

Evaluating Collaborative Search for a Learning-Oriented Search Task

Master's Thesis

Sindunuraga Rikarno Putra

Evaluating Collaborative Search for a Learning-Oriented Search Task

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Sindunuraga Rikarno Putra
born in Surakarta, Indonesia



Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
<http://wis.ewi.tudelft.nl>

Evaluating Collaborative Search for a Learning-Oriented Search Task

Author: Sindunuraga Rikarno Putra
Student id: 4624467
Email: SindunuragaRikarnoPutra@student.tudelft.nl

Abstract

Web search has become a convenient option for seeking information related to learning, therefore understanding how to facilitate human learning through a search engine has the potential to improve the quality of informal education and online learning. One less understood aspect of *search as learning* is the effect of collaboration in a search session focused on learning. Given the benefits of collaboration during both searching and learning, this thesis aims to evaluate the benefits of collaborative search for a learning-oriented search task.

We have developed a collaborative search system—SearchX—which adopts essential features of collaborative search as found in the literature. In the design process, we focused on providing support for research needs (such as running crowd-sourced experiments and fast prototyping). We deployed SearchX in a crowd-sourced user study on collaborative and single-user search for a learning-oriented search task with a total of 76 participants. Our results show that workspace awareness helped users in understanding the search task at the early stages of the search session and in selecting useful documents. However, we found little evidence that collaboration brought significant changes to the way users search and to users' learning outcomes. We found though that a good rapport between collaborators resulted in higher learning gains, highlighting the importance of social presence for online learning. Finally, we found indications that having a less difficult search topic results in less time spent on understanding search results and more exploration of the topic domain.

Thesis Committee:

Chair: Dr. Claudia Hauff, Faculty EEMCS, TUDelft (*supervisor*)
Committee member: Prof. Geert-Jan Houben, Faculty EEMCS, TUDelft
Committee member: Dr. Julian Urbano Merino, Faculty EEMCS, TUDelft
Committee member: Felipe Moraes, Faculty EEMCS, TUDelft (*supervisor*)

Preface

This document is a culmination of my two years as a computer science student at the Delft University of Technology. Throughout working on my thesis for almost a year, I was able to grow both scientifically and personally. This work would not be possible without the advice and support of the people that I would like to acknowledge here.

First and foremost, I would like to express my sincerest gratitude to my primary supervisor, **Dr. Claudia Hauff**, for facilitating my interest in a thesis topic related to learning and education. Working with her throughout the course of my thesis has given me a comprehensive experience in conducting computer science research, which for me is an invaluable learning experience. Without her guidance and feedback, I would not have gained as much insight and my research would not have taken shape.

Furthermore, I am deeply grateful to my PhD supervisor, **Felipe Moraes** for his time, advice, and patience. His dedication towards sharing his knowledge and his supportive attitude have motivated me throughout the course of my thesis.

I would like to thank **Lembaga Pengelola Dana Pendidikan (LPDP)** for providing me with a scholarship which made it possible to study in the Netherlands. As a token of my gratitude, I am deeply motivated to utilise the knowledge I have gained during my studies to improve the quality of education in Indonesia.

Last but not least, I would like to thank my friends and family for their support throughout the course of my studies. My partner in life, **Eashva Nazora**, who has accompanied me through my joy and struggles. My lovely daughter, **Sarah Estella Rasyad**, who colour my world and erase away my stress and fatigue. My dear parents, **Karno** and **Sri Hartini**, who always support me and pray for my success. Also thanks to my fellow Indonesian comrades in Computer Science: **Helmi, Andre, Romi, mas Reza, Gilang** and **Ulin**; for the great moments we shared during the past two years, and for the support in my studies and projects. Finally, big thanks to all members of **Keluarga Muslim Delft (KMD)** and **Perhimpunan Pelajar Indonesia Delft (PPI Delft)**, for becoming my second family and making Delft feel like home.

Sindunuraga Rikarno Putra
Schiedam, the Netherlands
August 24, 2018

Contents

Preface	iii
Contents	v
List of Figures	vii
1 Introduction	1
1.1 Learning During Search	2
1.2 Collaboration During Search	3
1.3 Collaboration During Learning	4
1.4 Research Objective	5
1.5 Approach	5
1.6 Scientific Contribution	6
1.7 Outline	6
1.8 Publications	7
2 Related Work	9
2.1 Exploratory Search	9
2.2 Search as Learning	10
2.3 Collaboration in Online Learning	11
2.4 Collaborative Search	13
2.5 Collaborative Search Systems	15
3 SearchX	21
3.1 Architecture	22
3.2 Supporting Collaboration	24
3.3 Development	28
4 Research Design	31
4.1 Learning Task	31
4.2 Study Setup	34
4.3 Crowd-work Deployment	35
4.4 Study Participants	37
4.5 Evaluation Metrics	37

5	Results	45
5.1	Participants' Collaboration Experience	45
5.2	Search Behaviour	49
5.3	Learning Outcome	54
6	Discussions and Conclusions	61
6.1	Discussions	61
6.2	Limitations	63
6.3	Future Work	64
6.4	Conclusion	65
	Bibliography	67
A	Deployed Study	77
A.1	Study Briefing	77
A.2	Registration	79
A.3	Pre-test	80
A.4	Search Session	82
A.5	Post-test	83
B	Essay Annotation	87
B.1	Quality of Facts (<i>D-Qual</i>)	87
B.2	Fact and Statement Counting (<i>F-Fact</i> and <i>F-State</i>)	89

List of Figures

1.1	Information search process [101].	3
2.1	The Community of Inquiry framework [28].	12
2.2	Collaborative search design space.	17
3.1	SearchX architecture overview.	22
3.2	SearchX collaborative search interface. [A] Document metadata, [B] shared query history, [C] shared bookmarks, [D] chat, [E] document rating, [F] document comments.	25
3.3	Colour coding of the shared query history.	26
3.4	Guided interface walk-through	27
4.1	Task description for all conditions. The <u>underlinedgreen</u> phrases were only added in the CSE condition.	34
4.2	The task bar containing the timer and the task description.	35
4.3	Overview of the study setup. Two search conditions are evaluated: single user and (pairwise) collaborative search.	36
5.1	Collaborative feature ratings (5-point Likert scale) by CSE participants.	46
5.2	Comparison of search actions (mean value) for each of the five search stages. The stages are created by splitting a participant's search session into five equal sections. As the standard deviations are large (because of the small sample size), we do not display the error bars.	50
5.3	Keyword complexity over each search stages.	57
5.4	Document complexity over each search stages.	57
5.5	Connection between chat actions and realised potential learning. Each point is a participant in CSE.	59
A.1	Study terms and requirements in the study briefing for all conditions. Both The prolific web page for our study and the welcome page in SearchX contains this briefing. The green phrases were only added in the CSE condition.	77

A.2	Study description in the study briefing for all conditions. The green phrases were only added in the CSE condition. The <u>underlined</u> phrases were not present in the CSE condition.	78
A.3	Registration form.	79
A.4	Vocabulary assessment for the first topic out of four pre-test topics. Only the first three questions (out of ten) are shown.	80
A.5	Waiting page after the pre-test for the CSE condition. The participants are notified when their partner has started the pretest.	80
A.6	Additional questionnaire related to collaboration for the CSE condition. The questions are utilised to prime participants into a mood for collaboration.	81
A.7	Interface guide at the beginning of the search session.	82
A.8	The deployed search interface for the CSE condition. The SE condition has a similar interface but without colour coding, group chat and the query history.	82
A.9	Vocabulary assessment for the post-test. Only the first three questions (out of ten) are shown.	83
A.10	Study completion message. Participants can click on the link to confirm their completion and receive their payment.	83
A.11	Written assessment and study feedback for the post-test.	84
A.12	Feedback on the collaboration experience for the CSE condition.	85
B.1	Question structure for scoring the quality of facts.	87
B.2	Instructions for scoring the quality of facts.	88
B.3	Question structure for counting the number of facts and statements.	89
B.4	Instructions for counting the number of facts and statements.	90

Chapter 1

Introduction

Information seeking—which refers to the conscious effort of acquiring information in response to a need or gap in knowledge—is a typical and essential human behaviour, whether done individually or collaboratively [27, 57]. The digital age has brought changes to multiple aspects of human life, including the way people seek information and knowledge. Previously, the options present for information seeking were mostly limited to large silos of print media in the form of libraries and, to a lesser extent, personal mediums of curated information such as the newspaper or an encyclopedia. Library science—which looks into effective management of printed information resource—can be considered as the roots of modern information retrieval [53]. The subsequent development and adoption of the internet have resulted in a spike of publicly available information sources since it is now possible for anyone to publish content on the Web. To illustrate, Youtube receives an average of 5 hours of video content every second¹, whereas Wikipedia receives 10 updates to existing articles every second and an average of around 600 new articles every day across all languages². This rapid growth of content paired with advances in information retrieval (IR) technologies has made the Web an accessible source of information for anyone with an internet connection. Personal computers connected to the internet have now taken over the role of libraries as a primary source of information, with everyone from researchers to children relying on Web search to fulfil their daily information needs.

The adoption of the internet has also increased the quantity and quality of publicly available learning materials. User-contributed online wikis have been shown to provide information quality comparable to those in centrally-controlled Websites and textbooks [77]. With the recent rise of Massive Open Online Courses (MOOCs), university grade teaching materials are now also conveniently available to the public. This increase in high-quality contents has made the Web a vital information source for learning. Researchers have observed that learners of all ages are increasingly turning towards search engines to support their learning needs [79, 67]. Additionally, a log-based study of commercial Web search engines in 2012 [8] has shown that people have been using the Web for learning purposes, with nearly 30% of Web search queries showing an intent to learn, ranging from factual lookup to more complex information seeking such as exploring a topic in depth.

¹<https://www.youtube.com/yt/press/statistics.html> (Accessed on July 2018)

²<https://en.wikipedia.org/wiki/Wikipedia:Statistics> (Accessed on July 2018)

Despite this, current search engines are not designed to fulfil learning needs—the underlying goal for most IR algorithms is to maximise relevance and precision, which does not always result in better learning gains [16]. Additionally, relevance is typically optimised for a single query, whereas learning searches (i.e. searching for a learning-oriented task) may require relevance that spans multiple query iterations [58]. It is understandable that not all search engines need to be optimised for human learning, as not all search tasks are learning-oriented and many search needs can already be satisfied with a system designed to optimise general relevance. However, as people are starting to depend on Web search for learning-oriented tasks, there is a need to explore how current search engines can be adapted to better support learning.

In 2012, IR researchers started to recognise the importance of learning gains as a search outcome [3]. This research agenda was highlighted again in a 2014 Dagstuhl seminar [1] where researchers concluded that the connection between searching and learning warrants further exploration. The field started to mature in 2016 when the SIGIR (Special Interest Group on Information Retrieval) conference hosted the very first workshop on search as learning [38]. Since then, increasingly more researchers are putting their attention on developing this emerging field [101, 78, 23, 17]. This thesis aims at investigating one less understood aspect of search as learning, which is *the effect of collaboration during the search process towards learning outcomes*. Complex search tasks have been shown to benefit from collaboration [27], and many learning tasks are indeed complex.

1.1 Learning During Search

Initially, search engines were designed as a means to retrieve specific pieces of information from a large corpus, similar to database lookup. The adoption of the internet made search systems essential since an efficient method of browsing the Web was needed; however, the underlying system still functions on the same goal of efficient retrieval, requiring the user to know their information need and express it in a detailed query. As Web search engines have now become ubiquitous with more than 50% of the world population having access to the internet³, people expect it to serve other types of services beyond lookup. There is a trend towards more active engagement during the search process because often a person's information need can be vague or incomplete, therefore requiring a trial and error process of querying and browsing the results space in order to refine their information need iteratively. Marchionini [58] refers to this blend of searching and browsing as *exploratory search*. As opposed to only recalling pieces of information such as in lookup search, exploratory search is a more intricate process which involves discovery and learning while exploring the result set.

Learning is a complex topic involving research from various disciplines, including pedagogy, educational psychology, cognitive science, neuroscience and information science. The definition of learning itself varies greatly depending on the point of view used [45]. A generic definition of learning from the English Oxford Dictionary⁴ is "the acquisition of knowledge or skills through study, experience, or being taught". Learning thus happens when a person goes through a specific process and from that

³<https://www.internetworldstats.com/stats.htm> (Accessed on July 2018)

⁴<https://www.oxforddictionaries.com/> (Accessed on July 2018)

integrates new information into their current understanding. This point of view can be explained more clearly using the construction-integration theory by Kintsch [47], which states that humans understand a topic by first constructing an approximate mental representation of it from information sources, and then integrating it with their existing mental representations to create a coherent whole. Information that has been assimilated to a person's mental model is thus what we refer to as knowledge.

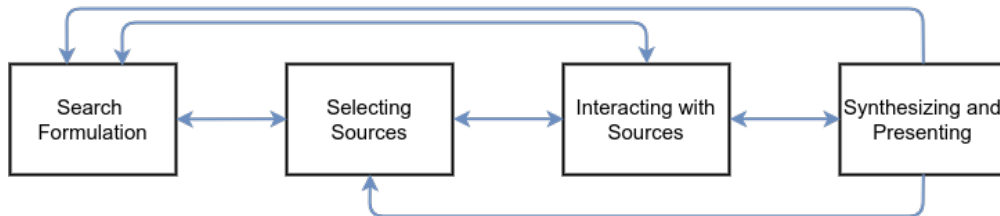


Figure 1.1: Information search process [101].

A similar process can be observed during an exploratory search session where a user's understanding regarding the search task and goal is continuously updated. Vakkari [101] described a generic model for information searching (Figure 1.1) which begins with formulating a search query that reflects the user's understanding of their own information need. From the result set, the user select documents that will satisfy their current understanding of the information need. The user then interacts with the selected documents in a sense-making process to familiarise themselves with the new information received. Many times, this results in them updating their information need according to their newly acquired understanding. If their information need changes a lot, the user will go back and refine the search terms. For a learning-oriented search task, the user will additionally update their understanding of the task domain. Table 1.1 provides two examples of queries issued during a learning-oriented search session, illustrating the process of gaining new knowledge over time as displayed by the use of new search keywords.

1.2 Collaboration During Search

Web search is generally seen as a solitary activity, as most mainstream search technologies are designed for single-user search sessions. However, for a sufficiently complex search task (i.e. consisting of multiple aspects or steps [40]), collaboration during the information seeking process is beneficial as it enables users to complement each other's skills to make up for an individual's lack of knowledge [100, 27]. Two surveys by Morris [63, 64] conducted in 2006 and 2012 has shown that collaborative search (CSE) is an increasingly common activity, albeit through the use of ad hoc solutions such as email and instant messaging. There was a significant increase in the number of people who collaboratively search on a regular basis, from 0.9% in 2006 to 11% in 2012. Additionally, there is a change in collaboration practices towards increasingly larger group sizes: there was an increase in people who collaborated in groups of three or more, from 19.3% in 2006 to 68.8% in 2012; with a significant increase in group sizes of more than four, from none at all in 2006 to 22.1% of all respondents in 2012. These are indications that there is a need to better support collaborative search systems.

Table 1.1: Two search sessions with their respective sequence of queries, issued by participants during a 20 minutes learning-oriented search session in our study. Underlined text indicate new query terms.

Learning Topic	Search Queries
Radioactivity	<u>what is radioactivity</u> → <u>types of</u> radioactivity → <u>how does</u> radioactivity <u>occur</u> → radioactivity <u>resources online</u> → <u>uses for radiation</u> → <u>university</u> resources radioactivity → radioactivity <u>explained</u> → <u>electromagnetic</u> radiation → radioactivity <u>harmful</u> → <u>ionising and non ionising</u> radiation → radioactivity <u>studies physics journals</u> → radioactivity <u>exposure</u> → <u>what creates gamma waves</u> → <u>xray</u> radiation <u>type</u> → <u>dangers</u> of radioactivity → uses of radioactivity
Industrial Biotechnology	industrial biotechnology → bioplastics → bioplatics biotechnology → bioplatics <u>production enzymes</u> → <u>polylactic acid</u> bioplastic → industrial biotechnology → <u>biofuels</u> → <u>biocatalysts</u> → industrial enzymes → <u>biological large scale process</u> → biological large scale <u>processes</u> → industrial biotechnology → enzyme <u>engineering</u> → <u>enzymes in</u> biotechnology → industrial biotechnology <u>worth</u> → biotechnology <u>industry</u> → biotechnology industry <u>market</u> → biotechnology industry <u>size</u>

The increasing use of CSE has been picked up by the research community, where CSE has now been an active area of research for many years. Workshops that explicitly focus on collaborative search—and more generally information seeking—have started to appear in 2008 [73] and continue to do so to this day [7].

1.3 Collaboration During Learning

Collaboration is often a natural choice when we face a complex task, but it is not always necessary [19]. Learning is a task which has been consistently shown to benefit from cooperation and collaboration according to past meta-studies [42, 43, 93]. Collaborative learning can refer to any learning session in which learners work in groups to complete a common goal and maximise each others' learning gains. Although Dillenbourg [20] differentiates between collaborative and cooperative learning based on the level of interdependence between collaborators, both types of learning rely on the same premise of two or more individuals working together to learn. Apart from the benefits of social interaction, collaborative learning also offers additional learning opportunities through the interaction between learners. Student-student interaction has long been suggested to be an essential factor in learning success [42], and even more so for online learning where there is a lack of social presence [61, 94].

1.4 Research Objective

Given the similarity between the process of searching and learning, and people's dependence on Web search nowadays, we believe that the development of a search system that optimises learning gains can help improve the quality of informal education and online learning. However, understanding of the connection between searching and learning is still in its infancy—researchers have just started to explore this through the emerging research field of search as learning. One less explored aspect in this field is investigating how collaboration affects learning-oriented search tasks. Given the positive benefits of collaboration during both searching and learning, it raises a question: does the benefit of collaboration hold true for a learning-oriented search task? White and Roth [103] have argued that the exploratory nature of learning searches makes collaboration a natural choice during learning, but to our knowledge, no research so far has looked into the concrete benefits of collaboration in this context.

In response to this, we investigate the effects of collaboration during learning-oriented search sessions, specifically in Web search. To do so, we implemented a collaborative search system, ran a crowd-sourced user study for both collaborative search and single-user search, and then analysed the collected logs. We investigate the following research questions in this thesis:

RQ1 *How does the design and implementation of a collaborative search system based on modern Web technologies look like?*

As the available systems for collaborative search are not suitable for our use case (elaborated in Section 2.5.3), our first step in this thesis is to design a collaborative search system built on modern Web technologies, and therefore more feasible for a crowd-sourced user study. We analysed how prior systems accommodate collaboration during search, and then adopted their implemented features to our use, resulting in the SearchX collaborative search system.

RQ2 *How does pair-wise collaborative search compare to single-user search for a learning task in terms of learner's search behaviour?*

Collaboration has been shown to promote more diverse queries and less document overlap for information gathering tasks [44, 87]. We investigated whether introducing collaboration during a learning-oriented search task also results in a similar effect on search behaviour.

RQ3 *How does pair-wise collaborative search compare to single-user search for a learning task in terms of learner's knowledge gain?*

We looked into the learning outcomes of both pair-wise collaborative search and single-user search, and evaluated the effectiveness of each method in relation to the achieved learning outcome.

1.5 Approach

In order to investigate learning during search, we conducted a user study in which participants were asked to utilise SearchX for a learning task. Specifically, we assigned an exploratory search task in which participants were instructed to learn a topic in which

they were unfamiliar. The list of topics was adopted from online video lectures as they have been carefully designed for online learning. We measured the learning outcome using a pre-test and post-test as it offers a precise method for calculating the learning gain.

We deployed the user study on a crowd-sourcing platform as it is both time and cost efficient, therefore allowing us to accommodate a higher number of study participants compared to a lab-based study. Participants were assigned either a collaborative search session or a single-user search session. We then compared the two search conditions using user interaction logs collected by SearchX.

1.6 Scientific Contribution

With our work, we bring several contributions to the fields of search as learning and information retrieval.

1. We created a comprehensive analysis of prior collaborative search systems and described the design decisions taken in developing a system for collaborative search.
2. We introduced a new collaborative search system that is built using modern Web technologies. SearchX is the second only open-sourced system for collaborative Web search, and does not require additional user-side installation, making it suitable for large-scale studies. Furthermore, we put additional effort to make the system more accessible to other researchers as elaborated in Section 3.2.1. We also highlighted the challenges we encountered during the development of the system and its deployment on a crowd-sourcing platform.
3. We provide insight on how collaboration affects the outcomes of a learning-oriented search task. We found that for a difficult learning topic, the differences in the way participants search and in their achieved learning were not significant. However, from a more in-depth analysis, we found that collaboration did help participants in understanding the topic at the beginning of the session. Additionally, we found that good discourse between collaborators has a positive effect on the learning outcome.
4. We found indications that task difficulty has a significant effect on exploratory search behaviour. Participants receiving a less difficult topic were able to explore more of the topic domain compared to participants who received a topic that was difficult for them to comprehend.

1.7 Outline

The organisation of the thesis is as follows. In Chapter 2, we summarised literature related to our study and present key ideas we adopt in our work. Chapter 3 elaborates on the design decisions and challenges involved in developing a collaborative search system that is suitable for our study. In Chapter 4, we describe the setup of our user study and elaborate the deployment of our study on a crowd-sourcing platform, and

in Chapter 5 we present our analysis of the data we received from the user study. We conclude our results in Chapter 6 where we discuss the findings and limitations we observed in the results, as well as propose related future works.

