, “social injustice” dominates clusters, 9, 8, 6, and 4.

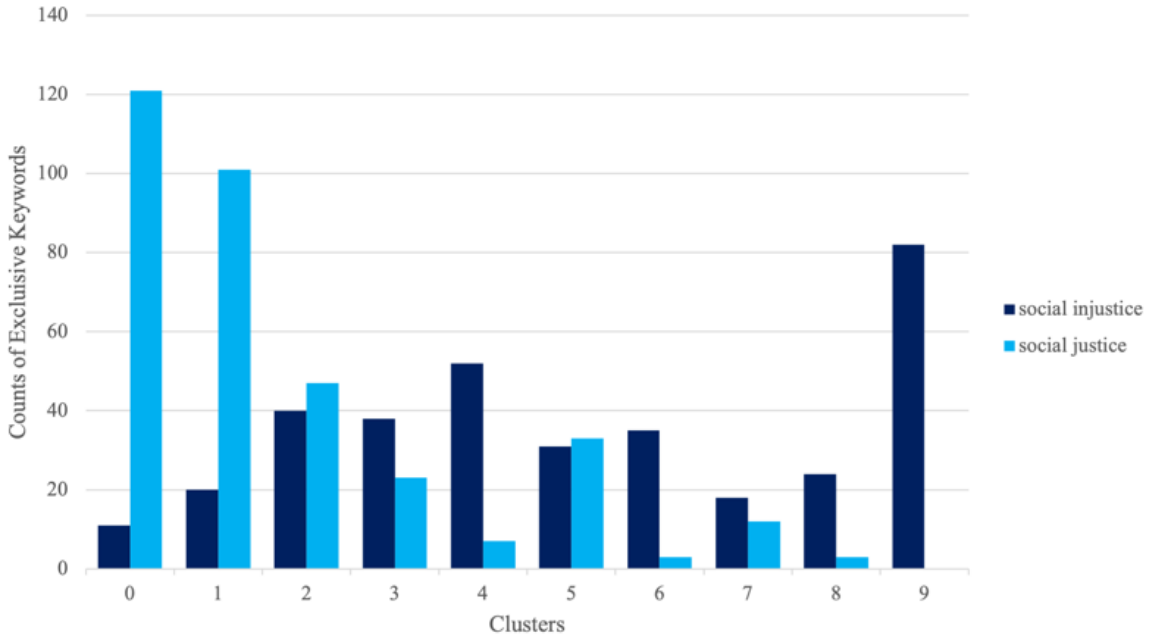

Figure 12. Keyword exclusion counts for both search results. N = 351 for of exclusive keywords for both search result.

Figure 13 shows the distribution of the entire dataset of 851 keywords across all clusters, with the lower nominal clusters containing significantly more keywords than the rest, except cluster 9. The Leiden algorithm created 22 clusters and after manual inspection of the low numbers of keywords in clusters 9 through 22, it was decided to combine these into one cluster as their indicator value was consistently close to -1 and cluster keyword counts were very small.

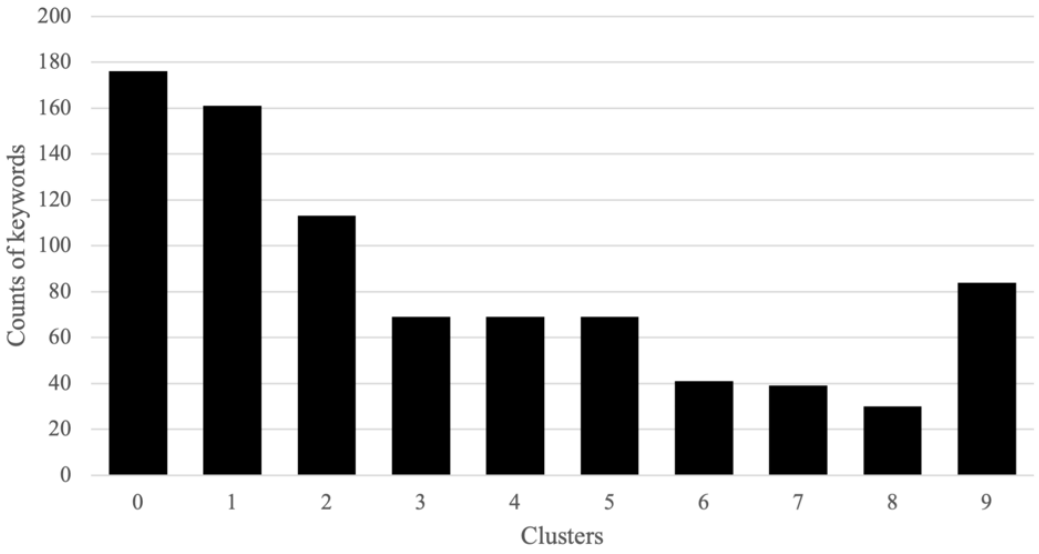

Figure 13. Distribution of all nodes, (n=851) across all clusters.

4.2.2 Clusters topics

The top 6 keywords by weight of each cluster are shown below in Table 3. These 10 clusters give us an idea of the breadth of topics within our data set. The *indicator average* also provides an idea of where these topics might be located on our graph. For example, topics in *cluster* 9 have an *indicator average* of -0.97, which suggests these topics are only found in "social injustice". Conversely, although clusters 0 and 1, as shown in Figure 12, are dominated by "social justice" exclusive keywords, the *indicator average* is more towards the center, suggesting that there are significantly shared keywords.

Table 3. 10 clusters of topics determined using Leiden clustering algorithm.

Cluster	**Keywords**	*Indicator Average*
0	justice, human rights, poverty, inequality, education, ethics	0.10937588
1	gender, race, higher education, diversity, intersectionality, social work	0.27378927
2	neoliberalism, environmental justice, democracy, migration, feminism, globalization	-0.1939534
3	equity, sustainability, participatory action research, epistemology, right to the city, cultural diversity	-0.1555504
4	social justice, affect, care, public policy, nancy fraser, emotions	-0.2541234
5	critical theory, critical pedagogy, advocacy, critical literacy, poetry, global education	-0.0140142
6	social injustice, nigeria, adolescents, school, emancipation, compensation	-0.8371718
7	inclusion, disability, inclusive education, exclusion, prejudice, ghana	-0.2155393
8	china, social inequality, air pollution, social problems, sociology, precariat	-0.7110917
9	political and moral claims, relational justice, market economy, social injustice and inequity, local development, migrant teachers	-0.9719533

4.2.3 Keyword clusters map

Figure 14 provides an overview of the cluster distributions within the same node positioning as Figure 10. Nodes, instead of being assigned color by indicator value are now identified by cluster color. Clusters 0 and 1 dominate the left side, or the “social justice” side, and nodes 4, 6, 8, and 9 dominate the right, or “social injustice” side.

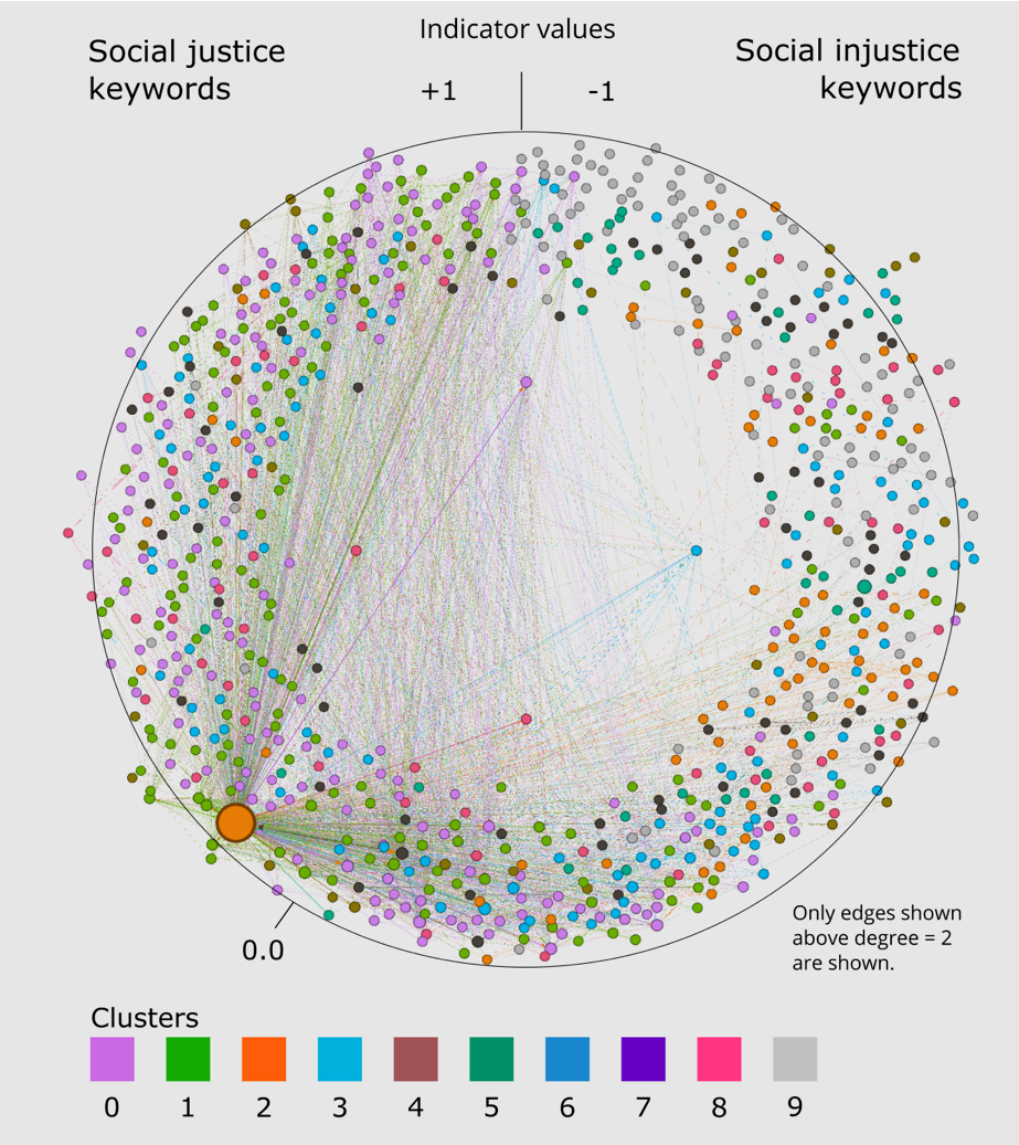

Figure 14. All nodes represented by cluster color.