1.8 Publications

During the completion of this thesis, three publications have been made based on the ongoing study. A list of the publications along with the author's contributions are listed below:

1. Sindunuraga Rikarno Putra, Felipe Moraes, and Claudia Hauff. SearchX: Empowering Collaborative Search Research. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2018.
A demonstration paper promoting SearchX as a tool for research. The author contributed to the development of the system and partially in the writing of the paper.
2. Sindunuraga Rikarno Putra, Killian Grashoff, Felipe Moraes, and Claudia Hauff. On the Development Of a Collaborative Search System. In *Proceedings of the 1st DESIRES (Design of Experimental Search & Information REtrieval Systems) Conference*. CEUR-WS, 2018.
A prototype paper describing the design decisions and implementation details of SearchX—similar to Chapter 3 in this thesis. The author contributed to the development of the system and the majority of the writing.
3. Felipe Moraes, Sindunuraga Rikarno Putra, and Claudia Hauff. Contrasting search as a learning activity with instructor-designed learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2018.
A full paper exploring how learning through search compares with instructor designed learning concerning learning outcomes. The author contributed to the development of the system and in conducting user studies for the collaborative search condition.

Chapter 2

Related Work

This thesis builds on research in the fields of (i) exploratory search, which is closely related to learning searches; (ii) search as learning, which studies the process of search with a learning intent; (iii) pedagogy, specifically on collaboration in online learning; and (iv) collaborative search, including the design of systems that support collaboration during search. This section elaborates on each of the fields mentioned above, highlighting how they relate to our work and how we expand upon prior findings.

2.1 Exploratory Search

Search activities are commonly divided into two broad categories: lookup search and exploratory search [58]. Lookup searches are basic kinds of search, where the information need is clear, and the search goal is discrete and well-structured. The majority of search engines are designed around the fast and accurate completion of lookup searches. Exploratory search, on the other hand, results from an open-ended or complex information need where there is a level of uncertainty, therefore requiring multiple query iterations and possibly multiple search sessions. White and Roth [103] define the exploratory search process as a combination of exploratory browsing and focused searching. Exploratory browsing is the more dominant activity and involves exploration of the search space with the goal of better defining the information need. Focused searching, on the other hand, requires a clear information need, and involves constant query refinements in order to extract information related to the search goal.

Learning searches are one particular type of exploratory search, involving cognitive processing and interpretation of the search results [58]. Heinström [41] analysed the search behaviour of master students working on their thesis and observed that in undertaking a predominantly learning-oriented task, students show characteristics of exploratory search—they undertake exploratory browsing when looking into new topics or to get an overview of a topic, and then switch to focused searching to fill in specific facts for a previously explored topic. Given this close relationship between learning and exploratory search, researchers have proposed the use of learning outcomes as a potential evaluation method for exploratory search [103, 1].

In supporting exploratory search, there are two main directions. The initial direction suggested by Marchionini [58] was implementing an interface that better supports exploratory search behaviours. Interface features that have been shown to have

a positive impact on exploratory search situations include: faceted search [52], features supporting multiple-session searches [35], automatic decomposition of complex queries [40], and interactive visualisations of the topic space [80, 49]. Another direction is to modify the retrieval algorithm to return a result set that is beneficial to exploration—mainly in the form of intrinsically diverse result sets which cover multiple topics related to the query [76, 16]. Recently, Athukorala et al. [5] investigated the difference between lookup search and exploratory search using measurable behaviours in an IR system (i.e. initial query length, the time spent on analysing the first result page, the scroll depth and task completion time), concluding that there are indeed measurable differences between the two search types. On a follow-up study, Athukorala et al. [6] leveraged their previous findings to develop a predictor that can classify the type of search session based on users' search behaviour, enabling the creation of a hybrid search system that can automatically adapt to the search task at hand.

Our work is in line with prior search behaviour observation studies such as [5] but aims at observing learning searches—which in recent years have been explored under the search as learning research field [23].

2.2 Search as Learning

Search as learning (SAL) is an emerging research field in the information retrieval, information science, and human-computer interaction communities which strives to better understand and support learning during the search process—notably through search engines. Researchers have started to recognise the importance of learning gain (i.e. the difference between the knowledge at the end and the start of the search session) as a search outcome in 2012 [3], and are now putting more attention into developing the field of SAL [78, 17]. The majority of research in SAL can be categorised into three research directions: (i) understanding the connection between learning and searching; (ii) investigating methods to measure and model learning during search; (iii) exploring ways to facilitate learning during search. In practice, the research directions often overlap, as the emerging nature of this field means that not many standards and benchmarks are agreed upon yet.

Several works have utilised measurable behavioural traces to characterise domain expertise. A typical method is to extract the logs of search sessions aimed at learning in order to derive session metrics (e.g. query complexity, diversity of domains on the SERP, document display time). White et al. [104] and Eickhoff et al. [22] utilised these metrics to investigate the connection between search behaviour and domain expertise. Zhang et al. [111] and Cole et al. [15] on the other hand used a similar method to analyse predictors of domain expertise—the former using log traces and the latter using eye tracking traces. The use of such data-driven methods have the advantage of being highly scalable (e.g. Eickhoff et al. [22] analysed more than 700K search sessions), but the resulting metrics are crude heuristics of learning gain.

Other works have opted to directly measure learning gain through the use of knowledge assessments (e.g. multiple-choice tests, writing a summary), which offer a precise way of measuring learning gain through comparing learning outcomes before and after the learning session. However, the quality of these methods is dependant on the assessment process, resulting in many researchers opting for a lab-based study where partici-

pants can be guided to complete the assessment correctly. Additionally, the choice and design of assessment matters: multiple choice questions if not appropriately designed tend only to measure recall instead of higher order learning [39], whereas open-ended questions and summaries are hard to quantify. Wilson and Wilson [105] proposed and explored methods to objectively measure summaries generated as a result of an exploratory search session, concluding that there is no "one-size-fits-all" approach to measure the learning outcomes of written summaries.

Not much research has been done on designing systems to facilitate learning during search. Most research on the interface side focuses on supporting the more generic exploratory search as elaborated in section 2.1. Collins-Thompson et al. [16] conducted a user study to investigate whether certain types of search results (single-query result sets, multiple-query result sets, and intrinsically-diversified result sets) are more beneficial for learning. They measured learning outcomes via manually assessed open-ended questions as well as self-reports and concluded that intrinsically diverse search results offer a considerable advantage to learning compared to the other two conditions. Additionally, they found that users' perceived learning correlates highly with their actual learning outcome. On the algorithm side, one recent work by Syed and Collins-Thompson [95] proposed a document ranking model optimised for learning (as opposed to relevance as standard ranking models) that was shown to improve learning outcomes for a *vocabulary learning task* (i.e. understanding domain-specific terms) in comparison with a commercial search engine. However, the setup of the study was rather artificial: participants were not given the ability to search and were only tasked with evaluating a fixed result set (produced by variants of the document ranker) from a predefined query—explicitly avoiding variance introduced by users' differing search behaviours. A follow-up study [96] analysed features of the returned document set associated with improved short-term and long-term learning outcomes. A different work by Yamamoto and Yamamoto [106] looked into the effect of query priming and found that suggesting specific query keywords (e.g. survey, comparison, evidence) can promote search behaviours which indicate critical thinking and higher-order learning.

Our work is similar to [16, 95, 106], which involves conducting a user study to investigate methods of supporting learning-related search tasks—specifically, we look into the effects of providing collaboration capabilities. Chi et al. [13] have created a similar study that explores the process of knowledge learning during collaborative search using indirect measures of knowledge (click and query complexity). They assume that infrequently accessed documents and query keywords indicate a higher level of complexity, which in turn indicates a higher level of achieved knowledge. In contrast to this, our work includes direct measurements of knowledge adopted from [95] to more precisely measure the achieved knowledge.

2.3 Collaboration in Online Learning

Generally, studies that investigated the benefits of collaborative work over individual work in education report improvements in academic achievements, quality of interpersonal interactions, and study retention [75]. There are a few explanations regarding the effect of collaboration towards learning performance. Cohen [14] suggests that the free exchange of ideas that happens during a collaborative learning session stimulates

conceptual understanding. Scardamalia and Bereiter [81] argue that the interactions introduced when collaborating promotes a more active form of learning which enables the co-construction of knowledge within the group of learners. Additionally, Gokhale [32] reported that collaborative learning resulted in better retention of knowledge, and is beneficial in enhancing critical thinking and problem-solving skills—which is essential for higher-order learning [30].

A precursor to research in online learning is the exploration of distance education, which aims to support education when face-to-face learning is not possible—mainly through replacing face-to-face interaction with a remote alternative such as through post or online. Moore [61] identifies three types of interaction during face-to-face learning that needs to be taken into account when designing distance learning: learner-resource interaction, learner-teacher interaction, and learner-learner interaction. This was further developed by Garrison et al. [30] into the *Community of Inquiry* framework which defines an online learning experience as the intersection of three main elements as can be seen in Figure 2.1. *Cognitive presence* is the primary goal of the framework and refers to a state where learners actively seek knowledge. *Social presence* refers to the establishment of personal and purposeful relationships within a group of learners. *Teaching presence* refers to the existence of structure and organisation in the learning experience, whether directly through instruction or indirectly through a well-designed course. Concerning the framework, a collaboration between students is the product of proper social and teaching presence, leading to effective discourse which results in improvements to the cognitive presence within the community of learners [28].

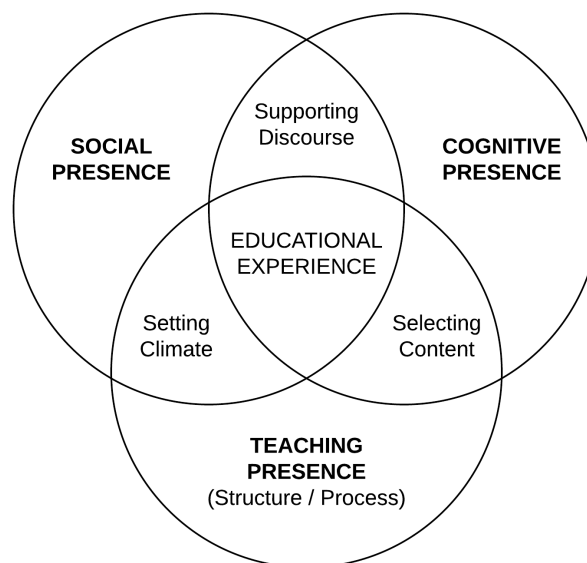


Figure 2.1: The Community of Inquiry framework [28].

Several studies have looked deeper into how each component of the community of inquiry framework affects the online learning experience—mostly focusing on university online courses within a long time-span. Picciano [71] looked into the relationship between social presence and learning performance, and found a strong correlation between the perceived quality and quantity of interaction with learning outcomes. The learning outcome of the written assessment was especially significant for learners who

demonstrate high levels of interaction. Garrison and Cleveland-Innes [29] compared students' learning outcome in four online courses that differ in the intensity of instructor presence and the level of interactions. The results indicate that the existence of social presence alone is not enough to produce meaningful discourse—the existence of interactions does not necessarily translate into better cognitive presence. A well-designed teaching presence is required to guide learners' towards a productive discourse properly. Akyol and Garrison [2] analysed the development of each type of presence over time in an online course, and found that each type of presence was distinguishable and that social presence did not have as much of an impact on learning outcome compared to both cognitive and teaching presence.

We adopt the community of inquiry framework to analyse online learning similar to [71, 29, 2], but using collaborative search as the medium. Curtis and Lawson [18] have shown that the choice of medium, as well as learners' familiarity with the platform, has a strong influence towards learners' interactions and social presence. In our interpretation, learning through collaborative search is an online learning experience which involves more social presence than teaching presence; therefore we focus mainly on the social presence aspect of collaborative search.

2.4 Collaborative Search

CSE is a subset of the more generic field of collaborative information seeking (CIS), which is an interdisciplinary domain drawing from research on the field of information science, information retrieval, human-computer interaction, and computer-supported collaborative work. Research on CIS aims at exploring and supporting the collaborative aspect of information tasks such as browsing, searching, filtering and navigating [27].

The most common approach in analysing the effects of collaboration on search involves testing different search conditions through user studies and using data collected from the search sessions for a comparative study. One research by Joho et al. [44] compared single user search (SE) and pair-wise CSE for an information gathering task (i.e. gathering as many relevant documents as possible) to analyse differences in search strategies and performance. Their results suggest that collaboration during the search process helps in diversifying the queries and avoiding duplicated efforts, but did not have a significant impact on search performance measured by precision and recall. Shah and González-Ibáñez [83] conducted a similar study that further divides the CSE condition into co-located and remote collaboration. Unlike [44], they did not use an existing corpus and instead let the study participants search the open Web. To compute search performance measures, they used the documents accessed by participants as the document space. The observed results were similar in that there were no significant differences in search performance between SE and CSE, but CSE produces a more diverse set of retrieved documents.

González-Ibáñez et al. [36] expanded on [83] by investigating the effect of different collaboration conditions (co-located, remote, and asynchronous) on search performance measured in four different aspects: communication, productivity, information synthesis, and cognitive load. Their results suggest that the different collaboration conditions resulted in no significant difference in sense-making, but did affect query

diversity: participants in the co-located condition demonstrated more similar queries, whereas participants in the asynchronous condition demonstrated the most diverse queries. Capra et al. [11] investigated in more detail what search behaviour can be observed during asynchronous CSE by utilising think-aloud data and screen recordings. They identified three common behaviours throughout the search sessions—individual, parallel, and divergent—but no general pattern was observed because of the small sample size. Chen et al. [12] looked into the effect of awareness mechanisms (e.g. query history) for asynchronous collaboration, observing that for an asynchronous setting, awareness resulted in less query diversity, but more unique documents and less duplicated efforts.

Several studies looked into the effect of the type of search task towards search behaviour in CSE. Yue et al. [108] compared search behaviour between two search task types (information gathering and decision making) for both SE and CSE, and found that decision making tasks such as travel planning involve more use of direct communication, whereas information gathering tasks such as academic research involve more individual work. A follow-up analysis [109] concluded that queries were most affected by different actions for each task type: bookmarks for the information gathering task, and chat for the decision making task. Chen et al. [12] on the other hand investigated the effects of task orientation (open-ended or recall-oriented) towards search behaviour, but observed no significant differences in search performance. The recall-oriented task though resulted in more unique documents and less document viewing time, which we believe to be because participants focused more on coverage rather than sense-making.

All the previously mentioned research was limited to pair-wise collaboration. Some prior works have opted to explore team sizes beyond two, although to our knowledge only triads (i.e. groups of three) were investigated. Tao and Tombros [98] specifically looked into the search behaviour of triads through an observational user study, and found that the larger team size promoted a structure-driven search strategy which involves creating a structure of the task domain and then having each collaborator focus on the different sections. Additionally, the existence of a shared document for information synthesis resulted in less time taken to finish the task and a higher knowledge increment, but less communication between collaborators. Shah et al. [87] proposed an evaluation framework to analyse exploratory search behaviour and used it to compare the search performance of different team sizes (single user, pairs, and triads) for an information gathering task. They found few significant differences in search performance when looking at the user level, but at a team level, CSE produced better performance than SE for nearly all their measured metrics, with triads achieving better performance than pairs. A follow-up study [88] looked into the same team sizes, but for a fact-finding task, which is non-divisible (i.e. difficult to divide into sub-tasks). The results indicate that an increased team size did not have a significant advantage on search effectiveness and efficiency. Increasing the team size though did have a positive effect on response precision (i.e. the accuracy of answers) but at the cost of longer task time.

In this thesis, we are interested in comparisons between SE and CSE. We have summarised the conclusions from prior studies in Table 2.1. In general, research has shown that for an information gathering task, collaboration has a beneficial effect on search behaviour, mainly in diversifying queries and reducing document overlap.

Table 2.1: Summary of comparisons between CSE and SE. Signs indicate significance of CSE: more (+), less (-), and non-significant(=).

Source	Search Task	Conclusions regarding CSE
[44]	Information gathering	(+) Distinct queries, Keywords (-) Overlapping documents (=) Performance (relevance)
[83]	Information gathering	(+) Query diversity, Useful documents, Unique documents (=) Cognitive load, Performance (relevance)
[87]	Information gathering	(+) Team-level metrics (=) User-level metrics
[88]	Fact finding	(+) Response precision (-) Performance (efficiency & effectiveness) (=) Coverage, Query diversity

However, for a fact-finding task where the division of labour is much more difficult, there were no significant differences in search behaviour while there was a decrease in search performance. We are therefore interested in investigating the effects of collaboration during search for a learning task—which is similar to an information gathering task but emphasises gaining knowledge from documents more than the number of covered documents. In doing so, we adopt the same comparative approach as prior studies on CSE which involves comparing data on different search conditions.

2.5 Collaborative Search Systems

Golovchinsky et al. [34] have characterised the collaborative aspect of online CIS along four dimensions: *intent* (explicit or implicit), *mediation* (user interface or algorithm), *concurrency* (synchronous or asynchronous), and *location* (remote or co-located). Morris [64] suggested two additional dimensions: *role* (symmetric or asymmetric) and *medium* (traditional or emerging devices). All six dimensions can be considered as the design space of CSE systems and is visualised in Figure 2.2¹.

In our interpretation, *CSE is scoped around collaboration of explicit intent, with the mediation and role dimension explored in designing collaboration in the system, and the other three dimensions (concurrency, location, medium) determining the domain of the system.*

In this thesis, we are interested in prior collaborative search systems proposed in the literature. In regards to the previous taxonomy of dimensions, we limit the domain to the most common Web collaboration practice—which according to a survey in 2012 refers to remote synchronous collaboration using traditional devices (laptops and desktops) [64]. Table 2.2 compare the features and availability of the prior systems we analysed and our proposed system—SearchX.

¹The diagram is adopted from a presentation by Morris in <https://www.youtube.com/watch?v=MdMzileZY3Y>.

Table 2.2: Feature comparison of existing remote collaborative search systems and SearchX (ordered by publication year of the first paper describing the system). A dash – indicates that this information is not available. Language and platform abbreviations: JS=JavaScript, BP=Browser Plugin, IE=Internet Explorer, FF=Firefox, GC=Google Chrome. Note that we only list programming languages listed in the respective papers (if no open-source code is available). † The Coagmento iOS app is only available in Apple’s US app store.

	Search Together [65]	CoSense [69]	Coagmento [85]	Querium [35]	ResultSpace [10]	CollabSearch [108]	CoZpace [54]	SearchX
Division of Labour								
Group Chat	✓		✓	✓		✓	✓	✓
Document Sharing	✓		✓	✓				
Sharing of Knowledge								
Bookmarking			✓					✓
Document Rating	✓	✓		✓	✓		✓	✓
Document Annotation	✓	✓	✓				✓	✓
Awareness								
Query History	✓	✓	✓	✓	✓	✓	✓	✓
Document History		✓	✓	✓	✓		✓	✓
Document Metadata	✓	✓	✓	✓	✓		✓	✓
Group Summary		✓		✓		✓	✓	✓
Colour Coding		✓				✓		✓
System Mediation								
	Split Search			Ranked Doc. History	Re-ranked Search Results			
Tool Availability								
Functioning	✗		✓					✓
Open Source	–		✓					✓
Last Update	2009		2018					2018
Language	–		PHP & JS	JS	PHP		JS	JS
Platform	BP (IE)		iOS†, Android	Web	Web	Web	Web	Web
			BP (FF, GC)					



Figure 2.2: Collaborative search design space.

2.5.1 Prior Systems

SearchTogether by Morris and Horvitz [65] was one of the first attempts at a system for collaboratively searching the Web. Based on a survey of collaboration practices during Web search, the design of the system focused on supporting awareness, division of labour, and persistence. They designed the system for synchronous search sessions, i.e. a search session in which collaborators work together on the task at the same time. From a user study involving the system, they found that awareness was the most vital aspect in supporting collaboration—with features intended for division of labour and persistence used by participants as a tool for awareness instead. Shah et al. [85] built upon the weaknesses of *SearchTogether* by introducing novel features to *Coagmento*: the ability to snip Web pages and to share Web pages for further discussion with collaborators. Kelly and Payne evaluated the longitudinal use of *Coagmento* in an everyday setting [46], and highlighted issues specific to real-world search use: time-criticality of search results and privacy issues on sharing collaborators’ data. Other concerns that they raised included sense-making of the collaboration product and the interaction cost of collaborative features. Paul and Morris [69] built *CoSense*, an extension for *SearchTogether* which aims to improve sense-making by providing additional views (i.e. interface layouts) that focuses on different aspects of the search session (e.g. browsing history, chat). The existence of elaborate views was shown to improve asynchronous search sessions as it allowed new collaborators to catch up with results from prior sessions more easily.

More recent systems were created to explore specific aspects of online collaboration. Golovchinsky et al. [35] designed *Querium* to better support collaboration in exploratory search processes, specifically through providing a shared document history that rank documents based on users’ relevance feedback. Capra et al. [10] designed *ResultsSpace* to study asynchronous collaborations (though synchronous collaboration is also possible), therefore features for direct communication such as a chat were not added. Yue et al. [108, 109, 110] designed *CollabSearch* with no advanced features for division of labour such as algorithmic mediation in order to investigate how users coordinate and collaborate in a synchronous search session. Kruajirayu et al. [50] explored the effectiveness of visual snippets for sense-making by introducing the

SnapBoard feature into the CoZpace system.

Researchers also designed experimental systems for more specific collaboration conditions. Amershi and Morris [4] developed CoSearch for a shared computer setting and offered features that enabled one interface to be used efficiently by multiple users. Golovchinsky et al. [33] designed Cerchiamo for collaboratively searching through videos and implemented different interfaces for two different roles: the *prospector* role focuses on breadth by scanning the search space for promising directions of exploration, whereas the *miner* role focuses on depth by digging deeper into the results for more comprehensive information. Morris et al. [66] designed WeSearch to facilitate co-located collaboration through a table-top interface where all collaborators can freely arrange search results together. Teevan et al. [99] on the other hand designed O-SNAP to explore collaboration using mobile devices by utilising different screen orientations. As these experimental systems were not designed for the same domain as SearchX, we did not include them in the upcoming analysis.

2.5.2 Collaboration Design

As mentioned earlier, the design of collaborative search systems mainly explores two dimensions of collaboration: mediation and role.

There are two main directions in developing mediation for CSE: *interface mediation* adapts the search interface towards a multi-user context, usually in the form of a shared workspace; *system mediation* directly mediates the collaboration process, mostly through re-ranking the documents [35, 10] or modifying the distribution of documents [65]. Both types of mediation are complementary to each other. Support for collaboration can be categorized along three lines [26, 84]: *division of labour*, *sharing of knowledge*, and *awareness*. Table 2.2 provides a feature comparison of prior works [65, 69, 85, 35, 10, 108] in relation to these concepts.

Division of Labour refers to the distribution of workload across collaborators. This division can be left to the user (user-driven) or mediated by the system. The latter can be implemented at the user level through the assignment of roles, or at the document level by assigning different document subsets [90]. Prior systems mostly support user-driven division of labour through the provision of communication features. Group chat and document sharing (i.e. explicit recommendation of a document to a collaborator) are two features which have been shown to be favoured by users [65, 85].

Sharing of Knowledge refers to the ability to share ideas and information effectively between collaborators [107], and can be facilitated either through shared workspaces [74], or through the re-ranking of search results based on relevance feedback [25]. Prior systems support sharing of knowledge primarily through providing a shared workspace with features for collectively capturing information. Document rating and bookmarking are both relevance feedback mechanisms, with prior systems either implementing one or the other. Bookmarking promotes shortlisting, which involves forming and refining a shared list of potential resources [46], whereas document rating provides a finer granularity of feedback, which is desirable for algorithmic mediation [35, 10]. Document annotation supports the previous two features by communicating the rationale behind an action [65, 46].

Awareness is defined as "*the ability to maintain some knowledge about the situation and activities of others*" [55], encompassing knowledge of the workspace and

collaborators' actions, as well as the ability to notice changes on the work conducted instantaneously. Prior systems focus on providing lightweight information regarding collaborators' search activities (query history, document history, and colour coding) and the overall sense-making process (document metadata and group summary). All features were found to be useful in past experiments, except for the document history which Kelly and Payne [46] reported to provide too much information.

The role dimension is less explored as it is dependant on the search task, in contrast to the mediation dimension which has been developed in the broader field of collaborative systems. Some predefined roles that researchers have explored in CSE include prospector/miner [72] and gatherer/surveyor [86]—both are similar and reflect the two stages of exploratory search described by [103]: one role focusing on exploratory browsing and the other role on focused searching. Both studies employ algorithmic mediation to adapt the search results according to the relevant documents found by each role. Tamine and Soulier [97] looked into the implications of predefined roles and found that it resulted in less coordination and constrained users too much, whereas user-driven roles—although needing more intense communication—achieved better coordination. Soulier et al. [91, 92] has recently explored the automatic assignment of roles based on users' search behaviour.

2.5.3 Suitability of Prior Systems

A majority of the mentioned systems were only described in publications, and not open-sourced (or even available as binaries). `SearchTogether` is available as a plugin for `Internet Explorer`, but we found it to be outdated and not compatible with current versions of most browsers. The only system that we found to be functional was `Coagmento` which has been open sourced and is currently under active development². The developers reported that `Coagmento` at its current state is not very accessible as a tool for research—requiring programming efforts to setup [60]. Additionally, `Coagmento` comes in the form of a browser plugin and mobile application, which makes it less practical for large-scale studies as it requires user-side installation. Since our work aims at deploying the system through a crowd-sourcing platform, the available tools are insufficient for our needs; therefore we opted to develop a new system instead.

²<https://github.com/InfoSeeking/Coagmento>

Chapter 3

SearchX

In contrast to single-user search where a number of up-to-date and open-source tools are readily available (e.g. `Terrier`¹ and `Elasticsearch`²), the CSE research community has currently just one maintained open-source option (`Coagmento`, cf. Table 2.2) despite the fact that researchers have designed and implemented a number of systems in the past ten years [65, 33, 4, 69, 85, 35, 10, 108]. While `Coagmento` provides an extensive collaboration feature set, it requires users to either install a browser plugin or an Android/iOS app, making it less viable for large-scale CSE experiments which are often conducted with crowd workers. Furthermore, we believe as researchers we should have a choice of tooling, instead of relying on a single one.

For these reasons, we have designed and implemented `SearchX`, a CSE system built on modern Web standards—allowing the system to be accessed from multiple platforms without the need for user-side installations as the system is compatible with all major browsers. We designed the system specifically for CSE research and provide comprehensive documentation to enable others to "implement" and run their own CSE experiments with `SearchX`. `SearchX` adopts common collaborative features found in prior works and presents them in a way that is familiar to generic users. Additionally, we emphasise support for research needs, specifically, support fast prototyping of new features and interfaces, and support for crowd-sourced studies.

`SearchX` is designed for the following experimental workflow: researchers first implement an experimental setup of their user study using `SearchX` (either relying on existing features or adding their own). Each study participant accesses the `SearchX` instance through a designated URL. The system then allocates a collaboration team to each set of $m \geq 2$ participants (m is a configuration parameter). Throughout a team's search session (which may include pre/post questionnaires), `SearchX` continuously captures fine-grained user activity logs.

Our main concern regarding the collaboration design is on the mediation aspect, as we aim to support generic collaboration with no predefined roles between collaborators. We first analysed how mediation was designed in prior works and used that as a starting point in implementing `SearchX`. Our analysis of previously proposed systems (cf. Table 2.2) is limited to those similar to `SearchX`—systems that support at least *synchronous* and *remote* collaborations. Additionally, we limit the scope to text

¹<http://terrier.org/>

²<https://www.elastic.co/>

retrieval systems, since it is the most common use case in Web search. As CSE solutions should require low additional effort compared to ad hoc solutions [35, 64, 46], we strove to implement features that look familiar to users (whom all use Web search engines) today.

We now discuss the architecture of SearchX and then elaborate on the two main design directions: supporting collaboration, and empowering research. Afterwards, we elaborate more on the development process of SearchX.

3.1 Architecture

In implementing SearchX, we chose to start from an existing system/interface to save development time. The options were limited as search engine interfaces are generally not open-sourced. We decided to use the single-user `Pienapple` search system [9]³ as a starting point, as it provides a generic Web search interface built with modern Web technologies. Their choice of technologies (`node.js`⁴, `React`⁵) have active developer communities and are supported by large companies, ensuring that the system will be relevant technology-wise for the upcoming years. Given this base system, we vastly expanded its functionalities for collaboration and experimentation, and then refactored the code base to be modular and reusable.

SearchX’s client-server architecture is shown in Figure 3.1. The front-end is responsible for presenting the interface, managing task sessions, and logging user activities; the back-end is responsible for communicating with the retrieval engine, and managing team creation and synchronisation.

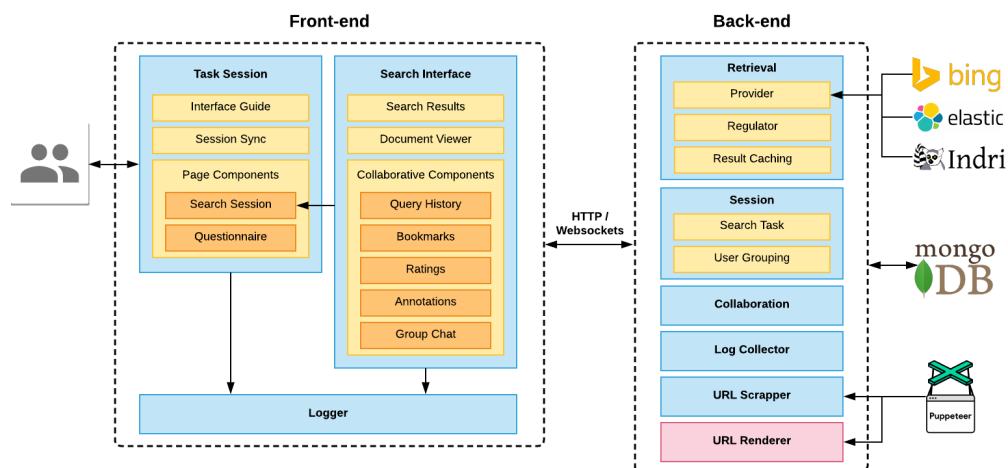


Figure 3.1: SearchX architecture overview.

Front-end. The front-end (shown in Figure 3.2) is developed using `React` (a JavaScript library). `React` manages a client-side data model, thus minimising communication with the back-end; and it enforces the creation of standalone view components, result-

³The authors kindly provided us with their source code.

⁴<https://nodejs.org/>

⁵<https://reactjs.org/>

ing in an interface simple to modify and extend. As the front-end is a Web application, any user with a modern browser can access it without requiring additional installation.

The front-end consists of three logical abstractions. *The search interface* is composed of features related to searching and collaboration, and is presented to the user during the search session. We implemented each feature as standalone components which makes changing the layout or design of the interface efficient. *The task session* defines the desired experimental setup, and controls the search task, team creation, and the experimental procedure (e.g. a pre-test, the search session, and then a post-test). An experimental procedure consists of a sequence of pages, which we bootstrapped by implementing template components for the search session, and questionnaires—as we found them to be the most commonly required templates in our experiments. *The logger* accepts activity data from each component, and regularly sends the logs to the back-end for storage. This abstraction provides a clear separation of concern between interface features, experimental setup, and data collection, making it clear which part of the system needs to be changed for a particular experimental need.

Back-end. We developed the back-end with the `node.js` server environment, which directly supports asynchronous I/O operations, making it suitable for applications requiring real-time updates. An added benefit of `node.js` is its language (JavaScript)—developing both the front-end and back-end in the same language made the development more manageable for us. The back-end provides the application data services which are made available to the front-end through APIs—implemented using the `express`⁶ framework for HTTP and the `socket.io`⁷ library for Web sockets. We chose these two libraries as they are currently the most common libraries for their respective role. We chose MongoDB⁸ for data storage as it uses a dynamic data schema, providing added flexibility during the development and modification of features.

The data services are categorised into four types. *Retrieval services* includes communication with the retrieval system through the *provider* and further processing of the retrieval results through the *regulator*. Currently, we support the Bing Web Search API for searching the Web, and both Elasticsearch and Indri⁹ servers for custom document collections. *Session services* handles team communication and assigning search tasks to users. *Collaboration services* includes the back-end logic of collaborative features in the front-end. *Utility services* includes data collection tools such as the log collector which stores user logs received from the front-end and the URL scrapper which scrapes all documents returned to the user. Additionally, we also have a URL renderer which makes it possible to load external Web pages inside our Web-based system (think of a browser inside a browser), allowing us to implement the front-end document viewer. The viewer makes it possible to keep users inside the system at all times, allowing the system to log user interactions within the opened documents as well. Both the URL scraper and URL renderer utilise a headless browser using Puppeteer¹⁰.

⁶<https://expressjs.com/>

⁷<https://socket.io/>

⁸<https://www.mongodb.com/>

⁹<http://www.lemurproject.org/lemur/>

¹⁰<https://github.com/GoogleChrome/puppeteer>

3.2 Supporting Collaboration

As can be seen in Table 2.2, only three prior systems implement system mediation, with each of them implementing a different type. In contrast, many interface mediation features are frequently found across systems. Accordingly, SearchX supports collaboration through interface mediation while facilitating custom implementation of system mediation through the addition of a regulator layer in the back-end.

We summarise the frequency of each interface mediation feature in Table 3.1. Based on our analysis of mediation in section 2.5.1, we have decided on the following implemented features for each mediation function in SearchX. *Division of Labour* is handled by a group chat; however, we did not implement a document sharing feature which was present in two prior systems—we argue that document sharing as described in prior systems can be achieved through the chat feature as well and thus does not warrant a separate UI element. Features for *sharing of knowledge* largely depend on the experimental setup, therefore SearchX implements all three in a way that toggling individual features is simple. *Awareness* is handled by the addition of query history, document metadata, and colour coding. As SearchX is designed for synchronous sessions, we did not implement a group summary which is most beneficial for asynchronous sessions [84].

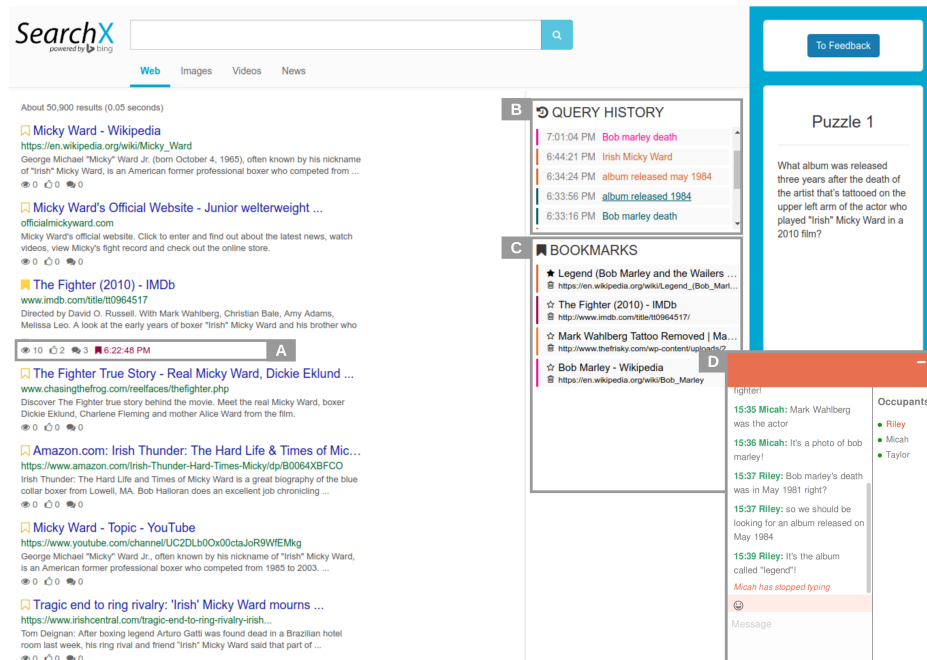
Table 3.1: Frequency of interface mediation features in prior CSE systems.

Function	Feature	Frequency (out of 7 systems)
Division of Labour	Group Chat	6
	Document Sharing	2
Sharing of Knowledge	Bookmarking	2
	Document Rating	5
	Document Annotation	4
Awareness	Query History	7
	Document History	3
	Document Metadata	6
	Group Summary	4
	Colour Coding	2

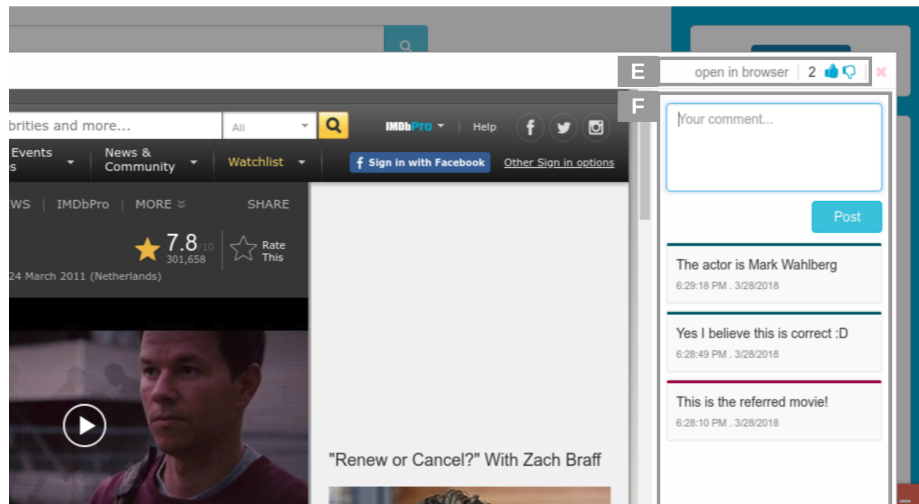
We now discuss each implemented feature. Figure 3.2 shows the UI of our interface with all features enabled.

Group Chat. Even though we already facilitated knowledge sharing through more specialised mediation features, direct communication is necessary for coordination and discussions. We opted for the familiar pop-up design where the chat window is always visible in the interface but can be minimised when not in use (to avoid cluttering the interface). It is implemented using `converse.js`¹¹ which provides a robust chat window out of the box. A downside though is the lack of APIs providing functional access to the internals, making it difficult to extend (e.g. we are not able to automatically assign usernames).

¹¹<https://conversejs.org/>



(a) Search result page



(b) Document viewer (zoomed in on the top right corner)

Figure 3.2: SearchX collaborative search interface. [A] Document metadata, [B] shared query history, [C] shared bookmarks, [D] chat, [E] document rating, [F] document comments.

Bookmarking. Apart from functioning as a means to save documents for later revisits, bookmarking also promotes the shortlisting strategy which involves curating a shared list of potential documents [46]. Given the central role of *bookmarking* in such collaborative efforts, we want to make it more accessible; therefore we implement the bookmark button directly next to each search result. The list of bookmarked documents

is always visible in the sidebar to promote awareness of collaborators' actions. Users benefit in the sense-making process when given the option to manage and rearrange their saved list [46]; therefore SearchX facilitates this using pinned bookmarks.

Document Rating. Document rating is mainly considered as fine-grained source of information for relevance feedback. To avoid cluttering of the SERP, we present the rating buttons not on the SERP but inside the document viewer; the benefit is that users can only rate once they have seen the document. Document rating is implemented as a like/dislike button to leverage users' familiarity with this type of interaction.

Document Annotation. Unlike existing systems, we implemented annotations as a message thread similar to chat interfaces as can be seen in Figure 3.2.b. This setup highlights the bidirectional nature of the annotation process, promoting sense-making through the exchange of opinions. The annotation interface is presented inside the document viewer, directly next to the document to make adding new annotations a quick process.

Query History. This feature has been implemented in all seven prior systems as can be seen in Table 3.1. It provides awareness of collaborators' search activities, allowing users to avoid duplication of effort and learn from their collaborators' choice of keywords [65]. Its implementation is similar across prior systems: as a list of queries that can be clicked on to open results for that query immediately. SearchX provides a scrollable list of recent queries in the sidebar.

Document Metadata. This feature is presented below each SERP entry to provide information about collaborators' activities on the document. Document metadata allows users to quickly identify documents that are considered relevant by their team. The information is presented using simple icons so that it does not take too much attention away from the actual search results.



Figure 3.3: Colour coding of the shared query history.

Colour Coding. We colour code elements of the interface associated with a particular collaborator's actions (such as querying and bookmarking). The colours allow users within the team to differentiate between the activities and contributions of each collaborator. We generate random-but-distinguishable colours to allow scaling to numerous users.

System Mediation. As stated before, SearchX does not implement a specific form of system mediation, but facilitates such an implementation if needed. In Section 2.5 we

outlined that system mediation is usually performed in the form of modifications to the retrieved list of results (re-ranking). A research collaborator has designed the retrieval service in the back-end also to contain a *regulator layer* that enables us to adjust the SERP sent to each user in the same team based on the actions of the team’s members. The regulator layer can fetch and aggregate data related to the search results from the database in order to re-rank them accordingly.

3.2.1 Empowering Research

We now elaborate on how SearchX is designed as a tool for research.

Availability and Accessibility. SearchX is open-sourced for use and development by other researchers^{12,13}. Installation of the system is made simple and effective—it requires executing one command to get the system up and running. We provide three example implementations of different experimental setups (synchronous collaborative search, asynchronous collaborative search, single-user search). We also put significant effort into extensively documenting how researchers can modify the system, e.g. by adding new UI features, or by changing the retrieval system in the back-end.

Interface Guide. Prior works report that users did not explore some features of their system because they do not know or understand it [65, 35]. We solved this issue by adding a guided interface walk-through of the interface (built using `Intro.js`¹⁴) which explains step-by-step what each feature is meant to do. This interface guide is launched when a user first starts the search session, ensuring that they are aware of the features we want them to use.