The following Figure 15 shows the same clusters but plotted in an array so that one may see the distributions separately. As noted in Figure 12, nodes 0 and 1 populate the “social justice” side whereas nodes 4,6,8, and 9 are situated predominately in the “social injustice” side of the indicator range. Interestingly, node 4 contains the keyword, “social justice” and while that node is positioned on its respective side, you can see that its community lies mostly on the “social injustice” side. Nodes 2, 3, 5, and 7 seem to be even distributed across both search indicator

ranges. Remember that the Leiden algorithm was running on co-occurrences in documents, so it clustered together nodes that were frequently found together. As one document may have a keyword with a high co-occurrence, that same document may also have a very unique keyword, that is exclusive to one search or the other. In general, the clusters identify a progression of node co-occurrences by document from one extreme to the other. Practically, this means that using the keywords from cluster 0 to refine your search, will most likely find a document that is about social justice.

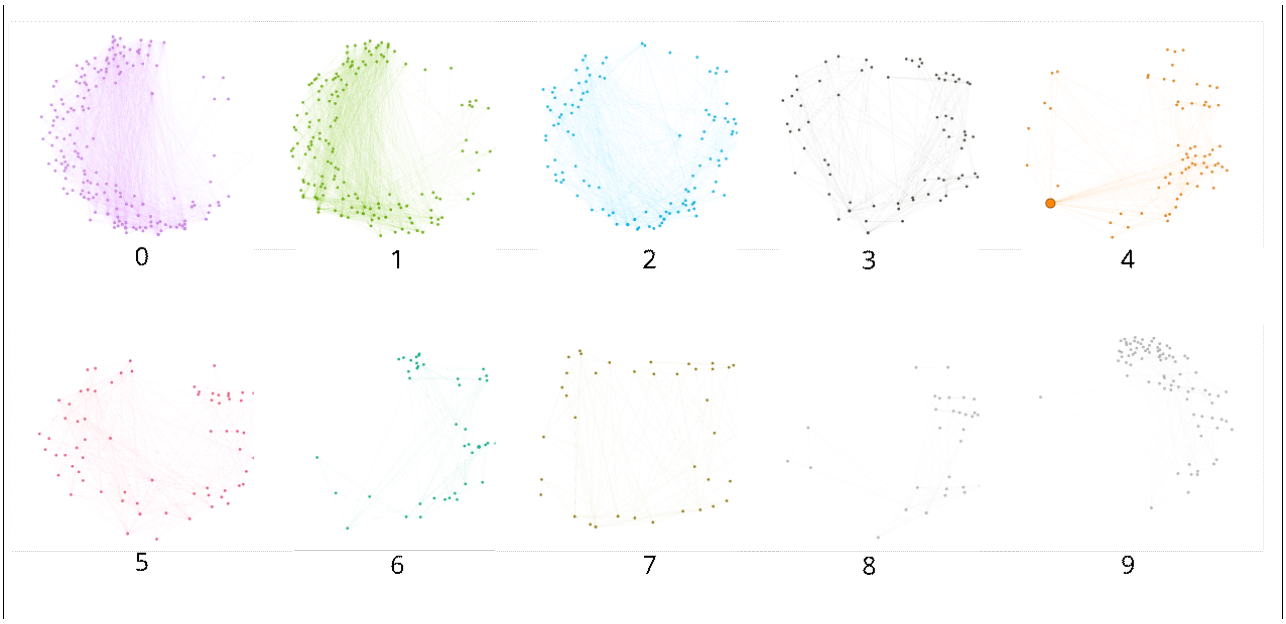

Figure 15. 10 clusters plotted individually to show distribution across indicator range.

4.4 Publication map

4.4.1 Document-based clusters using bibliographic coupling

Figure 16 shows bibliographic coupling using Gephi's force directed algorithm for layout. It reveals a central cluster of high degree relationships and an outside cluster with low to zero degree. Search origins of "social justice" or "social injustice" are indicated by node color. There

are relatively even distributions of documents from both searches in the central cluster, (giant component) as well as the outer cluster. As the visualization was created using the force-directed algorithm, Force Atlas 2, distance has some meaning as "proximity indicates communities" (Noack, 2009 cited by Jacomy et al., 2014, p. 2), and cropping outlying nodes at this level does not interrupt the structural interpretation.

Figure 16: The publication map, or bibliographic coupling, showing documents coded by their search origin, either "social justice" in pink or "social injustice" in green, (n = 931). Image has been cropped slightly to adjust for outliers whose spatial distributions are far away from the central clusters.

4.4.2 Giant component view

A detail view, seen in Figure 17, shows the central core of bibliographic coupled documents. This visualization has had filters applied to limit number of nodes and edges, (degree range of >100 and edge weight of >1) to focus on just the highly linked documents. In this figure, label size indicates normalized citations and node size indicates total link strength. Label and node color correspond to the search source as in Figure 10. At the center of this cluster, are the

documents with the highest number of keywords co-occurring (total link strength) and the highest number of links.

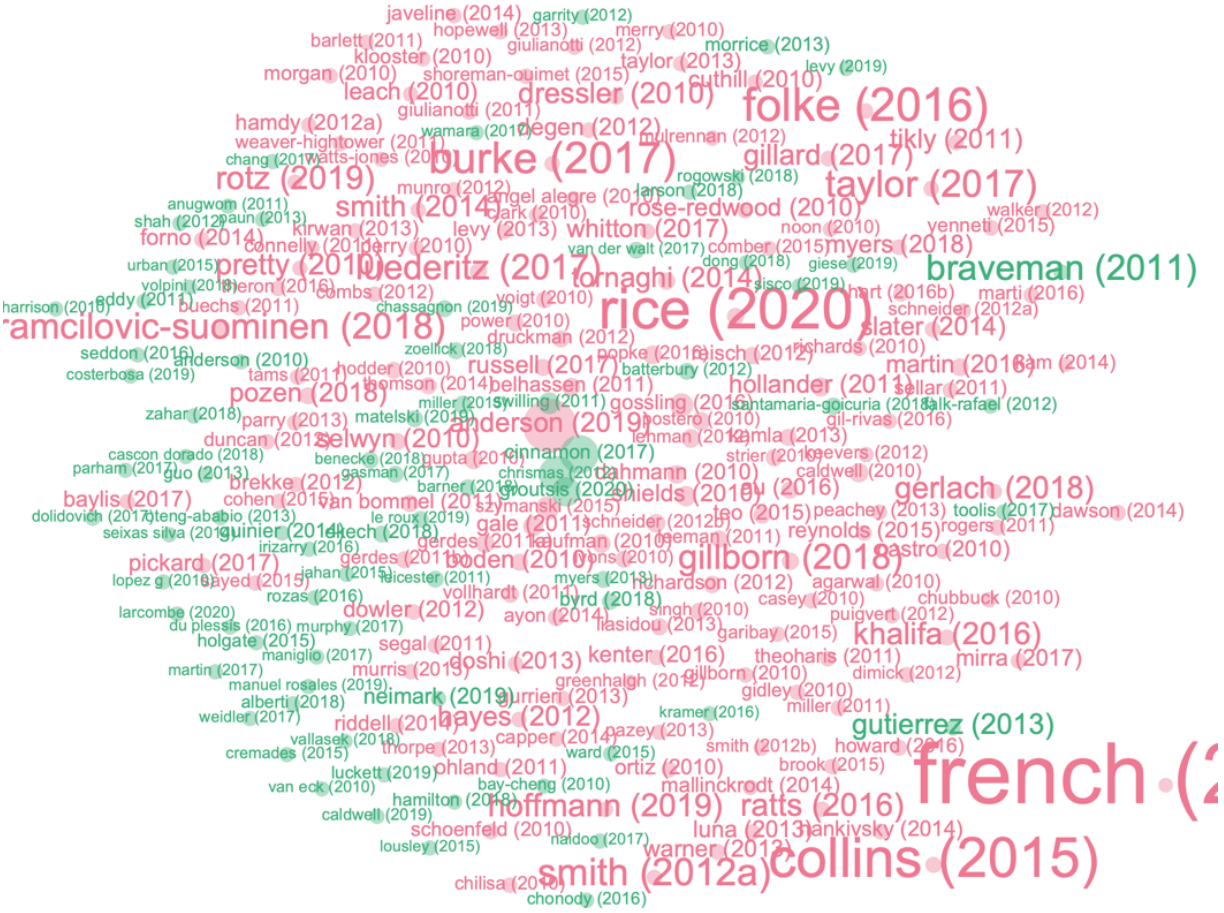

Figure 17: Detail of the giant component of bibliographic coupling from Figure 16 with the same node and label color. Label size indicates normalized citations (n=229).