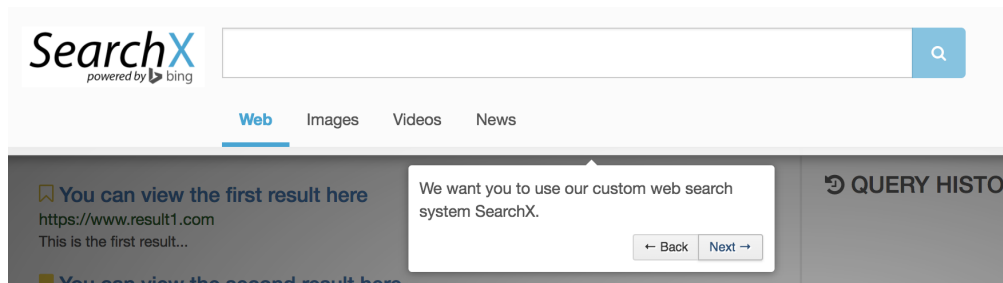


Figure 3.4: Guided interface walk-through

Data Collection. A requirement for a CSE user study is the collection of user activity logs. In SearchX, we have added logging to all interactive components of the system so that it records when a user hovers over or directly interacts with a component (e.g. clicking, querying, opening a document). We also log session related data (e.g. starting/finishing the search session, submitting a questionnaire) and interactions with the browser (e.g. changing tabs), which helps to understand all actions executed by a user. All logs related to user interaction are defined and implemented in the front-end, while the back-end handles storage of logs from the front-end as well as logs related to the

¹²<https://github.com/felipemoraes/searchx-frontend>

¹³<https://github.com/felipemoraes/searchx-backend>

¹⁴<https://introjs.com/>

retrieval engine, making it easy to modify or create additional logs according to what we want to analyse from the user.

User Study Design. Currently, modifying the system requires programming knowledge as we do not provide a graphical interface to create user studies yet. However, we have created reusable implementations of common components in the experimental setup: questionnaires and the search session. The questionnaires are implemented using SurveyJS¹⁵, which allows defining questionnaires directly using JSON. We have created a React component that abstracts over SurveyJS, adding logging features and flow control. We also did the same for the search session, which abstracts over the search interface, adding session-related logs, flow control, and a taskbar to describing the search task. We found this to simplify the creation of new user studies since it takes away much of the boilerplate code needed in configuring the experimental procedure. As an example, the set up of a basic search interface only requires a few lines of code as can be seen in Listing 1.

```

1 //Session.js
2 export default class Session extends React.PureComponent {
3   render() {return (
4     <TaskedSession collaborative={false}>
5       <div className="box" style={{
6         marginBottom: '20px',
7         textAlign: 'center'
8       }}>
9         <Link className={"btn btn-primary"} to="/submit" role="button">
10          Submit Itinerary
11        </Link>
12      </div>
13
14      <div className="box" style={{flexGrow: '1'}}>
15        <h3 style={{textAlign: 'center'}}>Task Description</h3>
16        <hr/>
17        <p>Please search more information regarding "{constants.topic}"</p>
18      </div>
19    </TaskedSession>
20  )}
21 }

```

```

1 //App.js
2 export default class App extends React.Component {
3   render() {return (
4     <Router history={history}>
5       <div>
6         <Route exact path="/" component={Session}/>
7         <Route path="/submit" component={Submit}>
8       </div>
9     </Router>
10  );}
11 }

```

Listing 1: Code required for a search interface with a basic task bar.

3.3 Development

The development of SearchX was a collaboration between two Master students (including me) and one PhD student in our faculty. In implementing SearchX, we adopted

¹⁵<https://surveyjs.io/>

an agile development strategy—involving multiple iterations of development—since we wanted to start the development early while our research design was being fleshed out. The base system we used as a starting point already provided us with a good search interface; therefore our priority in the development was to refactor the code in preparation for implementing the collaborative features. Afterwards, we started to put more structure on the development process. The software requirements—based on our research design—were translated into actionable items in our development backlog, which were then iteratively implemented. During the development process, we followed a few software engineering best practices:

Source Code Management. We released our system as an open source software in GitHub¹⁶—a version control system that is often used to host and develop open source software. We kept a stable build of our software in the master branch and did all our development in a new branch. When the development has finished, we merge the new code into our master branch through a pull request. In resolving a pull request, we require at least one other developer to review the code to keep the code quality high.

Clean Code. We strive for high code quality to make it easier for other people to develop. In general, we prioritised keeping the code readable instead of efficiency. We kept a logical directory structure and minimised interconnected components (by separating business logic from code dependencies) to keep the code robust to change.

Testing. Testing is useful in preventing breaking changes, which is essential for an open source software. It was rather straightforward to test the backend through unit tests, but testing the frontend required more effort as it is less common in practice compared to backend testing. Additionally, we implemented an automatic test system to check for breaking changes on every pull request automatically.

We now reflect on the implementation process and elaborate on the challenges we faced during development as well as the limitations of our final implementation.

3.3.1 Challenges

We now discuss some issues that arose during the development process.

Iterating on the experimental setup. We initially implemented the basic version of SearchX with a paper deadline in mind, which resulted with a functional but not very modular version of SearchX. For each of our experiment, multiple files in both the front-end and back-end required changes. Over time, we started fixing this in the front-end by separating out all code related to the experimental setup from the search interface and encapsulating them into reusable React components. While this simplified the experimental setup, the communication with the back-end remained a complex issue. We now limit the responsibility of the back-end to only team management and synchronisation, allowing us to directly implement the limited range of functionality inside the task components. If we had spent more time on the initial design to improve on the architecture and the interactions among components, we would have saved substantial development time.

¹⁶<https://github.com/>

Deploying a crowdsourced study. As an effort to support online studies, we adapted SearchX for crowd-sourced studies. During an initial CSE pilot study on Crowdflower, we found crowd workers to be unmotivated to properly execute our assigned collaborative search tasks, as they normally perform micro-tasks. We found two ways around this issue: (i) a new platform and (ii) actively encouraging proper behaviour. We switched to the research-focused Prolific¹⁷ platform which was shown to provide higher quality data [70]—something we found to be true as well in our work. We also spent significant development time on monitoring workers’ attentiveness and actively keeping them on track. We logged browser interactions (change tabs, context menu) and notified workers about their tab changes in real-time (more than n tab changes results in non-payment). We also added quality control questions and disabled copy-paste operations in the questionnaires. All these steps improved the quality of data we collected.

Synchronising team sessions. Running synchronous search sessions through a crowd-sourcing platform is tricky since workers are not available right away; therefore a type of “waiting room” is needed for the grouping so that workers assigned to a single team start their search session at the same time. This problem becomes particularly intense as the team size increases—an experiment with 20 workers requires 20 workers to accept the task at roughly the same time. Another issue we encountered was that workers were disconnected from the grouping process when the page was refreshed/closed during the waiting period, resulting in the worker not being able to continue the study. We currently warn workers that attempt to refresh/leave the Web page running SearchX.

Implementing a document viewer. Ensuring that crowd workers remain within SearchX (and otherwise rescind the payment) is a good way of ensuring compliance, but this idea breaks down when we want the workers to interact with the search results (and click on links to view documents in another browser tab). We thus needed to implement a document viewer that allows users to view the document within SearchX. It is not possible to render another Web page directly inside SearchX because of CORS (cross-origin resource sharing) restrictions. We thus had to render the URL in the back-end and pass the rendered HTML to an `iframe` in the front-end. This solution though is imperfect since the resulting page is static with most interactive elements disabled, and at times the rendering is not perfect.

¹⁷<https://prolific.ac/>

Chapter 4

Research Design

This chapter elaborates on the methodology to answer our research questions. In line with recent works on *search as learning* [95, 106], we conducted a user study through a crowd-sourcing platform to observe the effects of providing collaboration capabilities towards search behaviour and learning outcome. We will first elaborate on how we designed a measurable learning task for the study. We then elaborate on the study setup, the evaluated conditions, and the evaluation metrics.

4.1 Learning Task

Discovering more information about a specific topic has been shown to be a common Web search task [8]. We build upon this to design a learning task which is relevant to people in general—study participants are required to browse the Web to learn a specific topic. We adopt a setup similar to Syed and Collins-Thompson [95] in which participants need to complete a pre-test, the search session, and then a post-test. These steps enable us to measure *learning gain* by comparing the results of the two tests. We also adopt the use of vocabulary-based tests—which require participants to demonstrate their understanding of domain-specific terms related to the learning topic—because it is straightforward and can be completed within a short time frame, therefore suitable for a crowd-sourced study where supervision is minimal.

4.1.1 Learning Topics

Online video lectures are now increasingly common, and with the widespread adoption of MOOCs, it is becoming a popular medium for online learning—a setting in which we believe search as learning will play a vital role. We adopt the topics used in our prior research [62], which are taken from high-quality video lectures and are therefore relevant to online learning. Additionally, it allows us to extract topic-specific vocabulary items directly from the video transcript. A research collaborator [62] selected the topics and vocabulary items through a process as follows:

1. Ten videos were chosen from three popular sources of educational video lectures

(TED-Ed¹, Khan Academy², and edX³). For TED-Ed and Khan Academy, the ten most popular videos were chosen; whereas for edX which lacks a popularity listing, ten undergraduate level STEM MOOCs were selected. For the chosen MOOCs, a lecture video with a length of fewer than 15 minutes from the first two course weeks was chosen (therefore less likely to have a specific learning prerequisite). This resulted in a diverse list of topics including *dystopia*, *sticism*, *photosynthesis*, and *radioactive decay*.

2. Vocabulary lists for each topic were extracted manually from each of the thirty candidate videos based on these criteria: (i) it was mentioned in the video at least once; (ii) it does not frequently occur outside of the domain-specific context according to the collaborator's judgment.
3. Three computer scientists annotated each vocabulary item with a score based on the *Vocabulary Knowledge Scale* (VKS) [102, 21] which reflects their familiarity towards a specific term (explained in section 4.1.2). The annotators were not given the lecture videos to avoid affecting their judgement.
4. The candidate topics and vocabularies were reduced to a size that was practical for the study: ten topics with ten vocabularies each. The selection was made by picking ten vocabulary items with the lowest VKS score for each topic candidate. From the list of topics, we selected ten topics with the lowest total VKS score (the addition of all VKS scores in their respective top ten vocabulary lists).

This process ensures that during the study, participants will find at least one unfamiliar topic, therefore offering more potential for knowledge gain. Additionally, all participants will start with a similar low prior knowledge on their given topic, which makes the learning gain between participants comparable. The final topics cover a wide range of domains as can be seen in Table 4.1.

4.1.2 Knowledge Assessment

In assessing participants' vocabulary knowledge, we avoid the use of closed multiple choice questions as, without the proper domain expertise, it is hard to design choices that accurately reflect participants' knowledge [59]. We opted to use the *Vocabulary Knowledge Scale* (VKS) test which has been shown to be a reliable indicator of vocabulary knowledge [21]. The test evaluates a participant's ability to recall a specific term by using an incremental scale consisting of 5 states [68] as follows:

1. *I don't remember having seen this term/phrase before.*
2. *I have seen this term/phrase before, but I don't think I know what it means.*
3. *I have seen this term/phrase before, and I think it means ____.*
4. *I know this term/phrase. It means ____.*
5. (Not used) *I can use this term/phrase in a sentence: ____.*

¹<https://ed.ted.com/>

²<https://www.khanacademy.org/>

³<https://www.edx.org/>

Table 4.1: Overview of topics per condition. Conditions: single-user search (SE) and collaborative search (CSE).

Topic	Participants		Source	Video length	Average VKS
	SE	CSE			
Radioactive decay	4	10	edX	6m53s	2.72
Qubit	5	2	edX	12m24s	2.81
Water quality aspects	2	0	edX	10m45s	2.88
Religions	0	0	TEDEd	11m09s	2.91
Sedimentary rocks	3	6	edX	5m03s	2.92
Anesthesia	4	2	TEDEd	4m55s	2.94
Glycolysis	5	20	Khan	13m29s	2.97
Urban water cycle	1	2	edX	7m40s	3.01
Depression	0	0	TEDEd	4m28s	3.02
Industrial biotechnology	2	8	edX	5m48s	3.02
Total participants	26	50			

States (1) and (2) indicate unfamiliarity with the vocabulary item, whereas states (3) and (4) require the recall and reproduction of the definition, therefore reflecting an understanding of the vocabulary item (differentiated by the level of certainty). State (5) is geared towards second language learners and evaluates the ability to use the term in a grammatically correct way. Since participants in our study are limited to native English speakers, we only utilised the first four states for the vocabulary assessment. Accordingly, the vocabulary assessment requires participants to fill in the appropriate VKS state for each of the ten vocabulary items in a topic.

The VKS test relies on participants' own ability to demonstrate an understanding of the term, therefore avoiding the use of cue seeking and exhaustive strategies to guess answers, a practice which is common with closed multiple choice questions [59]. However, the results depend on participants' self-assessment of their understanding, where it is possible to incorrectly report their understanding, especially for states (3) and (4) where the produced definition may be incorrect or too generic. A manual assessment by a research collaborator in [62] confirmed that participants' perceived learning outcomes (i.e. the VKS state) match closely with their actual learning outcomes (i.e. the produced definitions). This result is in line with [16]; therefore we can be confident in the accuracy of the reported VKS scores, and use them in our analysis.

In addition to the vocabulary assessment, we also employ a written assessment to evaluate higher-level understanding of the topic (i.e. understanding structure or concept related to the topic), specifically in the form of a summary and an outline of the topic. We require the summaries to be at least 50 words long to make sure it is long enough for meaningful analysis.

The questions used for both the vocabulary assessment and the written assessment can be seen in Appendix A.3 for the pre-test and Appendix A.5 for the post-test.

4.2 Study Setup

Every participant starts the study by viewing a briefing containing the terms and requirements as well as the study setup. Participants then undertake a pre-test for which the system randomly selects three of the ten learning topics. The pre-test only contains the vocabulary assessment and not the written assessment to keep it within a reasonable duration. The results from the vocabulary assessment are used to determine the topic for the learning task. Afterwards, participants are asked to undergo a search session which differs according to the assigned condition. During the search session, participants are instructed to explore the given topic using the search system we provided. We used the task description in Figure 4.1, which was adopted from previous studies [51, 16] with adjustments to emphasise learning. The task focuses on acquiring knowledge on the learning topic, with no mention regarding the post-test in the description to prevent participants from learning only the vocabulary items instead of the whole topic.

Imagine you are taking an introductory [**general topic, e.g. *Health and Medicine***] course this term. For your term paper, you have decided to write about [**specific topic covered in the video e.g., *the symptoms and treatments of depression***].

The professor requires all students to demonstrate what they learn about a particular topic by collaboratively conducting searches online and presenting their views on the topic. To prepare your term paper, you and your partner need to collect and save all the webpages, publications, and other online sources that are helpful for you to write a paper. After you and your partner have completed the search phase, you will be asked to complete 13 exercises; those exercises include questions about your term paper topic and the writing of an outline for your term paper. Those exercises are solved individually (without your partner).

Figure 4.1: Task description for all conditions. The underlined green phrases were only added in the CSE condition.

The search session has a minimum time which is made visible through a timer (as can be seen in Figure 4.2), and participants can start the post-test as soon as they reach the minimum time. Prior studies typically assign a search session of 25 minutes [36, 13, 12, 83] with a minimum of 15 minutes [44] and a maximum of 35 minutes [87, 88]. We settled on a twenty minutes search to provide enough time for learning while keeping the study short enough for crowd-workers. The study ends with a post-test which consists of both the vocabulary assessment as well as the written assessment. Appendix A provides more information on the exact setup.

We devised two search conditions where participants are required to carry out a learning-oriented search task. Figure 4.3 provides an overview of the study setup we employ for this thesis.

Condition 1: Single User Search (SE)

The first condition serves as a baseline, in which participants go through a search session individually. From the pre-test, a total score for each topic is computed by

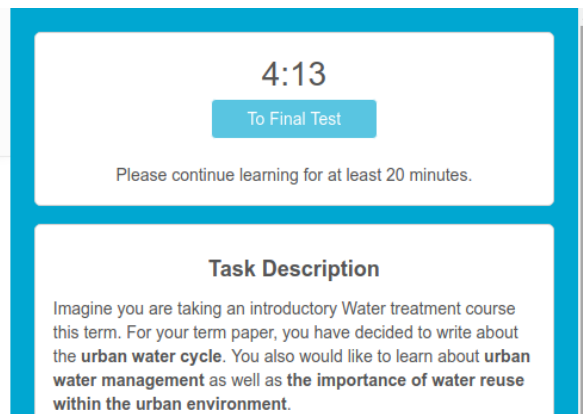


Figure 4.2: The task bar containing the timer and the task description.

summing all the individual vocabulary item VKS scores. Each participant is then assigned a topic from the pre-test topic with the lowest total VKS score (i.e. their most difficult topic). In case of a tie, we randomly choose one topic. The search interface provides generic Web search functionalities and uses the Bing search API to serve high-quality Web search results, simulating a typical Web search scenario.

Condition 2: Collaborative Search (CSE)

In the second condition, participants complete the learning task in teams of two. The two collaborators independently perform both tests and only collaborate during the search session. In determining the topic for the CSE condition, we take the common most difficult topic by summing up the topic scores for both collaborators and taking the topic with the lowest total score. To ensure that both collaborators start the search session at the same time, they enter a 'waiting room' until the system receives both pre-test results.

The search interface was adapted to provide collaborative features as described in Section 3.2. We provided an interface with a set of features commonly found in prior research [108, 109, 83]—shared bookmarks, shared query history, and chat. In the post-test, we also require participants to describe how they collaborated during the search session and rate the usefulness of the different collaborative features (as can be seen in Figure A.12).

4.3 Crowd-work Deployment

We deployed our pilot studies on the *CrowdFlower*⁴ crowd-sourcing platform. Through a series of initial deployments, we arrived at the study setup described in Section 4.2. Additionally, we added five compliance steps in our setup to ensure the search session goes as expected:

⁴www.figure-eight.com—The name of the platform has recently changed to *FigureEight*. We still refer to it as *CrowdFlower* in this thesis.

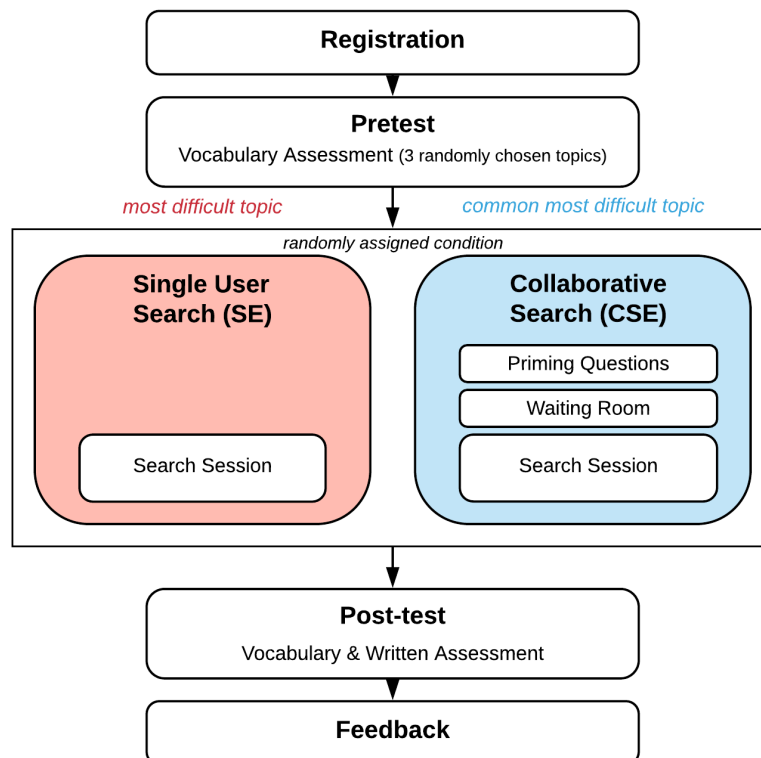


Figure 4.3: Overview of the study setup. Two search conditions are evaluated: single user and (pairwise) collaborative search.

- i We disabled the study if a participant accessed the system through a mobile device.
- ii We enforced a minimum word count of 50 words to the written assessments in the post-test.
- iii We included an easy topic in the pre-test for quality control and excluded workers who reported VKS knowledge level (1) or (2) more than four times. The chosen topic was *sports* with well-known vocabulary items such as *compete* and *baseball*. From this, we disqualified five participants from the user study.
- iv We disabled copy & paste on the search interface and alerted the participants every time they do a tab change during the pre-test and post-test. In the study description, we warn the participants that three such tab changes will lead to a disqualification and non-payment to dissuade participants from searching the Web for answers during the tests. Along the course of the study, we disqualified 6 participants due to this restriction.
- v Because crowd-workers in CSE have no pre-established rapport, we reinforced the collaborative nature of the upcoming search session through an additional

section at the end of the pre-test. The section includes examples of collaborative Web search in daily life and seven questions on participants' past collaboration experience taken from a prior survey in collaborative Web search practices [64], serving to produce a priming effect. The exact questions can be seen in Figure A.6).

For the deployment of the actual study, we switched to the *Prolific*⁵ crowd-sourcing platform, which has been shown to produce higher quality data and attention rates compared to *CrowdFlower* [70], making it more suitable for tasks which are cognitively demanding. This advantage is because crowd-workers in *Prolific* are used to receiving tasks which takes more time to finish, as opposed to crowd-workers in both *CrowdFlower* and *Amazon Mechanical Turk*⁶ which mostly receive micro-tasks.

4.4 Study Participants

During the deployment, we limited the participation pool to countries which have English as their first language (UK, USA, Canada, and Australia) given the reliance on English reading skills in the study. We ran the crowd-sourcing task until we received 26 participants for SE and 25 pairs for CSE for a total of 76 participants. We gave each participant compensation equal to £5.00 per hour for their participation in the study. As the CSE condition requires two participants to work synchronously (i.e. be online and start the task at the same time), we added a waiting period of at most 10 minutes after completion of the pre-test. If within the waiting period no other participant completed the pre-test for the same topic set, we release the waiting participant from the study and give a partial payment of £1.25 as compensation for completing the pre-test and waiting for a partner.