The top 10 papers by total link strength are shown below in Table 4 from the bibliographic coupling with examples from both search results (indicated by the search column) and keywords manually checked from the individual documents. These represent the very center of the giant component. Looking through the document keywords, multiple general themes can be seen: education, surveillance, migration, racism, transportation, sustainability and leadership. There

are inconsistencies with the data discovered, such as with the Anderson (2019) document, that I will discuss later in Chapter 5's section on limitations.

Table 4. Top articles by total link strength from the bibliographic coupling.

Search	Label	Normalized citations	Total link strength	Title	Document keywords
1	anderson (2019)	2.7682	9485	from transition to domains of transformation: getting to sustainable and just food systems through agroecology	social justice orientation, social justice education, professional development, graduate student development
-1	cinnamon (2017)	0.9654	5847	social injustice in surveillance capitalism	corporate surveillance dataveillance social justice big data personal data
-1	chrismas (2012)	0.1197	5431	the people are the police: building trust with aboriginal communities in contemporary canadian society	none
-1	groutsis (2020)	0.9396	3686	the 'new' migration for work phenomenon: the pursuit of emancipation and recognition in the context of work	emancipation, global economic crisis, Honneth, 'new' migration for work phenomenon, skilled migrants, social injustice
1	szymanski (2015)	0.6036	2283	race-related stress and racial identity as predictors of african american activism	racism, racial identity, African American, activism
1	gossling (2016)	1.1809	2266	urban transport justice	Equality, Justice, Transport infrastructure, Transport policy, Urban planning, Urban policy
-1	swilling (2011)	0.4356	1871	reconceptualising urbanism, ecology and networked infrastructures	sustainability; urban infrastructure; resource flows; urbanism
1	shields (2010)	1.9154	1574	transformative leadership: working for equity in diverse contexts	transformative leadership, social justice, power, critique, promise, social context

Search	Label	Normalized citations	Total link strength	Title	Document keywords
1	gale (2011)	1.5537	1506	social justice in australian higher education policy: an historical and conceptual account of student participation	education policy; higher education; social inclusion; social justice; student equity; widening participation
1	caldwell (2010)	0.6228	1276	critical incidents in counseling psychology professionals' and trainees' social justice orientation development	social justice orientation, social justice education, professional development, graduate student development

4.5 Results summary

In this chapter, I have shown results from the keyword map, the keyword clusters, and the publications map detail views and supporting tables. The keyword map (Figure 10) and its detailed view (Figure 11) show how co-occurring and exclusive terms can be visualized along a spectrum and is intended as a visual aid to help information seekers through the exploratory search process. This visualization provided an overview of the structure with emphasis on the proportions of the “social justice” search compared with the “social injustice” search. Table 2 supplements the structural view with the top 10 keywords by weight for each exclusion extreme and the co-occurring zone of the spectrum. Keyword clusters were discussed in more detail with a bar graph (Figure 12) showing how clusters are affected by keyword exclusions. Figure 13 continues the discussion clusters with a bar graph showing all nodes distributed across clusters. Table 3 explores the 10 topical clusters as determined by the Leiden algorithm. The keyword clusters map (Figure 14) shows all nodes colored by their cluster using the same layout as the keyword map. Figure 15 provides an improved partitioned view of the clusters and their node distributions across the indicator range in the same circular layout. The publication map (Figure

16) provided a different overall view of the data using bibliographic coupling and nodes colored by their search origin. The giant component view (Figure 17) shows the giant component from the publication map but with filters and node labels, revealing the documents that lie at the core with the most connections to others. Table 4 explores the core further by listing the top 10 papers by total link strength and their keywords. All together, the results show two methods, that of keyword co-occurrence clustered using the Leiden algorithm and that of bibliographic coupling. In the next chapter, I will discuss these findings in more detail and offer a perspective on how an exploratory information seeker might utilize additional information such as this to support decision making.

CHAPTER 5 DISCUSSION

5.1 Introduction

In the introduction, I presented a scenario in which a student, as the exploratory information seeker, engages in searching digital resources, but is faced with the problem of not being able to see the effect of their choices other than an ordered list of possible documents. As part of getting to know the field of social justice, they are unaware of what fields, subfields and topics are contained and how other seemingly associated search terms, such as “social injustice” are related. As more and more resources are available digitally every year, the gap between novice information seekers and subject experts will continue to increase. Crossing this gap is a significant cognitive burden (Kuhlthau, 1991) and it impacts the inexperienced disproportionately (Kuhlthau et al., 2012), negatively impacting international students (Cho & Lee, 2016, p. 590). At best, crossing the gap requires significant cognitive resources to navigate academic texts (Du & Evans, 2011) with disorientation commonplace (Dillon, 2000, p. 521), impacting the information seeker emotionally (Kuhlthau, 2004) with negative consequences such as abandoning the search (White & Roth, 2009). As such, there needs to be tools that aid the exploratory seeker in understanding the scope of the field as well as comparing how their search term selection affects what they can and can’t see. Constructions of cognitive maps (Demelo & Sedig, 2021) by mediations of a structural visualization (Kuhlthau, 2004), are enabled by providing the novice information seeker an opportunity to identify waypoints (Freksa, 1999) in formulating exploratory search goals (Kuhlthau, 2004; White & Roth, 2009). The keyword map (Figure 10) attempts to address this problem by providing a structural or overall view of author defined keywords across two searches enabling an information seeker to not only compare the overlap or infer similarity between two search terms, but also to explore the keywords and their

relationships with others. The purpose of the keyword map is to reduce the time to gain subject knowledge, reduce disorientation by identifying recognizable and associated keywords, and hopefully, reduce uncertainty which may lead to more successfully completed exploratory search sessions.

In this chapter, I will explore the visualizations and results and discuss how these benefit the exploratory information seeker to answer both RO1 and RO2. I'll be discussing Figure 10, the keyword map, as a static map, identifying how it satisfies a few of the benefits of information visualization and aids during exploratory search. I'll also discuss the keyword map imagined as an interactive display in which one can zoom in and explore and how this might satisfy more of the benefits of information visualization and expand its usefulness during exploratory search. I'll also discuss the results showing the keyword clusters and identify how the tables and partitioned cluster maps may be beneficial to the exploratory searcher, from both static and dynamic states. Both the keyword map and the keyword clustering support will be explained in how they both contribute to understanding the topical differences between search terms. For RO2, the discussion will also explore how the keyword map (Figure 10) is grounded by comparing its results with the publication map (Figure 16), as this bibliographic coupling method has been established as a means of identifying topical communities across a body of papers. I'll conclude with limitations of the current study and its data before summarizing interpretations of the data presented.

5.2 Research Objective 1

The first research objective investigates an effective way of visualizing the effects of search term choice so that included or excluded keywords might be compared between the two search results.

By employing a keyword co-occurrence analysis and a clustering algorithm, the keyword map (Figure 10) shows a truncated selection of 851 keywords from both searches that are color coded according to their relative occurrences within each set of search results.

5.2.1 The Keyword map, or structural view

"Social justice" and "social injustice" were chosen as search terms based on their perceived associations and their difference in scale. Figure 10 shows that edges equal to or above 2 populate most of the "social justice" side, suggesting there are more co-occurrences within the "social justice" keywords than the other. There is also a sense of scale with the largest node being that of the keyword "social justice". There is surprisingly little overlap between the two, those being represented by the spectrum of colors at the bottom of the dual circle layout but co-occurrence in this area is significant and meaningful as the search term, "social justice" is in this area represented by the largest node with the most edges. In contrast, "social injustice" is located at about the 3 o'clock position, is smaller in node size and has significantly fewer edges (with the filter applied).

As a static image, the keyword map provides an immediate read on the proportion of overlap between the two terms and the amount of exclusion. This is its most important attribute to answering RO1 – how do two search terms compare with regards to excluded topics and overlapping topics? The node color helps to understand the transitions from extreme values where exclusivity is occurring and the intermediary colors that suggest some level of co-occurrence. But it is the overall proportion of one to another that provides the information seeker with some sense of similarity between the two. As a static image, it may provide benefits of visualizing the overall structure, pattern recognition and facilitating cognitive map formation.

In contrast with the study by González-Valiente et al. (2021) which investigated quantitatively the similarity between two terms, the keyword map serves more as a qualitative tool to enable *discover, learning,* and *investigation* as part of the *exploratory browsing* stage rather than establishing thresholds for similarity. The question of thematic or topical similarity can be inferred loosely from *pattern recognition* and exploring *proximity relationships*, particularly in Gephi, where highlighting one term can help identify other co-occurrences. At this point, the need for interactivity becomes evident as this level of detail would likely exceed *display limitations.* While tables are useful, such as in Table 3, these are not currently part of the static visualization and would be much better suited to support an interactive visualization. As an example, the following Figure 18 shows how using an *interaction and filter* approach, can be useful to finding terms that cluster and co-occur together, supporting visualization benefits of *pattern recognition*, *proximity relationships*, and possibly enhancing the *cognitive map formation*, from the previous overall structural view.