The final set of 76 participants had a median age of 32 (minimum: 18, maximum: 66), and were 72.37% female. 43% of the participants reported a high-school diploma as their highest academic degree, 43% an undergraduate degree and the remaining 14% a graduate degree. From CSE participants, 76% reported having collaborated on the Web, in which of those, 67% collaborated in pairs, and 33% collaborated in teams of more than two. We show the distribution of participants across all learning in Table 4.1.

4.5 Evaluation Metrics

We adopt metrics utilised in prior studies for both search behaviour and learning in order to keep the result comparable. We calculated the metrics in both the individual and team level where appropriate. The team level metrics involve calculating the metric using data combined from both collaborators. To calculate team level metrics for SE, we generated all possible pairs of participants per learning topic similar to [83], resulting in 37 artificial SE teams. We can then observe the synergistic effect by comparing pairs that collaborated with pairs that did not have any group contact at all.

⁵<https://prolific.ac/>

⁶<https://www.mturk.com/>

4.5.1 Search Behaviour Metrics

We take inspiration from the evaluation framework proposed in [87] which analyse four aspects of exploratory. We adopted the two aspects related to search behaviour—*information search* and *information exposure*—and also metrics related to collaboration.

Information Exposure

Information exposure refers to the amount of information a user discovers through searching and browsing, which in Web search primarily corresponds to a user’s interaction with the SERP. Therefore, we computed metrics related to document coverage: the number of distinct viewed documents (*DistinctDocs_t*) and the number of bookmarked documents (*BookmarkedDocs_t*).

Since we deployed our study for a learning task, we are not only interested in the quantity of exposure, but also the quality of exposure. We assume that a longer reading time results in more time to learn from the documents. We therefore also compute participants’ total reading time (*ReadingTime_t*) and the number of documents that were useful (*UsefulDocs_t*). We define useful documents as having been read longer than 30 seconds [37], similar to [83].

Information Search

Information search refers to the way participants seek information, which in the case of Web search is reflected by how participants issue queries through a search engine according to their information need.

We computed simple query metrics for each user/team t : the number of distinct queries (*DistinctQueries*), the average number of words in each query (*QueryLength*), and the number of distinct keywords in all queries (*DistinctKeywords*). Additionally, we adopted a metric from [87] which measures the level of difference between each query in *DistinctQueries_t* (i.e. the diversity of queries). To calculate the difference, they used the character-wise Levenshtein distance between each pair of query strings S_a and S_b . They then computed the metric as the average of the resulting distances.

$$\begin{aligned} \text{QueryDiversity}_t &= \text{mean}\{\text{LevenshteinDistance}\{S_a, S_b\}\}, \\ &S_a \neq S_b \wedge \{S_a, S_b\} \in \text{DistinctQueries}_t \end{aligned} \quad (4.1)$$

Before processing the queries, we manually corrected misspelt query terms and standardised the spelling of similar terms (e.g. anaesthesia and anesthesia). Additionally, we cleaned up the terms by using stop words removal⁷ and a stemmer⁸ to remove redundancy and give more weight to unique keywords. We show two examples of the cleaning process in Table 4.2.

⁷using the `stopwords` corpus from the `nltk` Python package

⁸using `PorterStemmer` from the `nltk` Python package

Table 4.2: Two examples of the query cleaning process.

Steps	Example 1	Example 2
Original query	what is inhalational anesthesia	phases of glycolsis
Manual correction	what is inhalational anaesthesia	phases of glycolysis
Stop-words removal	inhalational anaesthesia	phases glycolysis
Stemming	inhal anaesthesia	phase glycolysi

Collaboration

As collaboration is not the result of a single action, we measure the impact of collaborative actions by the observable outputs of collaboration: communication between collaborators and overlapping resource usage.

We represent textual communication by the chat frequency within a team (*ChatFreq*), which we calculate as the total number of chat messages between the two collaborators. Additionally, we calculate the chat ratio (*ChatRatio*) to indicate the interactivity of the discourse, i.e. whether the chat session was one-sided or not. For two collaborators t_a and t_b in team t , we calculate the metrics as follows:

$$ChatFreq = Chat_{t_a} + Chat_{t_b} \quad (4.2)$$

$$ChatRatio = \begin{cases} \frac{\min(Chat_{t_a}, Chat_{t_b})}{\max(Chat_{t_a}, Chat_{t_b})}, & Chat_{t_a} \neq 0 \wedge Chat_{t_b} \neq 0 \\ 0, & otherwise \end{cases} \quad (4.3)$$

Overlapping resource usage reflects how collaborators affect each other's search output. We represent it by the number of overlapping queries (*QueryOverlap*), keywords (*KeywordOverlap*) and viewed documents (*DocOverlap*). However, these are crude metrics as we do not know precisely whether the overlap is a coincidence or a product of collaboration. We assume that most of the overlapping resources are a product of workspace awareness. Overlapping resource usage is calculated by the union of the resources as follows:

$$DocOverlap_t = DistinctDoc_{t_a} \cap DistinctDoc_{t_b} \quad (4.4)$$

$$QueryOverlap_t = DistinctQuery_{t_a} \cap DistinctQuery_{t_b} \quad (4.5)$$

$$KeywordOverlap_t = DistinctKeyword_{t_a} \cap DistinctKeyword_{t_b} \quad (4.6)$$

4.5.2 Learning Metrics

We report learning metrics for both types of assessment in the post-test: the vocabulary assessment and the written assessment. Additionally, we measure knowledge gain over the search session through implicit learning metrics.

Vocabulary Assessment

We adopted the same vocabulary learning metrics as [95]—absolute learning gain (*ALG*) and realised potential learning (*RPL*)—to enable direct comparison of our results with their work. ALG measures the aggregated knowledge gain across all vocabulary items between the pre-test and post-test. Knowledge gain, in this case, refers

to an increase in the VKS state, which reflects how much their understanding of the vocabulary items has improved. Hence, for a set of n vocabulary items v_1, \dots, v_n with an assigned knowledge score of vsk , ALG is then computed as follows:

$$ALG = \frac{1}{n} \sum_{i=1}^n \max(0, vsk_{v_i}^{post} - vsk_{v_i}^{pre}) \quad (4.7)$$

Since an increase from state (1) to (2) is natural (participants have seen all vocabulary items in the pre-test), we assign a score of 0 to state (1) and (2), and consequently decrease the score of state (3) to 1 and state (4) to 2. Note that in this way, we assume the amount of learning from state (2) to (3) and from state (3) to (4) are equal. This scoring scheme means that we interpret a score of 1 as having half of the vocabulary knowledge for a term (vaguely knowing the meaning), and a score of 2 as being thoroughly knowledgeable of the term.

RPL normalises ALG by the user's prior knowledge, which is calculated by dividing ALG with their maximum possible learning gain (MLG). MLG is calculated similarly to 4.7 but instead of the post-test score, we use the maximum possible score, i.e. for our case a score of 2 for an improvement from knowing the word (1) to understanding the meaning (3). Tables 4.3 and 4.4 show example calculations for both the test score and the learning metrics.

$$MLG = \frac{1}{n} \sum_{i=1}^n \text{MaxScore} - vsk_{v_i}^{pre} \quad (4.8)$$

$$RPL = \begin{cases} \frac{ALG}{MLG}, & MLG > 0 \\ 0, & otherwise \end{cases} \quad (4.9)$$

In computing the team level metrics for vocabulary assessment, we averaged the pre-test scores of both collaborators and took the maximum post-test score between the two collaborators. This approach is to take into account the possibility of team members focusing on different parts of the topic. For a team composed of participants B and C in table 4.4, the average pre-test score is 6, and the maximum post-test score is 12; therefore the team ALG is 6, the team MLG is 14, and the team RPL is 0.43.

Table 4.3: Two calculation examples for test score in the vocabulary assessment. Q1 until Q5 displays the reported VKS state and its corresponding score.

Participant	Q1	Q2	Q3	Q4	Q5	Test Score
A	2 (0)	1 (0)	3 (1)	2 (0)	3 (1)	2
B	3 (1)	2 (0)	2 (0)	4 (2)	4 (2)	5

Written Assessment

Measuring learning gain in the written assessment is much harder compared to the vocabulary assessment because of its non-quantitative nature. We adopt the sense-making measurements proposed by Wilson and Wilson [105] for analysing summaries

Table 4.4: Three calculation examples for ALG and RPL.

Participant	Pre-test Score	Post-test Score	MLG	ALG	RPL
A	4	8	16	4	0.25
B	4	12	16	8	0.5
C	8	12	12	4	0.33

produced from exploratory search sessions. They noted that the metrics were not useful in differentiating between high and low prior knowledge, however since in our study all participants were assigned with their most difficult topic (i.e. having low prior knowledge on the topic), the metrics are suitable for our case.

In processing the written assessment, since we observed that some participants mistakenly wrote most of their informational content in the outlines, we manually inserted statements from their outline into the end of their summary. The resulting average length of topic summaries were 105 words, which according to [105] are considered short summaries. Therefore, we adopt metrics suggested for short summaries which includes simple fact and statement counting (*F-Fact*, *F-State*), topic coverage (*T-Count*), and the quality of facts (*D-Qual*).

Table 4.5: Rating criteria for quality of facts (*D-Qual*) [105].

Rating	Description
0	Facts are irrelevant to the subject; facts hold no useful information or advice.
1	Facts are generalised to the overall subject matter; facts hold little useful information or advice.
2	Facts fulfil the required information need and are useful.
3	A level of technical detail is given via at least one key term associated with the technology of the subject; statistics are given.

F-State is calculated by counting individual statements, which for summaries typically translates to directly counting the number of sentences. *F-Fact* requires the identification of facts (defined as individual information pieces [105]) within the statements. *D-Qual* relies on human judgement to rate the quality of the summary as a whole. To keep *D-Qual* more objective, a set of rating criteria was proposed in [105] as can be seen in Table 4.5.

T-Count is more complicated as it involves establishing a set of subtopics that were covered by participants for each learning topic and counting the number of subtopics covered by each participant. Since the resulting set of subtopics differ in sizes for each topic, we standardised the resulting subtopic count by the total number of subtopics to keep the metric comparable across topics.

Table 4.6 shows two summaries taken from our study and their corresponding metrics. In example 1, the summary is well structured, and each sentence contains a fact;

Table 4.6: Two example summaries with their resulting metrics.

Summary	F- Fact	F- State	D- Qual	T- Count
1 This paper will explain the process of Glycolysis in detail. Glycolysis is the breakdown of sugars, and is a part of Cellular Respiration. There are 2 major phases: the preparatory phase, and the the payoff phase. The first 5 steps are the preparatory phase, since they consume energy to convert glucose into two three-carbon sugar phosphates (G3P). The second half of Glycolysis is known as the payoff phase, and is characterised by a net gain of the energy-rich molecule ATP and NADH. Since glucose leads to two triose sugars in the preparatory phase, each reaction in the payoff phase occurs twice per glucose molecule	5	5	3	3
2 I have learned that there are two different forms of radiation: Ionising and Non-ionising, and there are various forms: alpha, beta, gamma, Xray, neutron. I have learned that everything emits radiation through vibrations. Radioactivity is all around us in both harmful and non-harmful forms. I have learned that gamma is the most destructive and powerful form of radiation, occurring in black holes. Radiation occurs both naturally and can be man-made. It can be utilised for energy and medical breakthroughs but must be treated with respect and care. It has been harnessed in some good ways, and led to many scientific breakthroughs, but caution is needed as several disasters have shown.	6	8	2	3

therefore $F - State$ and $F - Fact$ are equal. Because the explanation is very technical with the use of many topic-specific keywords, we gave it a $D - Qual$ score of 3. They covered three topic areas: the definition of Glycolysis, steps of Glycolysis, and the product of Glycolysis. For example 2, the first and last sentence are not facts; therefore $F - Fact$ is two less than $F - State$. The coverage of the topic is quite broad, but was unclear and not very technical, therefore a $D - Qual$ rating of 2 is given. There are also three topics covered: Definition of radioactivity, types of radiation, and properties of radiation. We found the scores to be partially subjective, but the guidelines provided in [105] was clear enough; hence we believe the variance would not be significant.

Implicit Learning Gain

The use of implicit learning gain measures allows us to estimate knowledge gain over the course of the search session. We adopt the approach from [13] which estimates a

user's learning gain from their search behaviour, specifically on the required knowledge in accessing specific queries and documents. They assume that a user with higher knowledge on a topic is more likely to recognise *hard-to-find* documents (i.e. documents that were accessed by fewer participants) and issue *domain-specific* keywords. Therefore, a user's knowledge level at a certain point in the search session is quantified as the average complexity (in terms of *likelihood of discovery*) of their accessed relevant documents and issued queries.

Because in our study, we do not have data on document relevance, we estimate the relevance of documents based on how useful the document is for learning. Following the previous definition of useful documents, we calculate the document complexity using only documents that were read for longer than 30 seconds by any one user. For a user population of N , we calculate the document complexity (*DocComplexity*) of document d_i as follows:

$$DocComplexity_{d_i} = \log \frac{N}{n_{d_i}} \quad (4.10)$$

Since there were a lot fewer queries per participant in our study compared to [13], we decided to use keyword complexity (*KeyComplexity*) instead of the query complexity. For a keyword k_i , the metric is calculated as follows:

$$KeyComplexity_{k_i} = \log \frac{N}{n_{k_i}} \quad (4.11)$$

n_{d_i} indicates the number of participants who accessed document d_i , and n_{k_i} indicates the number of participants who issued keyword k_i .

Chapter 5

Results

We start this chapter by elaborating on the feedback we received regarding the usability of SearchX as a collaborative search system (**RQ1**). We then investigated differences in search behaviour caused by collaboration during the learning task (**RQ2**), and its effect on learning outcomes (**RQ3**). Because our sample size is not large and we do not have enough information to assume a specific distribution confidently, we utilise the two-tailed *Mann-Whitney U* test ($\alpha = 0.05$) to identify significant differences between the two populations.

5.1 Participants' Collaboration Experience

In order to create a suitable system for our user study, we have adopted how prior systems accommodate collaboration during a search session (in Section 2.5.1). We now assess the ability of the resulting system (as described in Chapter 3) at facilitating collaborative search based on participants' feedback.

As the role of both the shared bookmarks and query history in facilitating collaboration are mainly in providing workspace awareness, we cannot precisely measure the actual use of both features from the interaction logs. Therefore, we rely on participants' self-reports regarding the importance of each feature in helping them collaborate with their partner. In Table 5.1, we identified the feature referenced in each CSE participants' self-reports regarding how they collaborated (each self-report may reference more than one feature).

Table 5.1: Self-reported use of collaborative features among CSE participants.

Feature	# of Participants	% of Participants
Chat	18	36%
Bookmarks	33	66%
Query History	15	30%
None	6	12%

Six participants reported that they did not collaborate at all—all of them due to the lack of rapport: either them not wanting to collaborate or their partner was not coop-

erating. The other 88% of participants reported the use of one or more collaborative features. Shared bookmarking was the most widely used collaborative feature with 66% of participants reporting its use in collaborating with their partner, followed by chat (36%) and the shared query history (30%).

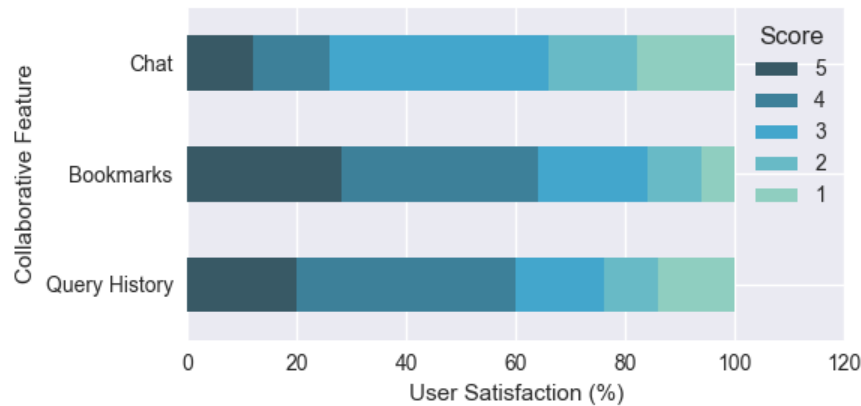


Figure 5.1: Collaborative feature ratings (5-point Likert scale) by CSE participants.

We display a more in-depth look at user's satisfaction regarding the collaborative features in Figure 5.1. We required CSE participants to express their satisfaction towards each of the three features (chat, bookmarks, and query history) using a 5-point Likert scale (from "strongly disagree" to "strongly agree"). The x-axis represents a stacked percentage (%) of participants for each different score. More than 60% of participants considered both the shared bookmarks and query history to be useful in collaborating with their partner. The result for the bookmarks is in line with the self-reports. However, the query history was not a widely used feature based on the self-reports. One explanation is that the query history functions as an awareness mechanism, as opposed to bookmarking which require participants' direct action. As an awareness mechanism, participants are likely not to feel directly utilising the query history, resulting with few participants reporting its usage.

On the other hand, the chat feature was more widely used according to the self-reports but received mixed ratings with 26% of participants agreeing that it is useful and 34% disagreeing. A closer look at the self-reports reveals that quite a few participants found chat to be unnecessary as the other two features were more than enough or their conversations were not fruitful. Some of the self-reports are listed below:

- *"I found checking the chat window, put me off my stride for searching for info."*
- *"We didn't really use the chat feature as my partner wasn't communicating. I would always prefer to do a collaborative search in person so we could discuss the subject properly."*
- *"The shared query history and bookmarks worked great. Those were helpful. The chat also worked great, but we didn't really use it. I don't know what we were supposed to talk about in the chat. We both searched, and got information*

individually."

- *"(We collaborated) Using the shared history and bookmarks. The chat was not needed as there were no questions that the searches didn't answer and no reason to converse to find the needed information."*

The low perception of group chat as a useful feature for the assigned task is in line with [108] which concludes that tasks which do not require immediate decision making depend more on workspace awareness (bookmarks and query history) compared to direct communication (chat). Another possible reason for the low satisfaction is because the collaborators did not know each other in advance and were not given time to build rapport before the search session, resulting with less motivation to communicate directly with their partner. This lack of rapport between collaborators is one area which we had less control over in order to make the crowd-sourced study feasible, unlike prior studies which recruited participants with the same background (i.e. same university) [83, 12] or even participants who already knew each other [44, 87].

To further analyse the discourse, we looked into the type of peer interactions that occurred through the group chat. We manually categorised each teams' chat discussions into four categories based on our observations: indications of verbal immediacy, coordinating on the task, suggesting resource related to the task, and actual discussions of the topic. Verbal immediacy refers to linguistic activities that convey positive messages of closeness (e.g. frequent use of emojis and humour or the sharing of personal information), and has been shown to be a good indicator of social presence [94]. Table 5.2 shows one example discussion in our study for each type of peer interaction.

Table 5.2: Example discussion for each peer interaction category.

Category	Example
Verbal immediacy	A : any background in biology? B : I did it at high school but that was many years ago A : same here! :D
Coordination	A : I find that the websites for kids explain plainer than scientific ones. B : having more trouble with why are they important to the environment B : I agree, but presumably we are to write a college level paper?
Resource suggestion	A : need to find out the types and how it happens. B : its an unstable atomic nucleus A : just read that on the bookmark I made
Topic discussion	A : how many major stages have you found? A : some places say 3, some 4 B : to be honest it's all gobbledy gook but I only find reference to 5 steps

The results in Table 5.3 show that although collaborators in our study did not know each other, 76% of the participants were still able to have a meaningful conversation, with 60% displaying verbal immediacy behaviours. These results indicate that most participants were willing to build a rapport for the task at hand.

Table 5.3: Occurrence of peer interactions through chat within CSE teams. The frequencies are shown in number (#) of teams and percentage (%) over all teams. Each team can have more than one category of peer interactions.

Category	# of Teams	% of Teams
No conversation	6	24%
Verbal immediacy	15	60%
Coordination	11	44%
Resource suggestion	8	32%
Topic discussion	7	28%

Two research collaborators from [62] have categorised the reported difficulties that participants experience in using *SearchX*. They used an open card-sort approach and independently sorted the user self-reports into groups of similar reported problems. The two collaborators then discussed the differences and combined the groups, resulting in the seven problem categories. Note that since each self-report can only be in one category, the numbers reflect each user's most dominant problem.

We see in Table 5.4 that CSE participants reported fewer problems, with nearly half of them not experiencing any problems. SE participants reported more problems related to the task (i.e. task setup and unclear focus), whereas CSE participants did not experience any problems related to the task at all—most likely because participants in CSE were able to leverage the existence of a partner to get a better grasp of the task at hand, whether through workspace awareness or direct communication (we found previously that 44% of teams coordinated in completing the task). As the task did not hinder CSE participants, they focused on their searching experience and therefore reported more difficulties in searching.

Table 5.4: Frequency of self-reported difficulties shown in % of total participants. [62]

Category	SE	CSE
No problems reported	38%	48%
Task setup	19%	-
Unclear focus	12%	-
Searching	12%	32%
Sensemaking	8%	12%
Credibility of sources	7%	8%
Attitude	4%	-

We can conclude that *SearchX* was able to function well as a collaborative search system, as indicated by its success in accommodating collaboration during search. The use of collaborative features was found to result in better overall experience and helped participants in understanding the task. Additionally, most teams were able to have a meaningful conversation, indicating that having good initial rapport is not a strict requirement for collaboration (although it helps).

5.2 Search Behaviour

In **RQ2**, we look into the differences in search behaviour between SE and CSE. We wanted to investigate whether the effect of collaboration on search behaviour for a learning task is similar to those of prior studies as elaborated in 2.5. We compared the conditions at both the individual and team level as suggested in [87]. For SE, we simulated the teams to make it comparable with CSE.

5.2.1 Collaborative vs. Single User Search

We started by comparing the effects of collaboration on participants’ information exposure as shown in Table 5.5. There is a significant difference in the use of bookmarks, but we believe this is a result of our system design—in *SearchX*, the same document cannot be bookmarked twice by participants in the same team. Thus, the existence of a partner in CSE decreases the need to bookmark many documents, since the responsibility of bookmarking relevant documents are split between the two collaborators. There is also a significant difference in the number of overlapping documents with participants in CSE having twice as much as SE. This difference is most likely because CSE participants have an awareness of their partner’s bookmarks, resulting in them being encouraged to check out the document.

Table 5.5: Statistics on information exposure across search conditions (mean). **Bold** indicates statistical significance ($p \leq 0.05$).

	User			Team		
	SE	CSE	p-value	SE	CSE	p-value
Total Reading Time	11m19s	10m11s	.34	-	-	-
Distinct Documents	10.35	10.54	.50	17.78	18.28	.59
Useful Documents	4.96	5.00	.93	9.00	8.04	.15
Bookmarked Documents	12.46	7.38	.0001	22.22	14.76	.0003
Document Overlap	-	-	-	1.22	2.80	.0001

Overall, participants who collaborated did not spend a significantly different amount of time opening and reading documents from the SERP. However, workspace awareness did affect how participants selected potential documents, with them opting to open documents bookmarked by their partner.

We then looked into the effects of collaboration towards participant’s information searching behaviour as can be seen in Table 5.6. We observe no significant differences between SE and CSE concerning querying behaviour—participants who collaborated during a learning-oriented search session did not demonstrate a difference in the way they issue queries compared to participants who searched independently.

To get a better understanding of the participants’ search behaviour over time, we plot the occurrences of individual search actions over different stages of the search session in Figure 5.2. We divided each participant’s session into five equal sections of around 4 minutes and 20 seconds. We used the first query occurrence as the beginning

Table 5.6: Statistics of information search across search conditions (mean).

	User			Team		
	SE	CSE	p-value	SE	CSE	p-value
Distinct Queries	6.38	4.92	.45	10.89	7.88	.08
Distinct Keywords	8.12	6.82	.34	11.76	9.88	.19
Query Length (Words)	2.87	2.46	.37	2.61	2.57	.81
Query Diversity	13.35	14.03	.86	13.71	14.77	.82
Query Overlap	-	-	-	1.49	1.96	.22
Keyword Overlap	-	-	-	3.65	3.76	.67

of the search session and the last interaction with the search page as the end. For CSE, we denote the start of the session by the first query issued by either collaborator.

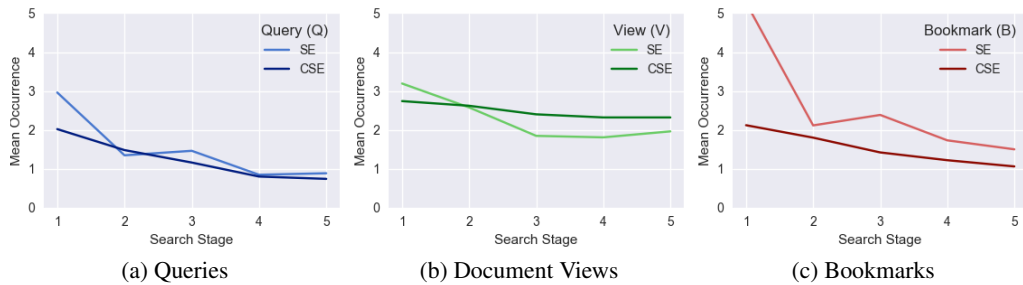


Figure 5.2: Comparison of search actions (mean value) for each of the five search stages. The stages are created by splitting a participant’s search session into five equal sections. As the standard deviations are large (because of the small sample size), we do not display the error bars.

From the three actions, we found significant differences in the number of bookmarks and queries. The number of bookmarks per search stage shows a consistent difference with CSE participants submitting fewer bookmarks than SE participants, similar to our previous analysis. However, the number of bookmarks is only significantly different at search stage 1 ($p = .00003$) and 3 ($p = .009$) with a spike in the number of bookmarks in SE, suggesting that SE participants explored more in the beginning and middle of the session.

This behaviour is further supported by a similar trend in the number of queries, with a small spike at search stage 1 and 3, but it is only significant for search stage 1 ($p = .02$) and not for stage 3 ($p = .89$). On a closer look, we found that most of the queries in search stage 1 were intended to explore the basics of the topic with most of the queries directly containing the topic name. Queries in search stage 3 and above on the other hand typically involve more advanced topics such as pyruvates (in *Glycolysis*) and gamma radiation (in *Radioactivity*). We give two examples of queries per search stage in Table 5.7.

We believe the lower number of queries at stage 1 for CSE is because a partner’s query history and bookmarks act as an additional source of information and validation

Table 5.7: Two search sessions with their respective sequence of queries, issued by participants during a 20 minutes learning-oriented search session in our study. The queries are divided per search stage.

Learning Topic	Stage	Search Queries
Radioactivity	1	what is radioactivity → types of radioactivity
	2	how does radioactivity occur → radioactivity resources online → uses for radiation
	3	university resources radioactivity → radioactivity explained → electromagnetic radiation
	4	radioactivity harmful → ionising and non ionising radiation → radioactivity studies physics journals
	5	radioactivity exposure → what creates gamma waves → xray radiation type → dangers of radioactivity → uses of radioactivity
Industrial Biotechnology	1	industrial biotechnology → bioplastics → bioplatics biotechnology
	2	bioplatics production enzymes → polylactic acid bioplastic
	3	industrial biotechnology → biofuels → biocatalysts
	4	industrial enzymes → biological large scale process → biological large scale processes
	5	industrial biotechnology → enzyme engineering → enzymes in biotechnology → industrial biotechnology worth → biotechnology industry → biotechnology industry market → biotechnology industry size

for participants, therefore decreasing the need for exploration as they found relevant sources of information faster. Participants in SE, on the other hand, are required to understand and validate the information they receive themselves; therefore, they do more "trial and error" in finding documents. This use of workspace awareness as validation was reflected in a few self-reports of CSE participants regarding their collaboration experience as follows:

- *I referred to partners bookmarks as a reminder of what needed to be done in the project. I also checked the search terms they entered, in case something useful came up that I did not think of.*
- *I looked at the search terms that my partner was using to get an idea of what they were searching to see if it would help my search, but that is all.*
- *I looked through their search history and bookmarks to see if there was information I had missed or to see if I was searching in the wrong direction for the information I needed.*

Overall, we only found significant differences in the number of overlapping docu-

ments and the number of initial queries, both of which is a result of workspace awareness. These results suggest that CSE participants mostly worked individually with no complex search strategies as the observed differences were a result of awareness. Although we previously found that 44% of CSE teams did coordinate, they were mostly coordinating on what the task meant, and not how to do the task.

5.2.2 Effect of Task Type on Search Behaviour

Prior studies have reported that collaborating during search resulted in the use of more diverse queries and keywords, as well as fewer overlapping documents [44, 83, 87]. In contrast, our observations suggest that collaborating results in fewer queries and more overlapping documents. We believe a difference in the search task causes this.

Our study was similar to the prior studies in that the participants needed to gather information to improve their understanding of a topic. We thus suspect that the cause was a difference in participants' familiarity with the topic. Prior studies typically assign topics related to a social issue such as gulf oil spills [83], climate change [87] or curbing population growth [44], which involve vocabularies that are more commonly known. In our case, we assigned a topic requiring uncommon knowledge; therefore participants found the task to be more difficult and ended up spending more time trying to understand the basics rather than exploring the topic domain.

To confirm the effect of topic difficulty, we conducted our study on five additional teams (CSE) in which we assigned the common easiest topic from the pretest instead of the most difficult. The average difference between the pretest score of the most difficult topic and the easiest topic was 8.75—meaning that on average, participants recognised around five more vocabulary items on the less difficult topic. We now compare the search behaviour of participants who received less difficult topic (C_{LD}) with participants who received a more difficult topic (C_{MD}).

Table 5.8: Comparison between CSE with an easy (5 pairs) and hard (25 pairs) topic (mean). **Bold** indicates statistical significance ($p \leq 0.05$).

	User			Team		
	Easy	Hard	p-value	Easy	Hard	p-value
Distinct Queries	7.40	4.92	.03	14.20	7.88	.03
Distinct Keywords	10.30	6.82	.02	16.20	9.88	.04
Distinct Documents	11.20	10.54	.94	20.80	18.28	.50
Bookmarked Documents	15.50	7.38	.001	31.00	14.76	.01
Query Overlap	-	-	-	0.60	1.96	.03
Keyword Overlap	-	-	-	4.00	3.76	.63
Document Overlap	-	-	-	1.60	2.80	.19

We can see in Table 5.8 that both the number of queries and keywords were significantly higher for C_{LD} , indicating that for a less difficult search topic, participants were able to discover relevant keywords quickly and thus do more exploration on the topic domain. However, participants in C_{LD} have one less overlapping query compared

to participants in C_{MD} , suggesting that they were less dependant on information from their partner in exploring the topic domain.

There was no significant difference in the number of viewed documents, but the number of bookmarked documents in C_{LD} was twice that of C_{MD} . We also see that for C_{LD} , the number of bookmarked documents was higher than the number of viewed documents (distinct documents). These results suggest that for a less difficult topic, participants were able to identify potentially useful documents easily and therefore they were quick to bookmark documents based on the caption and title alone.

Although we used a small sample size, the results indicate that the difficulty of the search topic has an effect on participants' exploration of the topic domain, but did not show an effect on the number of viewed documents. Additionally, we found that the average reading time for C_{LD} (34 seconds) was less than half of the time for C_{MD} (76 seconds), suggesting that since the topics were more understandable, participants spent less time on reading documents, and therefore had more time to explore the topic domain. Given that even the less difficult topics in our study are still considered a difficult topic (remember that we chose topics based on the number of unfamiliar vocabularies), more significant differences in search behaviour is expected for topics related to social issues (as in the prior studies) in which the vocabularies are more commonly known.

Concerning our study, participants' unfamiliarity with the given topic caused them to prioritise understanding the basics of the domain. We believe that given the limited task time, this encourages them to find an interesting subtopic to learn and delve deeper into, instead of covering many subtopics. In CSE, the task then becomes non-divisible as both collaborators needed to gain similar knowledge; therefore participants are likely to converge to the same area of focus, similar to the fact-finding task in [88].

5.2.3 Effect of Rapport on Search Behaviour

Rapport refers to a harmonious relationship between two or more people, resulting in better communication and a better understanding of each other¹. In Section 5.1, we have observed that there were mixed responses regarding participants' satisfaction on the group chat feature, with a low satisfaction rate being dominant. We now investigate whether bad rapport (as indicated by low chat frequency) affects the way participants search.

We used the mean chat frequency to group CSE participants into teams with low chat frequency and teams with high chat frequency. Table 5.9 shows search behaviour metrics for both groups. However, no significant differences between the two groups were found, with the means being similar. As we did not observe any significant effects on search behaviour caused by the rapport between collaborators, we believe that the effect of collaboration towards search is mainly from workspace awareness—this is consistent with our previous findings.

¹<https://en.oxforddictionaries.com/definition/rapport> (Accessed on July 2018)

Table 5.9: Search behaviour comparison (mean) between teams with low and high chat frequency (CF). CSE teams are divided into teams with **low** CF and teams with **high** CF based on the mean chat frequency.

	Low CF	High CF	p-value
Distinct Queries	4.88	4.96	.70
Distinct Keywords	7.25	6.42	.94
Distinct Documents	10.96	10.15	.72
Bookmarked Documents	7.17	7.58	.29
Query Overlap	1.83	2.08	.90
Keyword Overlap	3.83	3.69	.66
Document Overlap	2.33	3.23	.28

However, since all collaborators in our study did not personally know their partner beforehand, we cannot make any conclusions on the benefit of good rapport. There might be a positive effect on search behaviour caused by excellent rapport such as between collaborators who are friends with each other.

5.3 Learning Outcome

Our primary motivation in this research is to find out how collaboration affects learning outcome. In doing so, we replicated a knowledge assessment from similar studies along with the metrics they used, to quantify learning as elaborated in Chapter 4. In **RQ3**, we compare the effectiveness of CSE and SE concerning the learning outcome.

5.3.1 Knowledge Assessment

In our experimental setup, we implemented two types of knowledge tests: vocabulary assessment and written assessment (topic summary and paper outline).

Vocabulary Assessment

We calculated the learning gain for the vocabulary assessment by comparing results from the pre-test with the post-test. In Table 5.10, we show the average learning gain for all participants in each condition for the vocabulary assessment.

The mean ALG is 0.373 for SE and 0.308 for CSE, which means that on average, participants who worked independently learned one more vocabulary than participants who collaborated. The results for RPL—it is the same as ALG but is standardised by participants' prior knowledge—reflects a similar trend as ALG, with SE achieving a slightly higher gain. However, both measures did not show a significant difference between SE and CSE.

Additionally, we looked at the learning gain on a team level to see whether participants in CSE divided vocabulary expertise with their partner. We measured the team score by taking the average starting knowledge level and the maximum achieved knowledge level—this is to take into account specialisations within the team in which

Table 5.10: Vocabulary learning gain for both Absolute Learning Gain (ALG) and Realised Potential Learning (RPL). In addition to the metrics at user and team level, we display a metric in which we standardised the learning gains per topic before taking the average (scaled).

	User		Team		Scaled	
	ALG	RPL	ALG	RPL	ALG	RPL
SE	0.373	0.193	0.651	0.334	0.401	0.193
CSE	0.308	0.163	0.544	0.289	0.321	0.163
p-value	.36	.35	.26	.31	.47	.35

each collaborator learned different aspects of the topic. The differences, however, were also not significant. To make sure that there is no bias caused by the different learning topics, we scaled both measures using min-max scaling for each topic to standardise the score for each topic, but the differences were also not significant.

Although SE resulted in a higher average learning outcome than CSE, the high p-value means that we lack enough evidence to conclude that one method was better than the other regarding the learning outcome.

Written Assessment

For the written assessment, we did not conduct a pre-test for practical reasons. However, since we assigned each participant to the topics in which they achieved the worst result in the vocabulary pre-test, it is reasonable to assume that their starting knowledge on the topic was very low; therefore we can directly use the written assessment score to measure learning gain. We quantify the summary using sense-making measures adopted from a prior study as elaborated in Section 4.5.2.

We found that the scoring criteria for *F-Fact*, *F-State*, and *D-Qual* as defined in [105] was very clear and straightforward; therefore it is possible to rely on crowd-workers' judgement to rate the quality of participants' essays. Prior studies have shown that crowd-workers produce judgements comparable to expert annotations [48]—even for natural language tasks [89], which is similar to our case. We follow the suggested number of worker judgements [89] and assign five crowd-workers to rate each of the written assessments according to the given criteria. *T-Count* however requires knowledge of the topic domain, making it unsuitable for crowd-workers. As we also lack the appropriate expertise, we created the list of subtopics by manually comparing the content of all summaries for a given topic, and listing the subtopics that were covered by at least one summary. We then computed *T-Count* by counting the number of subtopics present in each of the summaries.

We chose the CrowdFlower platform which is especially suited for human annotation tasks. We utilised level 3 crowd-workers which consist of workers with a positive track record. We requested human judgements for 76 essays, each of which consists of three questions for *F-Fact*, *F-State*, and *D-Qual*. Additionally, we set up eight essays as test questions in order to filter out unsuitable judgements. We gave compensation equal to £0.04 for every judged essay (including the essays used as test questions). We

only accept judgements from crowd-workers with at least 70% of the test questions correct.

For *D-Qual* which measures the quality of facts in a written essay, we received judgements with an average agreement level of 67%, i.e. at least three of the five judgements were the same. Additionally, we only found two essays with divisive judgements, i.e. there was no dominant score and the two most voted scores were vastly different (e.g. two judgements for score 1, and two judgements for score 3). As for *F-Fact* and *F-State* which are numerical, the counts had an average standard deviation of 1.2 and 0.5 respectively, meaning that the judgements were not vastly different from each other. Since the majority of the judgements were conclusive and that all metrics were ordinal, we believe that the score assigned by the crowd-workers would not differ much from the actual score. Table 5.11 shows the averaged metrics for each condition.

Table 5.11: Sense-making measures calculated from post-test essays.

	SE	CSE	p-value
F-Fact	4.172	5.207	.07
F-State	5.451	6.139	.42
D-Qual	1.678	1.886	.16
T-Count	0.539	0.617	.28

Participants in both conditions wrote an average of five to six statements (*F-State*) with around four of the statements containing information related to the given topic (*F-Fact*). On average, SE participants covered half of the subtopics for their given topic, with CSE participants covering one more subtopic than SE participants. Both SE and CSE achieved an average *D-Qual* score of just below 2, which means that participants from both conditions wrote useful information related to the assigned learning topic. Overall, participants in CSE achieved a slightly higher average score for each metrics; however, the difference was not statistically significant. Additionally, we tried to linearly combine the different metrics into one single score, but the difference was also not significant.

5.3.2 Implicit Learning Metrics

We have plotted both the accumulated keyword complexity and document complexity in each of the five stages of the search session (in a similar way to Figure 5.2).

In figure 5.3, we see that the keyword complexity starts low and increases over time, that is, participants initially use common keywords at the start of the session (e.g. directly querying the learning topic) and over time used less common keywords in their search. At the user level, keyword complexity in both conditions did not show a significant difference, although different trends were present. Participants in CSE started with more common keywords, but over time achieved a similar keyword complexity to participants in SE. At the team level, the trends in both conditions are similar, but with SE consistently on top. However, using a Mann-Whitney U test, we found that the difference was not significant.

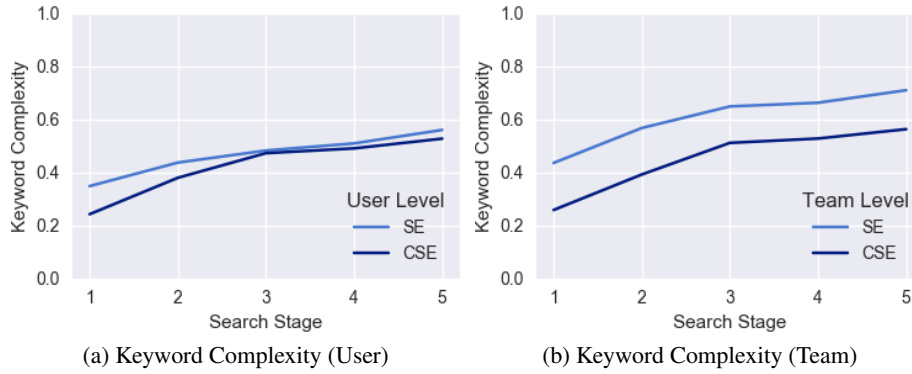


Figure 5.3: Keyword complexity over each search stages.

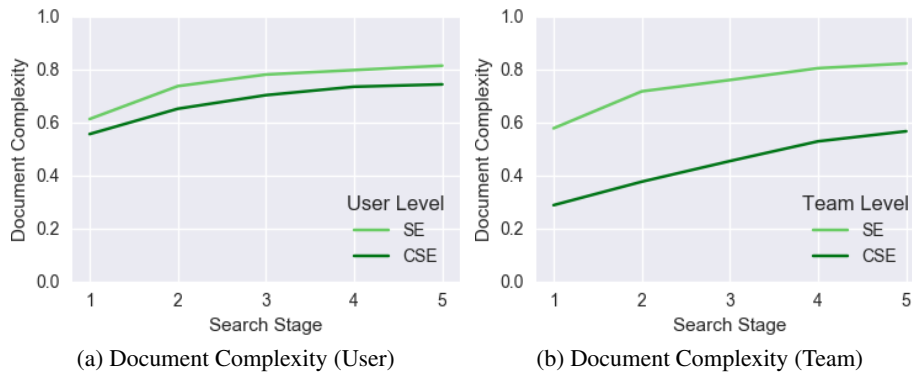


Figure 5.4: Document complexity over each search stages.

The overall trend for document complexity is similar to query complexity in which participants started with accessing more common documents, and over time moved on to less common documents as can be seen in Figure 5.4. However, we can see a more evident divide between SE and CSE at both user level and team level. Using the Mann Whitney U test, we found that the difference is significant at $p \leq 0.05$ for both user level and team level, concluding that collaboration resulted with participants accessing more common documents than participants who worked individually. In another way, this can be seen as participants in CSE reading documents that are considered resourceful by most other participants.

The results for both keyword complexity and document complexity was the opposite of results reported in [13] where collaboration resulted in a consistently higher query and document complexity. We believe that this difference is again because of the different search task assigned to the participants.