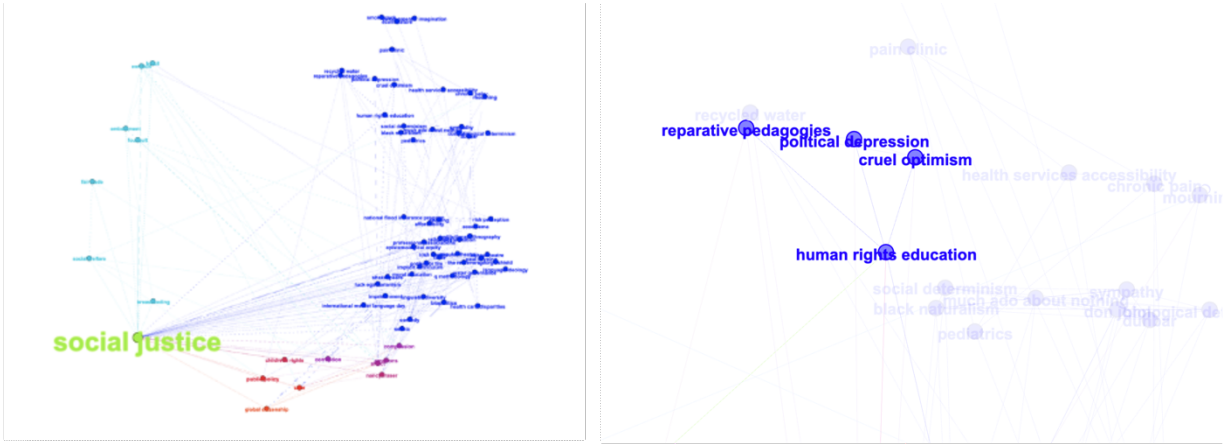

Figure 18. Cluster 4 isolated in Gephi (Left) and a subset of cluster 4 selected (right).

5.2.2 The keyword map detail view with node labels

The keyword map detail view in Figure 11 shows the two extreme ends of the scales representing keywords that are exclusive or near exclusive in their frequency of co-occurrence. Keywords on both sides of the exclusive areas signify general topics, such as conflict, whiteness, and inflation, as well as very specific topics, such as preservice teachers, simone de beauvoir, and accumulation by dispossession. Recall that the visualization is sorted by indicator value first and then by cluster, so proximity of terms is first organized by its frequency across both search results and then by its co-occurrence with other terms. While generalizations about the exclusive keywords are difficult to construct, the exclusive areas may be of interest to the information seeker, as they represent, 'what am I missing if I use the other search term?'

In contrast with faceted tools for precision retrieval or interfaces for extraction, this combination of overview, zoom and filter, and details-on-demand (Shneiderman, 1996) is appropriate for the exploratory information seeker by enabling recognition of terms to act as waypoints and places where the information seeker can return to identify and explore new keywords. I suggest this helps to alleviate uncertainty and its emotional states by facilitating a cognitive map, or building upon the 'personal construct' (Kelly, 1963 cited by Kuhlthau, 2004) that enable an exploratory searcher to create new waypoints on the way to a formulated goal (Demelo & Sedig, 2021; Freksa, 1999; Kuhlthau, 2004). Within the White & Roth exploratory search process, being able to *discover, learn*, and *investigate* at this level with keywords visible may lead the information seeker from *exploratory browsing* towards *focused search*, with new keyword knowledge being applied to *query formulation* and *result examination*. As noted in Chapter 2, this is a dynamic, back-and-forth process until information needs have been satisfied.

5.2.3 Conclusion for the keyword map

As an initial, structural view of the two search terms, “social justice” seems more connected and relevant to a wider range of topics than “social injustice” based on the static keyword map (Figure 10). There is also the impression, that these are two different things, even though they may be used in similar contexts. However, to get a more nuanced understanding of the keyword map, exploring topics supplemented by tables or by exploring the visualization in Gephi with zooming and filters, requires more than a static map might be able to provide. As an initial structural comparison, however, it provides value for seeing if there is significant overlap. In the case of “social justice” and “social injustice”, though they are linked through association, they seem different in their specificity.

Visually, the keyword map visualization has limitations as an overall view. Using Munzner & Maguires’s components of visual communication, there are issues of *scalability,* specifically, *cognitive capacity* and *display capacity*. While measuring *cognitive capacity* is beyond the scope of this study, I acknowledge there is a limit to how many terms that can be memorized before short-term memory is exhausted. The advantage of a static visualization is that one can return to it and it remains the same. *Display capacity* is affected as seeing the keywords is not possible at the overall map view. Adding the keywords table may be a possible solution, but it is more quantitative than aiding exploration as it may mislead with the list being perceived hierarchically. Solving the *display capacity* problem, requires an interactive visualization, where *interaction and filtering* can enable the keyword map to *visualize an overall structure*, *expand the working memory*, *enable pattern recognition*, and allow the *inference of proximity relationships*, and may lead to a greatly enhanced *cognitive map*.

5.3.4 Keyword clusters

Table 2 provides the top keywords by weight for each of the clusters along with the average indicator value, (average based on the indicator weights of the top 6, not the entire cluster). The clusters show a distribution of indicator values with cluster 0 indicating it is closer to the mid-point of 0 in contrast with cluster 9 being exclusively “social injustice” with an indicator of -1. While an overarching theme was not assigned or derived from each cluster, one can interpret the cluster 6 as a group representing the range of topics. As this study is from the perspective of exploratory search and not faceted, precision search, the keyword clusters should contribute to an understanding of the intellectual structure of the whole. For example, cluster 4 contains the keyword, “social justice”, along with other keywords of “affect”, “care”, “public policy”, “nancy fraser”, and “emotions”, which are related to social justice by frequency and co-occurrence. This does not suggest the information seeker will not find topics on emotions in other clusters, but it does provide a good starting point for identifying an area of the structure view that may provide keywords that can be used in *query (re)formulation* as the information seeker moves from *exploratory browsing* to *focused search*.

Figure 12 and Figure 13 reveal the extent of how nodes are distributed across clusters. Of particular interest to the information seeker, may be Figure 12 in which search results are defined by their distribution across clusters, providing additional information to build upon the keyword map detail view exploration. For example, clusters 1 and 2 are mostly results from “social justice” and clusters 4, 6, and 8 are mostly results from “social injustice”. As clusters are based on co-occurrences using the Leiden algorithm and proximity is sorted by occurrence and co-occurrence similarity, the information seeker may be interested to know if those clusters are widely distributed or confined to an area within the double circle layout. Figure 14 provides the

clusters visually utilizing the same node layout as seen in Figure 10, so that the viewer, already with a familiarity of the search result distribution, may get a sense of how clusters are mapped spatially. With 10 cluster colors, the initial impression may be identifiable patterns, or may look like confetti sprinkles on a cake, depending on the search terms being compared, as they do here. The keyword, “social justice”, (not to be confused with the search term but the specific keyword found in the author specified keywords in the WoS document record,) can be found just above the spectrum midpoint. As a static view in this context, it is not possible to understand the *overall structure*, or *identify patterns,* so this visualization has little value as is. However, a more nuanced and possibly informative view comes from Figure 15, in which the individual clusters can be seen as distributions across the spectrum. The keyword map (Figure 10) provides confirmation of Figure 12 with most cluster links in the “social justice” side of the search results. This suggests the highest weighted edges, (those with edge values > or = 2) are going to be found in “social justice” results. As expected from Table 3, clusters 0 and 1 are mostly located in the “social justice” side and clusters 2, 3, and 5 distributed all around the circular layout. For the information seeker, if keywords in these topics are of interest to them, then both search terms will serve your needs well and provide a wide range of topics. However, if your area of interest lies in cluster 9, then you would want to use “social injustice” as the exclusions are significant.

5.3.5 Conclusion for keyword clusters

The intent behind mapping the clusters of keywords is to understand how topics, as clusters of keywords, as located around the map and to aid awareness of excluded topics. Remember, that each node represents one of the most frequent keywords, and simultaneously, represents multiple documents. So, understanding where highly connected keywords are located in exclusive areas

also suggests there are exclusive documents and topics that an information seeker may want to acknowledge.

Looking at the keyword clusters from Table 3 and the keyword clusters partitioned maps of Figure 15, these may provide some sub-structure to aid in cognitive map construction particularly with awareness of excluded areas. The brevity of the table may be beneficial for understanding where to start looking, supporting an *overview of the sub-structure* of the keyword map, and *pattern recognition* and *proximity relationships* may have more meaning with the partitioned maps Figure 15. As static maps individually, they may be too cumbersome, but combined with the keyword map (Figure 10) as an interactive display, they may help to *improve working memory* and c*ognitive map formation*.

The cluster information in Table 3 and partitioned maps of Figure 15 from an exploratory search process perspective, may be helpful, particularly for crossing the knowledge gap from *investigation* to *query (re)formulation*. This is a critical stage at which uncertainty and its companion, anxiety, are replaced with increasing confidence. Being able to quickly focus in on a cluster may be of benefit that may merit inclusion within an interactive visualization.

5.3.6 Answering RO1

The comparison of the keyword map (Figure 10), Table 3, and the keyword clusters maps (Figure 15) provide a view of the intellectual structure of the search results from two independent searches and satisfy RO1 which may present an effective way of visualizing the effects of search term choice so that included or excluded keywords can be compared between the two search results. Instead of a simple visualization, however, I have provided a more thorough means of *discovery, learning*, and *investigation*, to support the information seeker

during exploratory search. This visual artifact as a static map is sufficient to be considered as a theoretical contribution for further analysis. However, it is evident that a static map is not sufficient and that an interactive map supported with tables and cluster maps would be needed to fulfil all six benefits of information visualization. With the goal to support exploratory search, particularly in reducing the uncertainty and anxiety to reduce the time it takes to move from *exploratory browsing* to *focused search*, the static maps would not satisfy this need and would be burdened with issues of *scalability*. From this investigation, the visualization method of a dual circle layout using keyword co-occurrence, clustered with the Leiden algorithm provided seems valid, but would need to be developed into an interactive display to meet *validity* and *scalability* needs and to fulfill its potential as an effective visualization.