Chi et al. [13] used data from two different tasks: creating a report on social media (an information gathering task) and planning a trip to Helsinki (decision making task). Both tasks assigned a general topic to participants and asked them to gather more information. Because the topics were familiar to participants, they did not need much effort to process information on a single document and therefore can afford to go for more breadth in searching. In contrast, we assign learning topics which are un-

familiar to participants and involves uncommon terms. Therefore, participants in our study prioritise depth more by finding documents that are comprehensive and spending more time understanding the content. The focus of collaboration in our search task is therefore in suggesting useful learning materials to each other, resulting in a lower document complexity.

To evaluate its suitability as a learning indicator, we tested the correlation of both document and keyword complexity with RPL at the user level. Both measures had low correlation with RPL (0.02 for keyword complexity and 0.12 for document complexity), indicating that the metrics do not accurately reflect participants' vocabulary knowledge.

5.3.3 Effect of Collaboration on Learning

As we did not observe a significant difference in vocabulary learning gain between SE and CSE, we now look into whether specific collaboration actions have a negative effect on knowledge gain. It has been shown in [24] that in collaborating, there is an additional cognitive load which might reduce the expected benefit of collaboration. An example in our case is that while discussions can help in increasing understanding regarding a topic, it also takes time and detracts participants from actual searching.

We investigate the effect of collaboration by comparing the RPL between participants with more collaboration and participants with less collaboration, as measured by the value of each collaboration metric. For a collaboration metric, we take its mean within CSE teams and use it to group the population into teams with high collaboration activities and teams with low collaboration activities. Table 5.12 shows the average RPL of both low and high collaboration groups for each of the metrics.

Table 5.12: Mean learning gain (RPL) of **low** and **high** collaboration groups. The mean of each collaboration metric is used to categorise each group into either low or high collaboration. **Bold** indicates statistical significance ($p \leq 0.05$).

		RPL		p-value
		Low	High	
Chat Frequency	User	0.119	0.204	.042
	Team	0.215	0.359	.047
Chat Ratio	User	0.122	0.191	.179
	Team	0.214	0.340	.075
Query Overlap	User	0.139	0.184	.423
	Team	0.258	0.319	.513
Keyword Overlap	User	0.132	0.209	.095
	Team	0.243	0.360	.126
Document Overlap	User	0.142	0.179	.161
	Team	0.269	0.306	.395

There was a significant difference in RPL between participants with a lower chat frequency compared to participants with a higher chat frequency. The result was con-

sistent at both the user and team level, concluding that more frequent communication between collaborators results with higher achieved learning. We can conclude that while more frequent communication between participants might take time away from searching, it has a positive impact towards learning performance. One possible reason is that frequent communication results in better rapport, which leads to both collaborators being more motivated to learn and therefore taking the task more seriously. However, the chat ratio was not statistically significant, indicating that while an unbalanced chat ratio might affect the quality of discourse, more frequent discourse is more important as there won't be any meaningful discourse when there are too few chats.

In Figure 5.5, we look into the connection between chat frequency and chat ratio regarding the achieved learning. In general, there is no direct correlation between both chat frequency and chat ratio towards learning. Participants without much discourse can still achieve an average score; however, participants with the highest learning gain displayed both a high chat frequency and a balanced chat ratio. We can say that on average, more discourse results in a slightly higher learning outcome; however, a high-quality discourse has the potential to result in significantly higher learning outcomes.

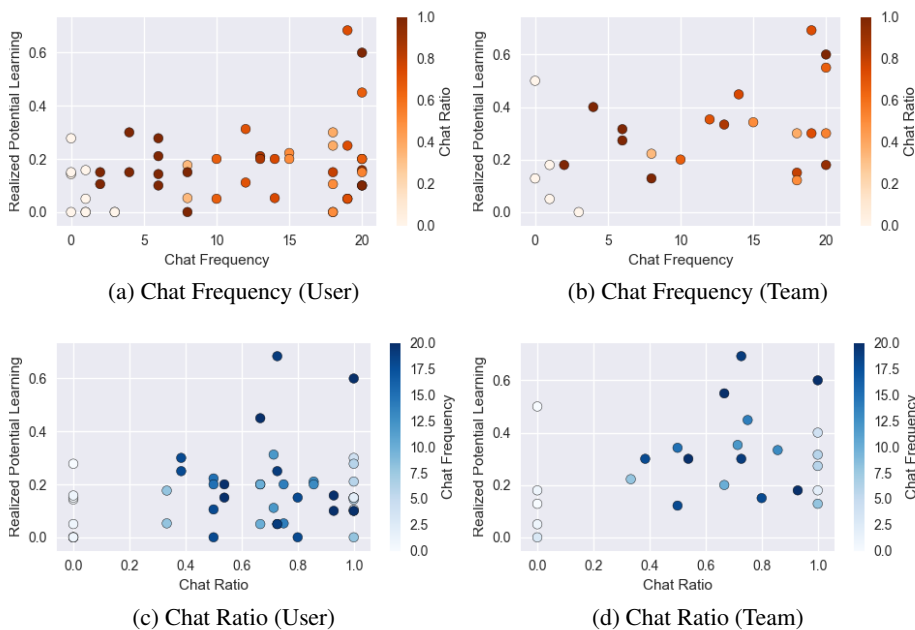


Figure 5.5: Connection between chat actions and realised potential learning. Each point is a participant in CSE.

To get a better understanding of the effect of communication towards learning, we now look into how different types of chat interactions affect the achieved learning gain. In doing so, we used our previous data from Section 5.1 to group the teams based on the existence of specific types of chat interactions. In Table 5.13 we can see that the only significant difference was on teams displaying verbal immediacy—members of teams that were able to build a rapport (as indicated by a display of closeness) achieved twice the learning gain of participants from teams that had a bad rapport. This result confirms our previous hypothesis that the cause of high learning gains for teams with

high chat frequency is the excellent rapport that they built. We believe that the role of good rapport in learning (at least in our case) is to motivate the collaborators to learn more seriously.

Table 5.13: Mean learning gain (RPL) of groups that displayed certain type of chat interactions (**True**) compared to groups that did not (**False**). **Bold** indicates statistical significance ($p \leq 0.05$).

		RPL		p-value
		True	False	
Verbal Immediacy	User	0.200	0.107	.028
	Team	0.350	0.199	.026
Coordination	User	0.175	0.154	.128
	Team	0.282	0.296	1.00
Resource Suggestion	User	0.199	0.146	.287
	Team	0.322	0.275	.619
Topic Discussion	User	0.159	0.164	.795
	Team	0.305	0.284	.926

Chapter 6

Discussions and Conclusions

The primary goal of this research was to investigate the effects of collaboration on a learning-oriented search task. We are interested in investigating the effects of collaboration in terms of search behaviour and learning performance. We have conducted a comparative study in which we assigned two different search conditions to study participants: single user search (SE) and collaborative search (CSE). In this section, we reflect on the results and discuss our findings. Furthermore, we elaborate on the limitations of our study, and we present future research directions based on our findings.

6.1 Discussions

Effect of Collaboration on Search Behaviour

We found that participants in CSE bookmarked fewer documents compared to SE at both the user and team level. However, we argue that this is mostly the result of a limitation enforced by SearchX as opposed to an effect of collaboration. We provided a shared bookmarks list in which no two collaborators can bookmark the same document, resulting in fewer options for bookmarking in the same SERP. Additionally, as collaborators received a combined list of their bookmarks and their partner's bookmarks, participants might be demotivated from bookmarking additional documents once the bookmarks list gets long. Therefore, we cannot attribute this decrease in bookmarked documents as an effect of collaboration.

Another significant difference was in the number of overlapping documents with groups in CSE having on average four more documents in common between their members compared to any two SE participants. We attribute this difference to the awareness of collaborators' bookmarked documents motivating participants to check out what their partner has marked as useful.

We did not detect any other significant differences in the way participants search between CSE and SE. However, if we look into the number of search actions over time, we found that SE participants issued one more query and a lot more bookmarks at the start of the search session compared to CSE participants. The number of viewed documents though were not significantly different. These results suggest that CSE participants required less exploration at the beginning of the session in order to get a basic understanding of the topic compared to SE participants. We argue that this behaviour is because CSE participants received help in understanding the topic through workspace

awareness features (i.e. shared query history and bookmarks), as the features provide feedback on the validity of their search efforts. This feedback can help CSE participants to more quickly identify documents and queries which are potentially useful for learning compared to SE participants.

To summarise, we observed the positive effect of workspace awareness towards participants' search behaviour. We believe that the additional information provided by the shared bookmarks and query history helped participants in selecting useful documents and validating their search efforts to keep them on track. The self-reports also reflect this conclusion with CSE participants reporting no problems related to the task setup and fewer problems in general. However, the fact that only workspace awareness affected search behaviour and that there were no effects of rapport towards search behaviour suggests that CSE participants mostly worked independently, as the metrics do not reflect a division of labour.

Effect of Collaboration on Learning

For the vocabulary assessment, the mean values suggest that CSE participants achieved slightly worse learning gain compared to SE participants. The opposite was observed in the written assessment, where CSE participants achieved slightly better learning gain compared to SE participants regarding their understanding of the topic. These results suggest that collaborating during search helped participants to understand the topic better as a whole, but at the expense of a slightly worse understanding of specific vocabularies. The differences though were not statistically significant; therefore we are unable to conclude these effects as a result of collaboration. However, since the differences were consistent at both user and team level and consistent across the various metrics, conducting the study at a larger sample size can potentially confirm our observations.

Since there were indications that collaboration resulted in worse vocabulary learning gain, we looked further into the effect of specific collaboration actions on learning gain as we suspect there might be a negative effect caused by the additional cognitive cost of collaboration. The results however indicate the opposite in that a higher chat frequency between collaborators leads to higher overall learning gain. Although frequent communication between collaborators take away time from searching and learning, it is actually beneficial to learning. We looked further for into how having a higher chat frequency affects search behaviour, but did not find any significant difference. This shows that the benefit of frequent communication is not in promoting specific search behaviour, but rather in affecting how participants perceive the task. Frequent communication is a sign of good rapport, which might motivate participants to learn more seriously.

From a closer look at the effect of specific chat interactions on learning, we found that teams displaying verbal immediacy behaviours—and thus better social presence [94]—achieved a higher learning gain compared to those who did not. Although we cannot determine the exact cause of high communication frequency between collaborators, we can say that a good rapport between collaborators in general has a positive effect on learning. This observation is consistent with [71] which states that social presence is an essential factor for a successful online learning experience.

Effect of Search Task

Some of our findings were inconsistent with those reported in prior studies. Concerning search behaviour, we observed that CSE participants issued fewer queries at the start of the session and had more document overlap; however prior studies found that CSE resulted in more and diverse queries as well as fewer document redundancy [44, 83, 87]. Additionally, we found that participants in CSE accessed more commonly found documents, whereas prior research concluded that participants in CSE accessed less commonly found documents and queries [13]. These discrepancies in the effect of collaboration are likely to be a result of the different experimental setup between our study and prior studies.

One possibility is that prior research conducted their study in a lab environment where collaborators have a similar background and can personally build rapport, whereas our study relies on crowd-workers who have diverse backgrounds and do not meet each other physically at all. Our analysis of the chat interactions, however, suggest that the majority of CSE teams were able to achieve meaningful discourse and social presence, indicating that bad rapport is not an issue in our study. Additionally, we did not observe a difference in search behaviour between groups with good and bad rapport.

Another possibility we investigated was the difference in the search task. Both our study and prior studies [44, 83, 87, 13] required participants to discover information and learn more about a topic. Since there are few differences in the instructions, we suspect that the leading cause of the inconsistencies is on the difficulty of the search topic. Prior studies generally assign a social topic which does not require much domain specific knowledge, whereas, in our study, we assign participants with the topic in which they the least familiarity. We suspect that this difference in the required knowledge to understand documents related to the assigned topic is the reason behind the differences.

From an analysis involving five additional CSE teams, we found that participants with the less difficult topic issued more queries, used more keywords, bookmarked more documents, had less overlapping queries, and spent less time reading the found documents. These results suggest that a less difficult search topic offers more opportunities for participants to explore the topic domain compared to participants with a more difficult-to-understand topic, as indicated by a higher number of queries and keywords, but with a similar number of viewed documents.

Although we cannot confirm the effects of topic difficulty towards collaboration, we found indications that topic difficulty has a significant effect on participants' exploratory search behaviour. An implication of this is that research on exploratory search cannot be compared directly to each other when the difficulty of the search task is different. More research is needed to further understand the effects of search task difficulty towards exploratory search behaviour.

6.2 Limitations

We acknowledge some limitations in our study that affect the reliability of the results we presented.

Design of the knowledge assessment. In our study setup, we had CSE participants

work together with their partner during the search session, but not during the post-test. We put this restriction to make the assessment directly comparable between SE and CSE. However, this setup can potentially reduce the benefits of collaboration, as [56] argued that collaboration is meaningless without the power to implement final decisions. For the information gathering task in prior studies, this condition is easier to satisfy as the researchers computed search performance from both collaborators' bookmarked/rated documents. Given that in our study we did not find statistically significant differences in learning gain, we believe that designing a suitable method for collaborators to demonstrate their learning as a group would better reflect the synergistic effect of collaboration.

Crowd-sourcing the study. For practicality, we chose to deploy our study through a crowd-sourcing platform, which has more limitations compared to a controlled lab study. One limitation is that it is harder to condition study participants for collaboration as participants do not meet physically at all—which potentially have implications on the quality of social presence within collaborative groups, reducing the output of collaboration [31]. In our study, we have minimised this limitation by priming participants for collaboration right before the start of the search session, which has resulted in the majority of participants producing meaningful discourse. However, we did not measure the effectiveness of the method, and thus the lack of social presence might still affect the reliability of our results.

Another limitation is on the task time. The less controlled nature of crowd-sourced studies compared to lab-based studies means that it is unsuitable for long task times as we cannot guarantee participants will still be focused. Because of this, we assigned a task time which can be considered as short compared to similar studies. However, Shah and González-Ibáñez [82] has shown that collaborations lasting longer and done over multiple sessions may produce significantly different results and user experiences. Moreover, our study assigned a difficult learning topic which, as we have shown, takes more time for participants to understand, therefore assigning a longer task time may produce different observations.

The amount of study participants. For the majority of our tested conditions, we believe that the number of participants is sufficient as it is comparable to similar studies. For the additional condition of CSE with an easy topic, we were only able to conduct the study for 10 participants (5 pairs) because of resource limitations. However, our 'easy' topic is still difficult as we purposely chose topics which have uncommon vocabularies; therefore we believe that we can expect more significant differences from assigning well-known topics similar to prior studies.

6.3 Future Work

There is still much room for improvements in our study to explore in future research.

- **The effect of topic difficulty on exploratory search.** Our results suggest that a less difficult topic allowed participants to demonstrate more exploratory search behaviours. As a result, we found that our findings were inconsistent with the reported benefits of collaboration in prior studies [44, 83, 87]. A potential research

direction would be investigating further how different search task difficulties impact exploratory search behaviour.

- **Explore ways of building good rapport in remote collaborative studies.** A good social presence between peers is essential in building a successful online learning experience [71]. Our results suggest that for a crowd-sourced collaborative study requiring above average cognitive efforts (i.e. online learning), a good rapport between collaborators helps in keeping participants motivated throughout the study, resulting in better task performance. Another line of future research should explore ways of building good rapport in collaborative studies where collaborators do not know each other and are remotely located. Insight on methods to build up online rapport will be valuable for future online collaborative studies.
- **Observe learning over longer sessions.** Recently, a study has shown the possibility of conducting longitudinal studies through crowd-sourcing in order to observe long-term learning [96]. Additionally, we have shown that for a difficult search topic, participants demonstrated less exploratory search behaviours. Observing learning over longer sessions or even multiple sessions can potentially give very different insights. It can be argued that a search session involving a longer task time or multiple sessions are more typical than our study setup; therefore research on this would give a better understanding of real learning circumstances.

6.4 Conclusion

With the high rate of internet penetration in most countries nowadays, Web search has become a convenient option for seeking information related to learning. As people started to depend on searching the Web to fulfil their learning needs, the field of search as learning has emerged with the agenda of investigating how to adapt search engines to support human learning better. One area in which we aimed to explore concerning this is how collaboration affects the outcome of a learning oriented-search session—as collaboration, in general, has been shown to be beneficial. To this end, we have engineered a collaborative search system and deployed it on a crowd-sourced study.

In Chapter 3, we elaborated on the design decisions we made in engineering our collaborative search system—SearchX. We designed SearchX with the goal of accommodating collaboration during a search session, specifically in a way that would make deployment of user studies on a crowd-sourcing platform feasible. From our analysis of participants' feedback, we conclude that SearchX was able to accommodate collaboration during a search session by providing mediation features for workspace awareness and direct communication. Although there was mixed feedback on the usefulness of the chat feature, an analysis of chat interactions revealed that the majority of participants were still able to achieve meaningful discourse.

Our analysis of user interaction logs reveals that collaborating during a learning-oriented search session did not result in much differences in search behaviour compared to single-user search. A more in-depth analysis of participants' search actions revealed that workspace awareness helped collaborative participants in selecting use-

ful documents and in understanding the task at the beginning of the search session. The results though were not consistent with prior studies, and from an additional pilot study, we were able to conclude that the difficulty of the search topic had a significant effect on exploratory search behaviours. We observed that participants with a less difficult learning topic needed less time to understand the content of documents and therefore were able to explore the topic domain more.

For the learning outcome, we see indications of collaboration resulting with a better understanding of the topic domain but is detrimental to vocabulary learning. However, our results were not conclusive as the observed differences were not statistically significant. From an analysis of the effects of specific collaborative actions, we found that frequent communication between collaborators resulted in higher learning gain. We identified the cause of this higher learning gain to be mainly because of a good rapport between collaborators.

To summarise, we found indications that collaboration has a positive effect on gaining understanding of the topic, and that social presence is relevant in learning through search. Although we did not observe the same benefits of collaboration on search behaviour as reported in prior research, we did observe a similar trend to prior research on collaborative learning: good collaboration promotes higher-order learning.

Bibliography

- [1] Maristella Agosti, Norbert Fuhr, Elaine Toms, and Pertti Vakkari. Evaluation methodologies in information retrieval dagstuhl seminar 13441. In *ACM SIGIR Forum*, volume 48, pages 36–41. ACM, 2014.
- [2] Zehra Akyol and D Randy Garrison. The development of a community of inquiry over time in an online course: Understanding the progression and integration of social, cognitive and teaching presence. *Journal of Asynchronous Learning Networks*, 12:3–22, 2008.
- [3] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. In *ACM SIGIR Forum*, volume 46, pages 2–32. ACM, 2012.
- [4] Saleema Amershi and Meredith Ringel Morris. Cosearch: A system for co-located collaborative web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1647–1656. ACM, 2008. ISBN 978-1-60558-011-1.
- [5] Kumaripaba Athukorala, Dorota Glowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. Is exploratory search different? a comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology*, 67(11):2635–2651, 2016.
- [6] Kumaripaba Athukorala, Alan Medlar, Antti Oulasvirta, Giulio Jacucci, and Dorota Glowacka. Beyond relevance: Adapting exploration/exploitation in information retrieval. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 359–369. ACM, 2016.
- [7] Leif Azzopardi, Jeremy Pickens, Chirag Shah, Laure Soulier, and Lynda Tamine. Second international workshop on the evaluation of collaborative information seeking and retrieval (ecol'17). In *CHIIR '17*, pages 429–431, 2017.
- [8] Peter Bailey, Liwei Chen, Scott Grosenick, Li Jiang, Yan Li, Paul Reinholdtsen, Charles Salada, Haidong Wang, and Sandy Wong. User task understanding: a web search engine perspective. In *NII Shonan Meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems, Kanagawa, Japan*, 2012.

- [9] Martynas Buivys and Leif Azzopardi. Pienapple search: An integrated search interface to support finding, refinding and sharing. In *ASIST '16*, pages 122:1–122:5, 2016.
- [10] Robert Capra, Annie T. Chen, Katie Hawthorne, Jaime Arguello, Lee Shaw, and Gary Marchionini. Design and evaluation of a system to support collaborative search. *ASIST*, 49(1):1–10, 2012.
- [11] Robert Capra, Annie T Chen, Evonne McArthur, and Natalie Davis. Searcher actions and strategies in asynchronous collaborative search. *Proceedings of the Association for Information Science and Technology*, 50(1):1–10, 2013.
- [12] Annie T Chen, Robert Capra, and Wan-Ching Wu. An investigation of the effects of awareness and task orientation on collaborative search. *Proceedings of the Association for Information Science and Technology*, 51(1):1–10, 2014.
- [13] Yu Chi, Shuguang Han, Daqing He, and Rui Meng. Exploring knowledge learning in collaborative information seeking process. In *CEUR Workshop Proceedings*, volume 1647, 2016.
- [14] Elizabeth G Cohen. Restructuring the classroom: Conditions for productive small groups. *Review of educational research*, 64(1):1–35, 1994.
- [15] Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. Inferring user knowledge level from eye movement patterns. *Information Processing & Management*, 49(5):1075–1091, 2013.
- [16] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 163–172. ACM, 2016.
- [17] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. Search as learning (dagstuhl seminar 17092). In *Dagstuhl Reports*, volume 7. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.
- [18] David D Curtis and Michael J Lawson. Exploring collaborative online learning. *Journal of Asynchronous learning networks*, 5(1):21–34, 2001.
- [19] Peter J Denning and Peter Yahlkovsky. Getting to we. *Communications of the ACM*, 51(4):19–24, 2008.
- [20] Pierre Dillenbourg. *Collaborative learning: Cognitive and computational approaches. advances in learning and instruction series*. ERIC, 1999.
- [21] Katherine A Dougherty Stahl and Marco A Bravo. Contemporary classroom vocabulary assessment for content areas. *The Reading Teacher*, 63(7):566–578, 2010.