5.4 Research Objective 2

To answer RO2, I compare the keyword-level analysis with a document-level analysis using bibliographic coupling to determine if topical exclusivity exists between search terms. While acknowledging that different analysis and mapping methods are going to produce a different narrative (Klavans & Boyack, 2017), what might an information seeker infer from this method with regard to comparing the two search terms? Do the two stories align?

5.4.1 Document-based bibliographic coupling clusters

As discussed in Chapter 2, bibliographic coupling creates topic similarity (Klavans & Boyack, 2017). Figure 16 shows the resulting publication map visualization of bibliographic coupling at the document level for the combined search results. The overall structural view is comprised of two main clusters. The giant component in the center suggests a tightly interconnected group of documents that are significant with very high total link strengths (the number of publications in

which two references are shared). The edge degree (or the link strength or number of references shared) in the central giant component are very high in contrast with the outer group, where edge degree is quite low and easily disappears as edge filters are applied.

Nodes are coded by their search origin and at this overall level, it appears that both search results are well distributed. Remember from Chapter 3, that the bibliographic coupling was run on a subset of documents, so the proportions of the two are not representative of the actual proportions of the two search results, (see Table 1). However, its usefulness as a static image provides an *overall structure* of the two results and there is a *pattern* consisting of two bodies, one giant component and a less cohesive body of low degree to non-degree edges. At this level, there is a different narrative beginning to take shape – that search results are well dispersed across the two bodies, with a core set being highly connected.

5.4.2 Giant component view

The giant component, as shown in Figure 17, is the central body seen in the publication map, (Figure 16). Nodes have been reduced, labels including the first author and publication date have been added, and label size indicates normalized citations. As it is, the giant component suggests highly important literature to both searches, although it does appear that the "social justice" category is more numerous. While the central nodes in Figure 16 were largest, (indicating total link strength), this does not necessarily relate to normalized citation strength. Though there are some highly cited documents, what interests us most are those with the most connections, thus the most relevance topically.

The results suggest that both search results are distributed across document-based clusters with some minor clustering for "social injustice" on the left and a more dominant presence of "social

justice" on the right side of the giant component. Where documents have a very high degree and total link strength, such as in the giant component in Figure 17, both search terms seem to provide connections to what may be considered the most significant literature.

Currently, however, it is unknown if this is the case with other search term results and the importance of including this visualization into the wholistic view for decision making. While an exploratory searcher may be a novice within a subject area, it may also include those that are expert researchers and would value understanding the bibliographic connections or have concerns about missing out on significant works. However, now the conversation has started to move towards formulation of desired outcomes and collection of specific works (Kuhlthau, 1991, 1999a, 2004) – a precision task that is outside the scope of this study. From the lens of the exploratory search model, this visualization may be useful after exploratory browsing, where additional keywords have been used to refine the search down to a smaller set of papers. IF there were a view, like the giant component detail view of Figure 17, or a list of the most central papers in Table 4, this may speed the exploratory searcher during the *focused search* stage.

Table 4 provides a quick view of some of the core papers by total link strength in the giant component. A manual review of these papers showed that most of them are argument papers and do not contain any systematic reviews, which may cause high numbers of links. As noted in the Results section, there is a variety of topics contained within the top 10 papers. From the perspective of an exploratory searcher within the context set in this study, this may be considered a benefit and an indicator that both search terms are equally useful.

5.4.3 Answering RO2

The second research objective was to compare the keyword-level analysis with a document-level analysis using a proven bibliometric method to determine if topical exclusivity exists between search terms. Examining the publication map (Figure 16) and its giant component view (Figure 17), along with the top 10 papers table (Table 4), produces a different narrative than what was interpreted from the keyword map and its supporting views. The publication map suggests that both search terms will lead the information seeker to the most significant literature, while the keyword map suggests that there are significant areas of exclusion and that the two search terms are referring to different, not synonymous meanings.

As an answer to RO2, I can't infer that the keyword map is enough on its own to understand how these two search terms affect the literature being returned from the academic search interface. Both used together, would provide a more thorough and balanced tool with which an information seeker during exploratory search could overcome their subject area knowledge gap with less time and more direction. However, as noted in answering RO1, visualizations of the keywords and the publications are very limited in their usefulness as static images due to *scalability* issues, and to present an effective tool to address the emotional impact of uncertainty during exploratory search, an interactive tool with keyword level and document level visualizations should be explored in future work.

5.5 Limitations

This exploration of a visualization method is limited in several ways, such as data processing methods, limitations of Gephi's layout methods, clustering methods, and citation-based selection of a data subset. Currently data processing is occurring in both a Python-based notebook and in

Gephi, which creates a workflow that would be challenging for others to reproduce. The solution would be to move this entirely to a web page with the Python notebook driving both the data interpretation and visualization.

This study is currently dependent upon Gephi's layout methods and algorithms. The Dual Circle layout was chosen for its ease of use, consistency, and ease of comprehension as it is just a linear scale wrapped in a circle. Other visual idioms should be explored for readability and tested by novice and expert information seekers. Given that the terms exist on a scale from -1 to +1, perhaps variations on linear visual idioms may be helpful. Additionally, other distance-based mapping clustering algorithms may provide improvement on interpretability.

Currently the keyword map (Figure 10) shows exclusions most saliently. As the number of nodes increases in the overlap areas, navigation and recognition of topics may be increasingly more difficult. Different means of creating clusters, such as by semantic similarity or formal field designations should be explored for its benefits during *exploratory browsing*.

The bibliographic coupling data subset was selected based on sorting by citation counts from the full records downloaded from WoS. So, the documents in bibliographic coupling are the most highly cited, and the keywords in the keyword map and clusters analyses are the most frequent. These two do not necessarily coincide and this may have presented two different narratives with what is essentially, two different data subsets. In future versions, it should be explored how to use the same documents for both analyses.

Another limitation to this study's interpretation is that the visualization is static and has disadvantages as noted by Buzydlowsky (2002). There is a need to make both analyses real-time with the addition of dynamic query so that search terms may be continually examined.

One inconsistency was found with the bibliographic results data in Table 4, in that the manually retrieved keywords of the top paper, that of Anderson (2019), do not match those from the downloaded dataset from WoS. It is unknown if keywords were indexed from another version of a paper, (such as a preprint). All other papers from the top 10 in Table 4 are consistent with the dataset downloaded from WoS. While it is known how many papers were dropped due to lack of keywords in the 'DE' column from the full record, it is not known how many are inaccurate.

5.6 Conclusion

The first research objective was to identify an effective means of visualizing and comparing the results of two different search terms. The keyword map (Figure 10), Table 3, and the keyword clusters maps (Figure 15) provide a view of the intellectual structure of the search results from two independent searches and satisfy RO1 which is to present an effective way of visualizing the effects of search term choice so that included or excluded keywords might be compared between the two search results. The method provided is available to someone with basic Python proficiency and experience with Gephi. The layout of a dual circle with nodes distance-mapped along a spectrum provides some measure of visual comparison especially with exclusive keywords and co-occurring keywords. This visualization attempts to address issues of uncertainty identified during the exploratory search process (Kuhlthau, 1999a, 2004; White & Roth, 2009). By visualizing the list-based results from an academic database, keywords mapped by their occurrences within each search may provide an opportunity for the information seeker to

construct cognitive maps (Demelo & Sedig, 2021) and identify waypoints (Freksa, 1999) in their formulation of an end goal (Kuhlthau, 2004). This visual artifact improves on current practices and may reduce cognitive load, as conceptualized by Sweller et al., (2011) and should be considered as a theoretical contribution for further analysis and investigation.

The second research objective is more difficult to answer as interpretations of results may be affected by what two search terms are compared. While the bibliographic coupling method shows that both search terms will provide significant results, the keyword co-occurrence analysis and visualization, shows there are significant areas of exclusions that affect these two terms. As an answer to RO2, I suggest the keyword map may need supplementary views, tables, or even other methods, such as the bibliographic coupling to support an understanding of how these two search terms affect the literature being returned from the academic search interface. As discussed in RO1 above, the visualizations of the keywords and the publications are very limited in their usefulness as static images due to *scalability* issues. To present an effective tool to address the emotional impact of uncertainty during exploratory search, an interactive tool with keyword level and document level visualizations should be explored in future work.

So, one method suggests, either term will work, and the other method suggests, they are quite different. This is not to be unexpected, as Klavans & Boyack state that maps are validated by "the observations and stories associated with the partitioning, structure, and dynamics of these maps, and our association of these observations with reality, are the things that give our maps face validity, make them compelling, and make us want to dig a little further" (Boyack & Klavans, 2010, p. 2393). González-Valiente et al. make a similar comment with "…term analysis should not only be based on merely quantitative measures since the most important is to

understand the relationships and meanings of words in their context" (González-Valiente et al., 2021, p. 343).

With these two perspectives mind, neither map suggests one search term over the other. Both are qualitative as they are being read visually for differences that provide some guidance and invite further inquiry. In this, they perhaps work equally well in discriminating where differences must be evident. Proportionally, it can be seen that "social justice" is more connected and is more common to both, than exclusive. At the document level, there is also an indication that "social justice" occupies a little more of the giant component. But the goal of this study has been to make evident what is excluded, so the keyword visualization clearly shows exclusions, just as the bibliometric map clearly shows which documents are excluded. The maps must be interpreted by the information seeker for their own complex, dynamic information requirements.

In this chapter, I have provided an interpretation of the results and discuss how the research objectives were met and have provided insights into future research directions. I finished up this chapter with identification of limitations found over the course of this study and how some of them provide insight for future work. In the next chapter, I summarize the position of this research, present key findings and suggest theoretical and practical applications, and provide thoughts on future research and developments.

CHAPTER 6 THE FINAL, AND QUITE THE LAST, CONCLUSION

6.1 Summary

This study compared keyword co-occurrence between two searches illustrating how seemingly synonymous terms contain not only many topics, but also have exclusions that are unique to each search. Thus, if one were to perform a search on only one of these search terms, then the user would not be aware of what had been missed. This visualization attempts to address issues of uncertainty identified during the exploratory search process (Kuhlthau, 1999a, 2004; White & Roth, 2009). By visualizing the list-based results from an academic database, keywords visualized by their occurrences within each search may provide an opportunity for the information seeker to construct cognitive maps (Demelo & Sedig, 2021) and identify waypoints (Freksa, 1999) in their formulation of an end goal (Kuhlthau, 2004) as they work through the exploratory search process.

6.1.1 Position of the research

This study was positioned within LIS and examined the challenges of exploratory search in academic libraries. I have applied quantitative methods to compare keyword distributions and clusters of similarity within a distance-based network map. While dependent on social constructivist theoretical perspectives (Kuhlthau, 2004) for situating the context of the information seeker (Agarwal, 2018) and their interaction with information mediations (Demelo & Sedig, 2021; Kuhlthau, 2004), I interpret the results of the visualizations qualitatively referencing White & Roths's model of exploratory search (White & Roth, 2009), which establishes the emotional impact of uncertainty during the information seeking process. Evidence for decisions was drawn from studies from LIS, BIM, HCI, CHI, as well as graphic design.

6.2.1 Chapter summaries

Chapter 1 introduced the study within the context of a scenario in which a student is engaged in exploratory search of an unknown topic. The problem was defined that current academic discovery systems provide list-based results pages, and this makes it cognitively difficult and time consuming to develop an overall view of the subject area. This can lead to cognitive overload when trying to get through the uncertainty during exploratory search. Information visualization has been shown to alleviate these problems.

Chapter 2 provided an overview of exploratory search, information seeking models and uncertainty, information visualization, and appropriate methods for this study, along with past studies and current applications that attempt to solve the same problem. The first section explained exploratory search and how White & Roth's (2009) definition of the process acknowledges uncertainty as presenting challenges to the information seeker. The second section on information seeking behavior models showed how exploratory search fits within information seeking behavior models, such as Hoeber & Shukla's modified White & Roth model (2022), with emphasis on Kuhlthau's ISP model (Kuhlthau, 1991, 1999a, 1999b, 2004; Kuhlthau et al., 2012). I also explain how cognitive map construction (Demelo & Sedig, 2021), through the use of visualizations may mitigate uncertainty and improve the exploratory search process. The third section revealed the benefits of information visualization (Börner, 2010; Card et al., 1999), reviewed an evaluative framework from Munzner & Maguire (2015), and introduced Kosslyn's (2006) eight principles of graphic design. The last section covered the history and usage of two methods, keyword co-occurrence analysis including the Leiden clustering algorithm, and bibliographic coupling. I also examined prior studies and applications using keyword-based or citation-based methods, where visualization is the main interaction component for the

exploratory searcher. I concluded this chapter establishing the delimitations and identified how this study's research objectives are unique.

Chapter 3 described the methods used to answer both research objectives using two analyses. To answer primary research objective (RO1), I used keyword co-occurrence analysis to create a keyword map and keyword clustering results. The second research objective (RO2) compared the topics revealed between each search term by using bibliographic coupling to create a publication map, which reveals intellectual structure by finding common references between documents.

Chapter 4 showed results from the keyword map and its supporting views and tables, the keyword clusters, and the publications map with supporting detail views and tables. The keyword map (Figure 10) and its detailed view (Figure 11) show how co-occurring and exclusive terms can be visualized. They provided an overview of the structure with emphasis on the proportions of each search result. Table 2 supplements the structural view with the top 10 keywords by weight for each exclusion extreme and the co-occurring zone. Keyword clusters were discussed in more detail with a bar graph (Figure 12) showing how clusters are affected by keyword exclusions. Figure 13 shows all nodes distributed across clusters, and Table 3 lists the 10 topical clusters. The keyword clusters map (Figure 14) shows all nodes colored by their cluster with Figure 15, the partitioned view of the clusters and their node distribution. The publication map (Figure 16) provides a different overall view of the data using bibliographic coupling and nodes colored by their search origin. The giant component view (Figure 17) shows the giant component from the publication map and Table 4 lists the top 10 papers by total link strength and suggested topics.

Chapter 5 framed the interpretation of the results to answer the research objectives from two points of view: the benefits provided to the information seeker and heuristically identifying how the visualizations fulfil the benefits of information visualization.

The first research objective was to identify an effective means of visualizing and comparing the results of two different search terms. The keyword map (Figure 10), and the keyword clusters maps (Figure 15) provide a view of the intellectual structure of the search results from two independent searches and satisfy RO1 which present an effective way of visualizing the effects of search term choice so that included or excluded keywords might be compared between the two search results.

The second research objective provided insight that the keyword map may need supplementary views, tables, or even other methods, such as the bibliographic coupling to support an understanding how these two search terms affect the literature being returned from the academic search interface. Another insight from RO2 is that static images are limited in their usefulness due to *scalability* issues. To present an effective tool to address the emotional impact of uncertainty during exploratory search, an interactive tool with keyword level and document level visualizations may be more effective and should be explored.

6.2 Key findings

The study resulted in the following four key findings:

- Visualizing and comparing the author-defined keywords from two independent search results is possible and may provide benefits to the exploratory search process.
- Search terms may have significant exclusions and co-occurrences that affect topical areas. Visualizing to compare these exclusions may benefit exploratory seekers and others who

need to understand how search terms may impact returned results from an academic database.

- Bibliographic coupling and keyword co-occurrence produce different narratives of topical inclusion or exclusion. Yet, both may be needed to provide a more complete picture for effective decision making during exploratory search.

- Static images are limited in their usefulness and future work must be directed towards interactive visualizations to achieve the full benefits of information visualization.

6.3 Importance of the study

6.3.1 Theoretical implications

This study draws heavily from the constructivist theory-based work of Kuhlthau (Kuhlthau, 2004), the definition of exploratory search by White & Roth (2009), and definitions of cognitive maps by Demelo & Sedig (Demelo & Sedig, 2021). While this work does not affect those building blocks, I recognize that I have placed much emphasis on uncertainty and the other emotions such as anxiety and frustration as motivation for this thesis. At this writing, I am unaware of research further defining to what extent those emotions affect information seekers detrimentally within LIS.

The theoretical implications of this study advances research related to visualization and cognitive load, as discussed earlier. In this study, I propose a new information search artifact which can be studied by information researchers who want to investigate its impact on cognitive load and subsequent emotional states.

6.3.2 Practical applications

This study's key findings may find three practical applications in the future with more development. The keyword map, as it is, may benefit the academic librarian with its simplicity, and ease of implementation, compared with more complex solutions that have been developed by

past studies to address similar problems identified in exploratory search. Reference librarians may wish to illustrate the effects of search term choice during exploratory search. As part of instruction, reference librarians may wish to explain the exploratory search process more fully, reassuring inexperienced information seekers that the emotional states during search are expected and a normal part of the process (Kuhlthau, 2004).

For curriculum design or strategic planning, this type of visualization comparing keywords may be useful for academic administrators in evaluating the accuracy of their program field descriptors, as shown by González-Valiente et al. (2021) where they evaluated the terminological similarities between information management and knowledge management. Similarly, it could also be used to analyze the similarity of terms in thesauri to aid decisions of narrow, broad, or related taxonomy relationships.

Returning to the original scenario, a keyword map visualization as seen in Figure 10 may also benefit the student who wishes to communicate their exploration process as a means of validating their decision-making process.

6.4 Future research

This study has yielded insights that may direct future research of improvements to the visualization, validation of its effectiveness, and applications.

6.4.1 Future development

The visualization is currently static and requires experience with Jupyter notebooks to implement. Future work will focus on creating an interactive visualization within an interface with an API to an open database, permitting dynamic queries to be compared. Other methods

using natural language processing to aggregate semantically close keywords should be explored. Furthermore, instead of comparing separate searches, another version of the Keyword Map should compare against results from utilizing Boolean operators. This approach was used in the terminological similarity study from González-Valiente et al. (2021).

6.4.2 Validation

Effectiveness of the Keyword Map and any supplemental forms needs to be measured. From White & Roth (2009), metrics of "engagement, enjoyment, novelty, time, and success, and learning" including cognitive load and fatigue should be evaluated longitudinally to understand how users of such a visualization, if integrated into a system, benefit when compared against existing systems. Furthermore, actual user feedback has not been performed, particularly with respect to Dillon (2000) who advised studying and consideration of the needs of individual differences in cognitive abilities and knowledge-base that may affect understanding of spatial or semantic cutes in a visual interface/map. Given that this is 2022 and the conversation about inclusivity, including support for accessibility in digital resources, has been going on for years, this study falls short of those considerations and should be addressed in future work. Methods of participatory research using Information World Mapping methods in which the information world views of novice information seekers which drives user-centered design of interfaces (Greyson, 2019) should be explored. Finally, more validation or comparison against other keyword extraction methods and clustering should be explored. How do the author keywords align with TF-IDF methods and would this increase or decrease the exclusions? Other means of clustering, such as cosign similarity with KMeans, Latent Dirichlet Allocation, or Ward Hierarchical, should be explored for comparison.

6.4.3 Applications

From citing the need to address the additional uncertainty faced by international students (Cho & Lee, 2016), a future study should perform usability testing to determine if this visualization is effective for international students, particularly those with self-identified ESL challenges. Currently, the data source privileges English language, so our dataset reveals English words as a mapping. Other databases, such as SciElo, Asian, African, or Canadian indigenous content, should be explored in the future with the intent to investigate this type of visualization for articles that may include a mix of languages.

REFERENCES

Agarwal, N. K. (2018). *Exploring context in information behavior: Seeker, situation, surroundings, and shared identities*. Morgan & Claypool Publishers.

Agostinho, S., Tindall-Ford, S., & Bokosmaty, S. (2014). Adaptive Diagrams: A Research Agenda to Explore How Learners Can Manipulate Online Diagrams to Self-Manage Cognitive Load. In W. Huang (Ed.), *Handbook of Human Centric Visualization*. Springer New York. https://doi.org/10.1007/978-1-4614-7485-2

Alsallakh, B., Aigner, W., Miksch, S., & Hauser, H. (2013). Radial Sets: Interactive Visual Analysis of Large Overlapping Sets. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2496–2505. https://doi.org/10.1109/TVCG.2013.184

Athukorala, K., Micallef, L., An, C., Reijonen, A., Peltonen, J., Ruotsalo, T., & Jacucci, G. (2017). Visualizing activity traces to support collaborative literature searching. *Proceedings of the 10th International Symposium on Visual Information Communication and Interaction*, 45–52. https://doi.org/10.1145/3105971.3105981

Ayers, W., Quinn, T. M., & Stovall, D. (2009). *Handbook of Social Justice in Education*. Routledge.

Bascur, J. P. (2019). An interactive visual tool for scientific literature search: Proposal and algorithmic specification. *BIR 2019 Workshop on Bibliometric-Enhanced Information Retrieval*, *2345*, 12.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media,* San Jose, California. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

Bates, M. J. (1989). The Design of Browsing and Berrypicking Techniques. *Online Review*, *13*(5), 407–424.

Bates, M. J. (2005). Information and Knowledge: An Evolutionary Framework for Information Science. *Information Research: An International Electronic Journal*, *10*(4). https://eric.ed.gov/?id=EJ1082014

Berube, M. A., Schorr, D., Ball, R. J., Landry, V., & Blanchet, P. (2018). Determination of in situ esterification parameters of citric acid-glycerol based polymers for wood impregnation. In *Journal of Polymers and the Environment* (Vol. 26, Issue 3, pp. 970–979).

Börner, K. (2010). *Atlas of Science*. MIT Press. https://mitpress.mit.edu/books/atlas-science

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404. https://doi.org/10.1002/asi.21419

Brooks, T. A. (1995). People, words, and perceptions: A phenomenological investigation of textuality. *Journal of the American Society for Information Science*, *46*(2), 103–115. https://doi.org/10.1002/(SICI)1097-4571(199503)46:2<103::AID-ASI4>3.0.CO;2-7

Buzydlowski, J. W., White, H. D., & LIn, X. (2002). Term Co-occurrence Analysis as an Interface for Digital Libraries. In K. Börner & C. Chen (Eds.), *Visual Interfaces to Digital Libraries* (1st ed., pp. 133–144). Springer-Verlag.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. Social Science Information, 22(2), 191–235. https://doi.org/10.1177/053901883022002003

Capra, R., Marchionini, G., Velasco-Martin, J., & Muller, K. (2010). Tools-at-hand and learning in multi-session, collaborative search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 951–960. https://doi.org/10.1145/1753326.1753468

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*. Morgan Kaufmann Publishers.

Case, D. O., & Given, L. M. (2016). *Looking for information: A survey of research on information seeking, needs, and behavior* (Fourth edition). Emerald.

Cho, J., & Lee, S. (2016). International Research International Students' Proactive Behaviors in the United States: Effects of Information-Seeking Behaviors on School Life. *Journal of College Student Development*, *57*(5), 590–603. https://doi.org/10.1353/csd.2016.0063

Citation Gecko. (n.d.). Retrieved February 2, 2022, from https://www.citationgecko.com/

CitNetExplorer. (n.d.). CitNetExplorer. Retrieved February 2, 2022, from https://www.citnetexplorer.nl//

Coats, S. (2020). Comparing Word Frequencies and Lexical Diversity with the ZipfExplorer Tool. In S. Reinsone, I. Skadiņa, A. Baklāne, & J. Daugavieti (Eds.), *Proceedings of the 5th Digital Humanities in the Nordic Countries Conference* (pp. 219–225).

Connected Papers. (n.d.). Retrieved February 2, 2022, from https://www.connectedpapers.com/

Cronin, B., & Sugimoto, C. R. (Eds.). (2014). *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*. The MIT Press.

Day, R. E. (2014). "The Data – It Is Me!" ("Les données–c'est Moi!"). In B. Cronin & C. R. Sugimoto (Eds.), *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (pp. 67–84). The MIT Press.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the Science citation index to cybermetrics*. Scarecrow Press.

Demelo, J., & Sedig, K. (2021). Forming Cognitive Maps of Ontologies Using Interactive Visualizations. *Multimodal Technologies and Interaction*, *5*(1), 2. https://doi.org/10.3390/mti5010002

Dervin, B. (1998). Sense-making theory and practice: An overview of user interests in knowledge seeking and use. *Journal of Knowledge Management*, *2*(2), 36–46. https://doi.org/10.1108/13673279810249369

Dervin, B. (1999). On studying information seeking methodologically: The implications of connecting metatheory to method. *Inf. Process. Manag.* https://doi.org/10.1016/S0306-4573(99)00023-0

di Sciascio, C., Sabol, V., & Veas, E. E. (2016). Rank As You Go: User-Driven Exploration of Search Results. *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 118–129. https://doi.org/10.1145/2856767.2856797

Dillon, A. (2000). Spatial-semantics: How users derive shape from information space. *Journal of the American Society for Information Science*, *51*(6), 521–528. https://doi.org/10.1002/(SICI)1097-4571(2000)51:6<521::AID-ASI4>3.0.CO;2-5

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, *37*(6), 817–842. https://doi.org/10.1016/S0306-4573(00)00051-0

Du, J. T., & Evans, N. (2011). Academic Users' Information Searching on Research Topics: Characteristics of Research Tasks and Search Strategies. *The Journal of Academic Librarianship*, *37*(4), 299–306. https://doi.org/10.1016/j.acalib.2011.04.003

Federico, P., Heimerl, F., Koch, S., & Miksch, S. (2017). A Survey on Visual Approaches for Analyzing Scientific Literature and Patents. *IEEE Transactions on Visualization and Computer Graphics*, *23*(9), 2179–2198. https://doi.org/10.1109/TVCG.2016.2610422

Freksa, C. (1999). Spatial Aspects of Task-Specific Wayfinding Maps. In J. S. Gero & B. Tversky (Eds.), *Visual and Spatial Reasoning in Design* (Vols. 15–32). University of Sydney. https://www.cosy.informatik.uni-bremen.de/spp/SPP_onlines/ProjektQ/Freksa1999.pdf

González-Valiente, C. L., Costas, R., Noyons, E., Steinerová, J., & Šušol, J. (2021). Terminological (di) Similarities between Information Management and Knowledge Management: A Term Co-Occurrence Analysis. *Mobile Networks and Applications*, *26*(1), 336–346. https://doi.org/10.1007/s11036-020-01643-y

Greyson, D. (2019). The Social Informatics of Ignorance. *Journal of the Association for Information Science and Technology*, *70*(4), 412–415. https://doi.org/10.1002/asi.24143

Greyson, D., O'Brien, H., & Shankar, S. (2020). Visual analysis of information world maps: An exploration of four methods. *Journal of Information Science*, *46*(3), 361–377. https://doi.org/10.1177/0165551519837174

Hartel, J., Noone, R., Oh, C., Power, S., Danzanov, P., & Kelly, B. (2017). The iSquare protocol. *Qualitative Research*, *18*(4), 433–450. https://doi.org/10.1177/1468794117722193

He, J., Ping, Q., Lou, W., & Chen, C. (2019). PaperPoles: Facilitating adaptive visual exploration of scientific publications by citation links. *Journal of the Association for Information Science and Technology*, *70*(8), 843–857. https://doi.org/10.1002/asi.24171

Hienert, D., & Lusky, M. (2018). Where Do All These Search Terms Come From? - Two Experiments in Domain-Specific Search. *ArXiv:1809.02407 [Cs]*. http://arxiv.org/abs/1809.02407

Hoeber, O., & Shukla, S. (2022). A study of visually linked keywords to support exploratory browsing in academic search. *Journal of the Association for Information Science and Technology*, *n/a*(n/a), 1–21. https://doi.org/10.1002/asi.24623

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE*, *9*(6), e98679. https://doi.org/10.1371/journal.pone.0098679

Jiang, Z., & Fitzgerald, S. (2019). Promoting Institutional Repositories via Visualizations: A Change-Point Study. *New Review of Academic Librarianship*, *25*, 1–18. https://doi.org/10.1080/13614533.2018.1547775

Kangasrääsiö, A., Chen, Y., Glowacka, D., & Kaski, S. (2016). Dealing with Concept Drift in Exploratory Search: An Interactive Bayesian Approach. *Companion Publication of the 21st International Conference on Intelligent User Interfaces*, 62–66. https://doi.org/10.1145/2876456.2879487

Kangasrääsiö, A., Głowacka, D., Ruotsalo, T., Peltonen, J., Eugster, M. J. A., Konyushkova, K., Athukorala, K., Kosunen, I., Reijonen, A., Myllymäki, P., Jacucci, G., & Kaski, S. (2014). Interactive Visualization of Search Intent for Exploratory Information Retrieval. *ICML '14 Workshop: Crowdsourcing and Human Computing*, 5.

Kessler, M. M. (1963). Bibliographic Coupling Between Scientific Papers. *American Documentation*, *14*(1), 10–25.

Kim, M. C., Zhu, Y., & Chen, C. (2016). How are they different? A quantitative domain comparison of information visualization and data visualization (2000–2014). *Scientometrics*, *107*(1), 123–165. https://doi.org/10.1007/s11192-015-1830-0

Kitzie, V., Wagner, T., & Vera, A. N. (2021). Discursive power and resistance in the information world maps of lesbian, gay, bisexual, transgender, queer, intersex and asexual community leaders. *Journal of Documentation*, *77*(3), 638–662. https://doi.org/10.1108/JD-08-2020-0138

Klavans, R., & Boyack, K. W. (2017). Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998. https://doi.org/10.1002/asi.23734

Kosslyn, S. M. (2006). *Graph Design for the Eye and Mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195311846.001.0001

Kuhlthau, C. C. (1991). Inside the Search Process: Information Seeking from the User's Perspective. *Journal of the American Society for Information Science*, *42*(5), 361–371. https://doi.org/10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-#

Kuhlthau, C. C. (1993). A Principle of Uncertainty for Information Seeking. *Journal of Documentation*, *49*(4), 339–355. https://doi.org/10.1108/eb026918

Kuhlthau, C. C. (1999a). The role of experience in the information search process of an early career information worker: Perceptions of uncertainty, complexity, construction, and sources. *Journal of the American Society for Information Science*, *50*(5), 399–412. https://doi.org/10.1002/(SICI)1097-4571(1999)50:5<399::AID-ASI3>3.0.CO;2-L

Kuhlthau, C. C. (1999b). Accommodating the user's information search process: Challenges for information retrieval system designers. *American Society for Information Science. Bulletin of the American Society for Information Science*, *25*(3), 12–16.

Kuhlthau, C. C. (2004). *Seeking meaning: A process approach to library and information services* (2nd ed). Libraries Unlimited.

Kuhlthau, C. C. (2008). Reflections on the development of the model of the information search process (ISP): Excerpts from the Lazerow Lecture, University of Kentucky, April 2, 2007. *Bulletin of the American Society for Information Science and Technology*, *33*(5), 32–37. https://doi.org/10.1002/bult.2007.1720330511

Kuhlthau, C. C., Case, D. O., Dervin, B., Bates, M. J., Cole, C., & Fisher, K. E. (2012). Crossing the divide: Putting information seeking research and theory into computer science practice to make information search systems and services more effective for the user. *Proceedings of the American Society for Information Science and Technology*, *49*(1), 1–4. https://doi.org/10.1002/meet.14504901078

Larsen, B., & Ingwersen, P. (2005). Cognitive Overlaps along the Polyrepresentation Continuum. In A. Spink & C. Cole (Eds.), *New directions in cognitive information retrieval* (pp. 43–60). Springer.

Leydesdorff, L., & Nerghes, A. (2017). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N < 1,000). *Journal of the Association for Information Science and Technology*, *68*(4), 1024–1035. https://doi.org/10.1002/asi.23740

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and Technology*, *57*(12), 1616–1628. https://doi.org/10.1002/asi.20335

Litmaps. (n.d.). Retrieved February 2, 2022, from https://www.litmaps.co/

Liu, Z., Sun, P., & Chen, Y. (2009). *Structured Search Result Differentiation*. VLDB. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151.8485&rep=rep1&type=pdf

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, *49*(4), 41–46. https://doi.org/10.1145/1121949.1121979

Mayr, P. (2016). How do practitioners, PhD students and postdocs in the social sciences assess topic-specific recommendations? *BIRNDL 2016 Joint Workshop on Bibliometric-Enhanced Information Retrieval and NLP for Digital Libraries*, *1610*, 84–92. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-47881-1

Mayr, P. (2020). *Greeting note for the 10th BIR workshop*. [Video] YouTube. https://www.youtube.com/watch?v=kNPVZZ7Mq0M

Mayr, P. (2021, April 14). *Openness, transparency, and inclusivity in science: What does it mean for IR?* [Video] YouTube. https://www.youtube.com/watch?v=0sOAyiPeThs

Mayr, P., & Scharnhorst, A. (2015). Scientometrics and information retrieval: Weak-links revitalized. *Scientometrics*, *102*(3), 2193–2199. https://doi.org/10.1007/s11192-014-1484-3

McCay-Peet, L., & Toms, E. G. (2018). *Researching serendipity in digital information environments*. Morgan & Claypool Publishers.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Munzner, T., & Maguire, E. (2015). *Visualization analysis & design* (1st ed.). CRC Press. https://doi-org.ezproxy.library.dal.ca/10.1201/b17511

Nepusz, T., Zanini, F., Horvát, S., & Traag, V. (2003). Igraph (0.8.x) [Python]. https://igraph.org/python/

Open Knowledge Maps. (n.d.). Open Knowledge Maps. Retrieved February 11, 2022, from https://openknowledgemaps.org/

Orduna-Malea, E., Martín-Martín, A., & Delgado López-Cózar, E. (2017). Google Scholar como una fuente de evaluación científica: Una revisión bibliográfica sobre errores de la base de datos. *Revista Española de Documentación Científica*, *40*(4), 185. https://doi.org/10.3989/redc.2017.4.1500

Paperscape. (n.d.). Paperscape. Retrieved February 2, 2022, from https://paperscape.org/

Rip, A., & Courtial, J.-P. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, *6*(6), 381–400. https://doi.org/10.1007/BF02025827

Scharnhorst, A. (2001). Constructing Knowledge Landscapes Within the Framework of Geometrically Oriented Evolutionary Theories. In M. Matthies, H. Malchow, & J. Kriz (Eds.), *Integrative Systems Approaches to Natural and Social Dynamics: Systems Science 2000* (pp. 505–515). Springer. https://doi.org/10.1007/978-3-642-56585-4_32

Scite. (n.d.). Scite.Ai. Retrieved February 2, 2022, from https://scite.ai

Shiffrin, R., & Börner, K. (2004). Mapping Knowledge Domains. *Proceedings of the National Academy of Sciences*, *101*(suppl. 1), 5183–5185. https://doi.org/doi\10.1073/pnas.0307852100

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In B. B. Bederson & B. Shneiderman (Eds.), *The Craft of Information*

Visualization (pp. 364–371). Morgan Kaufmann. https://doi.org/10.1016/B978-155860915-0/50046-9

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th Edition). Pearson. https://nsuworks.nova.edu/gscis_facbooks/18

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. https://doi.org/10.1002/asi.4630240406

Smith, L. C., & Wong, M. A. (Eds.). (2016). *Reference and information services: An introduction* (Fifth edition). Libraries Unlimited.

Spink, A., & Cole, C. (2006). Human information behavior: Integrating diverse approaches and information use. *Journal of the American Society for Information Science and Technology*, *57*(1), 25–35. https://doi.org/10.1002/asi.20249

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring Cognitive Load. In J. Sweller, P. Ayres, & S. Kalyuga (Eds.), *Cognitive Load Theory* (pp. 71–85). Springer. https://doi.org/10.1007/978-1-4419-8126-4_6

Traag. (2018). *Leidenalg* [Software]. https://github.com/vtraag/leidenalg

Traag, Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *ArXiv:1109.2058 [Cs]*. http://arxiv.org/abs/1109.2058

Verberne, S., Sappelli, M., Hiemstra, D., & Kraaij, W. (2016). Evaluation and analysis of term scoring methods for term extraction. *Information Retrieval Journal*, *19*(5), 510–545. https://doi.org/10.1007/s10791-016-9286-2

White, R. W., & Roth, R. A. (2009). *Exploratory search: Beyond the query-response paradigm*. Morgan & Claypool.

Wu, I.-C., & Vakkari, P. (2018a). Effects of subject-oriented visualization tools on search by novices and intermediates. *Journal of the Association for Information Science and Technology*, *69*(12), 1428–1445. https://doi.org/10.1002/asi.24070

Wu, I.-C., & Vakkari, P. (2018b). Effects of subject-oriented visualization tools on search by novices and intermediates. *Journal of the Association for Information Science and Technology*, *69*(12), 1428–1445. https://doi.org/10.1002/asi.24070

Zhang, J., Wolfram, D., Wang, P., Hong, Y., & Gillis, R. (2008). Visualization of health-subject analysis based on query term co-occurrences. *Journal of the American Society for Information Science and Technology*, *59*(12), 1933–1947. https://doi.org/10.1002/asi.20911

Zhang, J., Yu, Q., Zheng, F., Long, C., Lu, Z., & Duan, Z. (2016). Comparing keywords plus of WOS and author keywords: A case study of patient adherence research. *Journal of the Association for Information Science and Technology*, *67*(4), 967–972. https://doi.org/10.1002/asi.23437