- [22] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 223–232. ACM, 2014.
- [23] Carsten Eickhoff, Jacek Gwizdka, Claudia Hauff, and Jiyin He. Introduction to the special issue on search as learning. *Information Retrieval Journal*, 20(5): 399–402, 2017.
- [24] Raya Fidel, Harry Bruce, Annelise M Pejtersen, Susan Dumais, Jonathan Grudin, and Steven Poltrock. Collaborative information retrieval (cir). *New Rev. Inf. Behav. Res.*, 1(January):235–247, 2000.
- [25] Colum Foley and Alan F. Smeaton. Synchronous collaborative information retrieval: Techniques and evaluation. In *ECIR '09*, pages 42–53, 2009.
- [26] Colum Foley and Alan F. Smeaton. Division of labour and sharing of knowledge for synchronous collaborative information retrieval. *IPM*, 46(6):762–772, 2010.
- [27] Jonathan Foster. Collaborative information seeking and retrieval. *Annual Review of Information Science and Technology*, 40(1):329–356, December 2006. ISSN 0066-4200.
- [28] D Randy Garrison. Online community of inquiry review: Social, cognitive, and teaching presence issues. *Journal of Asynchronous Learning Networks*, 11(1): 61–72, 2007.
- [29] D Randy Garrison and Martha Cleveland-Innes. Facilitating cognitive presence in online learning: Interaction is not enough. *The American journal of distance education*, 19(3):133–148, 2005.
- [30] D Randy Garrison, Terry Anderson, and Walter Archer. Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of distance education*, 15(1):7–23, 2001.
- [31] Larry Gilbert and David R Moore. Building interactivity into web courses: Tools for social and instructional interactions. *Educational Technology*, 38(3): 29–35, 1998.
- [32] Anuradha A Gokhale. Collaborative learning enhances critical thinking. 1995.
- [33] Gene Golovchinsky, John Adcock, Jeremy Pickens, Pernilla Qvarfordt, and Maribeth Back. Cerchiamo: a collaborative exploratory search tool. CSCW, pages 8–12, 2008.
- [34] Gene Golovchinsky, Jeremy Pickens, and Maribeth Back. A taxonomy of collaboration in online information seeking. JCDL, 2008.
- [35] Gene Golovchinsky, Abdigani Diriye, and Tony Dunnigan. The future is in the past: Designing for exploratory search. In *IIIX '12*, pages 52–61, 2012.

- [36] Roberto González-Ibáñez, Muge Haseki, and Chirag Shah. Understanding effects of time and proximity on collaboration: implications for technologies to support collaborative information seeking. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 1805–1810. ACM, 2012.
- [37] Roberto González-Ibáñez, Chirag Shah, and Ryen W White. Pseudo-collaboration as a method to perform selective algorithmic mediation in collaborative ir systems. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–4, 2012.
- [38] Jacek Gwizdka, Preben Hansen, Claudia Hauff, Jiyin He, and Noriko Kando. Search as learning (sal) workshop 2016. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1249–1250. ACM, 2016.
- [39] Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333, 2002.
- [40] Ahmed Hassan Awadallah, Ryen W White, Patrick Pantel, Susan T Dumais, and Yi-Min Wang. Supporting complex search tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 829–838. ACM, 2014.
- [41] Jannica Heinström. Broad exploration or precise specificity: Two basic information seeking patterns among students. *Journal of the Association for Information Science and Technology*, 57(11):1440–1450, 2006.
- [42] David W Johnson. Student-student interaction: The neglected variable in education. *Educational researcher*, 10(1):5–10, 1981.
- [43] David W Johnson, Roger T Johnson, and Karl A Smith. Cooperative learning returns to college what evidence is there that it works? *Change: the magazine of higher learning*, 30(4):26–35, 1998.
- [44] Hideo Joho, David Hannah, and Joemon M Jose. Comparing collaborative and independent search in a recall-oriented task. In *Proceedings of the second international symposium on Information interaction in context*, pages 89–96. ACM, 2008.
- [45] David Jonassen. Reconciling a human cognitive architecture. *Constructivist instruction: Success or failure*, pages 13–33, 2009.
- [46] Ryan Kelly and Stephen J. Payne. Collaborative web search in context: A study of tool use in everyday tasks. In *CSCW '14*, pages 807–819, 2014.
- [47] Walter Kintsch. *Comprehension: A paradigm for cognition*. Cambridge university press, 1998.
- [48] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

- [49] Yamuna Krishnamurthy, Kien Pham, Aécio Santos, and Juliana Freire. Interactive exploration for domain discovery on the web. *Proc. of KDD IDEA*, 2016.
- [50] Hannarin Kruajirayu, Ake Tangsomboon, and Teerapong Leelanupab. Cozpace: a proposal for collaborative web search for sharing search records and interactions. In *Student Project Conference (ICT-ISPC), 2014 Third ICT International*, pages 165–168. IEEE, 2014.
- [51] Bill Kules and Robert Capra. Influence of training and stage of search on gaze behavior in a library catalog faceted search interface. *Journal of the Association for Information Science and Technology*, 63(1):114–138, 2012.
- [52] Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. What do exploratory searchers look at in a faceted search interface? In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 313–322. ACM, 2009.
- [53] Ray R Larson. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology*, 61(4):852–853, 2010.
- [54] Teerapong Leelanupab, Hannarin Kruajirayu, and Nont Kanungsukkasem. Snapboard: A shared space of visual snippets - a study in individual and asynchronous collaborative web search. In *Information Retrieval Technology*, pages 161–173. Springer International Publishing, 2015. ISBN 978-3-319-28940-3.
- [55] Olivier Liechti. Awareness and the www: An overview. *SIGGROUP Bulletin*, 21(3):3–12, December 2000. ISSN 2372-7403.
- [56] Scott London. Collaboration and community. *Richmond, VA, Pew Partnership for Civic Change, University of Richmond*, 1995.
- [57] Jens-Erik Mai. *Looking for information: A survey of research on information seeking, needs, and behavior*. Emerald Group Publishing, 2016.
- [58] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.
- [59] Paul McCoubrie. Improving the fairness of multiple-choice questions: a literature review. *Medical teacher*, 26(8):709–712, 2004.
- [60] Matthew Mitsui, Jiqun Liu, and Chirag Shah. Coagmento: Past, present, and future of an individual and collaborative information seeking platform. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 325–328. ACM, 2018.
- [61] Michael G Moore. Three types of interaction. In *The American Journal of Distance Education*. Citeseer, 1992.
- [62] Felipe Moraes, Sindunuraga Rikarno Putra, and Claudia Hauff. Contrasting search as a learning activity with instructor-designed learning. In *Proceedings of the CIKM International Conference on Information and Knowledge Management, CIKM '18*. ACM, 2018.

- [63] Meredith Ringel Morris. Collaborating alone and together: Investigating persistent and multi-user web search activities. In *Proceedings of international ACM SIGIR conference on research and development in information retrieval (SIGIR 2007)*, pages 23–27, 2007.
- [64] Meredith Ringel Morris. Collaborative search revisited. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1181–1192. ACM, 2013.
- [65] Meredith Ringel Morris and Eric Horvitz. Searchtogether: An interface for collaborative web search. In *UIST '07*, pages 3–12, 2007.
- [66] Meredith Ringel Morris, Jarrod Lombardo, and Daniel Wigdor. Wesearch: supporting collaborative search and sensemaking on a tabletop display. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 401–410. ACM, 2010.
- [67] David Nicholas, Ian Rowlands, David Clark, and Peter Williams. Google generation ii: web behaviour experiments with the bbc. In *Aslib proceedings*, volume 63, pages 28–45. Emerald Group Publishing Limited, 2011.
- [68] T Sima Paribakht and Marjorie Bingham Wesche. Reading comprehension and second language development in a comprehension-based esl program. *TESL Canada journal*, 11(1):09–29, 1993.
- [69] Sharoda A. Paul and Meredith Ringel Morris. Cosense: Enhancing sensemaking for collaborative web search. In *CHI '09*, pages 1771–1780, 2009.
- [70] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153 – 163, 2017. ISSN 0022-1031.
- [71] Anthony G Picciano. Beyond student perceptions: Issues of interaction, presence, and performance in an online course. *Journal of Asynchronous learning networks*, 6(1):21–40, 2002.
- [72] Jeremy Pickens, Gene Golovchinsky, Chirag Shah, Pernilla Qvarfordt, and Maribeth Back. Algorithmic mediation for collaborative exploratory search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 315–322. ACM, 2008.
- [73] Jeremy Pickens, Gene Golovchinsky, and Meredith Ringel Morris. Proceedings of 1st international workshop on collaborative information seeking. *CoRR*, abs/0908.0583, 2009.
- [74] Steven Poltrock, Jonathan Grudin, Susan Dumais, Raya Fidel, Harry Bruce, and Annelise Mark Pejtersen. Information seeking and sharing in design teams. In *2003 International ACM SIGGROUP Conference on Supporting Group Work, GROUP '03*, pages 239–247. ACM, 2003.

- [75] Michael Prince. Does active learning work? a review of the research. *Journal of engineering education*, 93(3):223–231, 2004.
- [76] Karthik Raman, Paul N Bennett, and Kevyn Collins-Thompson. Toward whole-session relevance: exploring intrinsic diversity in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 463–472. ACM, 2013.
- [77] Nicola J Reavley, Andrew J Mackinnon, Amy J Morgan, Mario Alvarez-Jimenez, Sarah E Hetrick, Eoin Killackey, Barnaby Nelson, Rosemary Purcell, Marie BH Yap, and Anthony F Jorm. Quality of information sources about mental disorders: a comparison of wikipedia with centrally controlled web and printed sources. *Psychological medicine*, 42(8):1753–1762, 2012.
- [78] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science*, 42(1):19–34, 2016.
- [79] Ian Rowlands, David Nicholas, Peter Williams, Paul Huntington, Maggie Fieldhouse, Barrie Gunter, Richard Withey, Hamid R Jamali, Tom Dobrowolski, and Carol Tenopir. The google generation: the information behaviour of the researcher of the future. In *Aslib proceedings*, volume 60, pages 290–310. Emerald Group Publishing Limited, 2008.
- [80] Tuukka Ruotsalo, Jaakko Peltonen, Manuel JA Eugster, Dorota Głowacka, Aki Reijonen, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. Scinet: Interactive intent modeling for information discovery. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1043–1044. ACM, 2015.
- [81] Marlene Scardamalia and Carl Bereiter. Computer support for knowledge-building communities. *The journal of the learning sciences*, 3(3):265–283, 1994.
- [82] Chirag Shah and Roberto González-Ibáñez. Exploring information seeking processes in collaborative search tasks. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–7, 2010.
- [83] Chirag Shah and Roberto González-Ibáñez. Evaluating the synergic effect of collaboration in information seeking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 913–922. ACM, 2011.
- [84] Chirag Shah and Gary Marchionini. Awareness in collaborative information seeking. *ASIST*, 61(10):1970–1986, 2010.
- [85] Chirag Shah, Gary Marchionini, and Diane Kelly. Learning design principles for a collaborative information seeking system. In *CHI EA '09*, pages 3419–3424, 2009.

- [86] Chirag Shah, Jeremy Pickens, and Gene Golovchinsky. Role-based results redistribution for collaborative information retrieval. *Information processing & management*, 46(6):773–781, 2010.
- [87] Chirag Shah, Chathra Hendaheewa, and Roberto González-Ibáñez. Two’s company, but three’s no crowd: Evaluating exploratory web search for individuals and teams. *Aslib Journal of Information Management*, 67(6):636–662, 2015.
- [88] Chirag Shah, Chathra Hendaheewa, and Roberto González-Ibáñez. Two’s not always company: collaborative information seeking across task types. *Aslib Journal of Information Management*, 69(1):22–35, 2017.
- [89] Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.
- [90] Laure Soulier and Lynda Tamine. On the collaboration support in information retrieval. *ACM Computing Surveys*, 50(4):51:1–51:34, August 2017. ISSN 0360-0300.
- [91] Laure Soulier, Chirag Shah, and Lynda Tamine. User-driven system-mediated collaborative information retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 485–494. ACM, 2014.
- [92] Laure Soulier, Lynda Tamine, and Chirag Shah. Minerank: Leveraging users’ latent roles for unsupervised collaborative information retrieval. *Information Processing & Management*, 52(6):1122–1141, 2016.
- [93] Leonard Springer, Mary Elizabeth Stanne, and Samuel S Donovan. Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of educational research*, 69(1):21–51, 1999.
- [94] Karen Swan. Building learning communities in online courses: The importance of interaction. *Education, Communication & Information*, 2(1):23–49, 2002.
- [95] Rohail Syed and Kevyn Collins-Thompson. Retrieval algorithms optimized for human learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 555–564. ACM, 2017.
- [96] Rohail Syed and Kevyn Collins-Thompson. Exploring document retrieval features associated with improved short-and long-term vocabulary learning outcomes. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 191–200. ACM, 2018.

- [97] Lynda Tamine and Laure Soulier. Understanding the impact of the role factor in collaborative information retrieval. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 43–52. ACM, 2015.
- [98] Yihan Tao and Anastasios Tombros. An exploratory study of sensemaking in collaborative information seeking. In *European Conference on Information Retrieval*, pages 26–37. Springer, 2013.
- [99] Jaime Teevan, Meredith Ringel Morris, and Shiri Azenkot. Supporting interpersonal interaction during collaborative mobile search. *Computer*, 47(3):54–57, 2014.
- [100] Michael B Twidale, David M Nichols, and Chris D Paice. Browsing is a collaborative process. *Information Processing & Management*, 33(6):761–783, 1997.
- [101] Pertti Vakkari. Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18, 2016.
- [102] Marjorie Wesche and T Sima Paribakht. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian modern language review*, 53(1): 13–40, 1996.
- [103] Ryen W White and Resa A Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–98, 2009.
- [104] Ryen W White, Susan T Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining*, pages 132–141. ACM, 2009.
- [105] Mathew J Wilson and Max L Wilson. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the Association for Information Science and Technology*, 64(2):291–306, 2013.
- [106] Yusuke Yamamoto and Takehiro Yamamoto. Query priming for promoting critical thinking in web search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pages 12–21. ACM, 2018.
- [107] Ke-Thia Yao, Robert Neches, In-Young Ko, Ragy Eleish, and Sameer Abhinkar. Synchronous and asynchronous collaborative information space analysis tools. In *1999 International Workshops on Parallel Processing*, pages 74–79. IEEE, 1999.
- [108] Zhen Yue, Shuguang Han, and Daqing He. Search tactics in collaborative exploratory web search. In *HCIR 2012*, 2012.
- [109] Zhen Yue, Jiepu Jiang, Shuguang Han, and Daqing He. Where do the query terms come from?: an analysis of query reformulation in collaborative web search. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2595–2598. ACM, 2012.

- [110] Zhen Yue, Shuguang Han, and Daqing He. Modeling search processes using hidden states in collaborative exploratory web search. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 820–830. ACM, 2014.
- [111] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. Predicting users’ domain knowledge from search behaviors. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1225–1226. ACM, 2011.

Appendix A

Deployed Study

A.1 Study Briefing

Requirements: <ol style="list-style-type: none">1. Check here if the version of your browser meets our requirements: Google Chrome version 47 (or higher) and Mozilla Firefox version 44 (or higher).
Payment: <ol style="list-style-type: none">1. Full Payment If you finished the study through the completion link. This is only possible if you were assigned a partner and together completed the learning phase and final test.2. Partial Payment If you completed the Diagnostic test and waited, but did not receive a partner. You will only receive the partial payment if you clicked on "Stop without completing" (instead of "I've finished" or "Submit study"). This payment will be delivered through a bonus payment which doesn't require you to finish the study.
Study description in Figure A.2
IMPORTANT! We will reject your participation if: <ul style="list-style-type: none">• your answers are shorter than the required word count• your answers are off-topic• during the Diagnostic test and the Final test you change to a different tab more than three times (you will receive a warning ahead of time). Note that during the search phase, tab changes are expected as the search results open in new tabs.• you become inactive (no searching/browsing/scrolling/reading web pages/video watching) for more than 5 minutes during the learning phase.

Figure A.1: Study terms and requirements in the study briefing for all conditions. Both The prolific web page for our study and the welcome page in SearchX contains this briefing. The **green** phrases were only added in the **CSE** condition.

In this study, you are tasked with learning about a given topic **in collaboration with a fellow Prolific worker**. This study is composed of three parts:

1. Diagnostic Test (by yourself).

This is a multiple-choice question test to find out what you already know. Please answer honestly. Your payment is not affected by the number of correct or incorrect answers.

Since this is a collaborative task, after the Diagnostic test you will need to wait for a partner. How much time that takes depends on how many other Prolific workers are active right now. We ask you to wait for 20 minutes. We will notify you when you have waited long enough. Then, please follow the instructions for a partial payment.

2. Collaborative Learning Phase.

We want you, **together with your assigned partner (another Prolific worker)**, to use our custom web search system (we call it "SearchX") to learn about a given topic. You are given 20 minutes to search for documents about that topic. You need to collect and save all the Web pages, publications, and other online sources that are helpful for you to learn about the topic.

Please use only SearchX to learn about the given topic. Do not use any other web search engine or search for an unrelated topic (e.g. your topic is computer science, we consider searches for tomorrow's weather, the latest news, movie reviews, etc. as severely off-topic). If you conduct such off-topic searches, we will cancel your participation.

In order to learn and search together, we provide you with: a chat window so that you can communicate with your partner (when asked for a chat name, choose any name you like), a shared query history so that you can see what your partner is currently searching for and a shared bookmarking list so that you can easily share worthwhile documents.

3. Final Test (by yourself).

We will give you 13 exercises to complete to see how much you have learned through the learning phase; those exercises include questions about the given topic and the writing of an outline for your paper about the given topic. Please answer honestly. Your payment is not affected by the number of correct or incorrect answers. Note that your answers must exceed a minimum word count and be on your assigned topic.

You will need approximately 40 **55** minutes to complete the whole study.

Figure A.2: Study description in the study briefing for all conditions. The **green** phrases were only added in the **CSE** condition. The underlined phrases were not present in the **CSE** condition.

A.2 Registration

Registration

Let's find out what you already know first.
First fill out this basic information about you.

Insert your Prolific ID here

What is your highest academic degree so far?

High School
 Bachelor
 Master
 Doctorate

Which subject areas you have university degree(s)?

Are you an English native speaker?

No
 Yes

What is your level of English?

Beginner
 Elementary
 Intermediate
 Upper-intermediate
 Advanced
 Proficiency

How often do you use Web search engine (e.g., Google, Bing, Yahoo) when you want to learn about something?

More than 10 times a day
 1-10 times a day
 Once a day
 Every few days
 Never

Figure A.3: Registration form.

A.3 Pre-test

Page 1 of 5

Diagnostic Test

Let's find out what you already know first.

Answer these questions about **Sedimentary rocks**:

How much do you know about "feldspars"?

- I don't remember having seen this term/phrase before.
- I have seen this term/phrase before, but I don't think I know what it means.
- I have seen this term/phrase before, and I think I know what it means.
- I know this term/phrase.

How much do you know about "terrigenous sediments"?

- I don't remember having seen this term/phrase before.
- I have seen this term/phrase before, but I don't think I know what it means.
- I have seen this term/phrase before, and I think I know what it means.
- I know this term/phrase.

How much do you know about "subsidence"?

- I don't remember having seen this term/phrase before.
- I have seen this term/phrase before, but I don't think I know what it means.
- I have seen this term/phrase before, and I think I know what it means.
- I know this term/phrase.

Figure A.4: Vocabulary assessment for the first topic out of four pre-test topics. Only the first three questions (out of ten) are shown.

Waiting for another Prolific worker to join...

You will be doing a collaborative search study. We are still waiting for your partner to join.

Please do not refresh/exit this page yet.

If after 20 minutes there is still no update, please stop the study.

Once you drop out after waiting, we will provide you with a partial payment for completing the Diagnostic test.


Your partner has just started their pretest...

Please wait a bit longer :)

Figure A.5: Waiting page after the pre-test for the **CSE** condition. The participants are notified when their partner has started the pretest.

Page 5 of 5

Collaborative search is when participants work together to complete a search task.
 Collaborating with other people can take many forms, a few examples are shown here: two people searching together on a single machine, several people searching towards a common goal on separate machines either in the same location or in different locations.



Have you ever collaborated with other people to search the Web?

No
 Yes

Think about the most recent time you collaborated with others to search the web.
 Describe the nature of the information need that prompted this collaborative search episode. (e.g. husband and wife planning a trip for the family, a group of students working on a writing assignment and sharing search results/findings, a couple shopping for a new sofa, etc.)

Which tools did you use to communicate with your collaborators(e.g. email, chat, Skype, Whatsapp, talking on the phone, etc)?

With how many others did you collaborate with (i.e. not counting yourself)?

How satisfied were you with the quality of the answer(s)?

Not Satisfied	1	2	3	4	5 Completely satisfied
---------------	---	---	---	---	------------------------

How satisfied were you with the ease of working collaboratively?

Not Satisfied	1	2	3	4	5 Completely satisfied
---------------	---	---	---	---	------------------------

Figure A.6: Additional questionnaire related to collaboration for the CSE condition. The questions are utilised to prime participants into a mood for collaboration.

A.4 Search Session

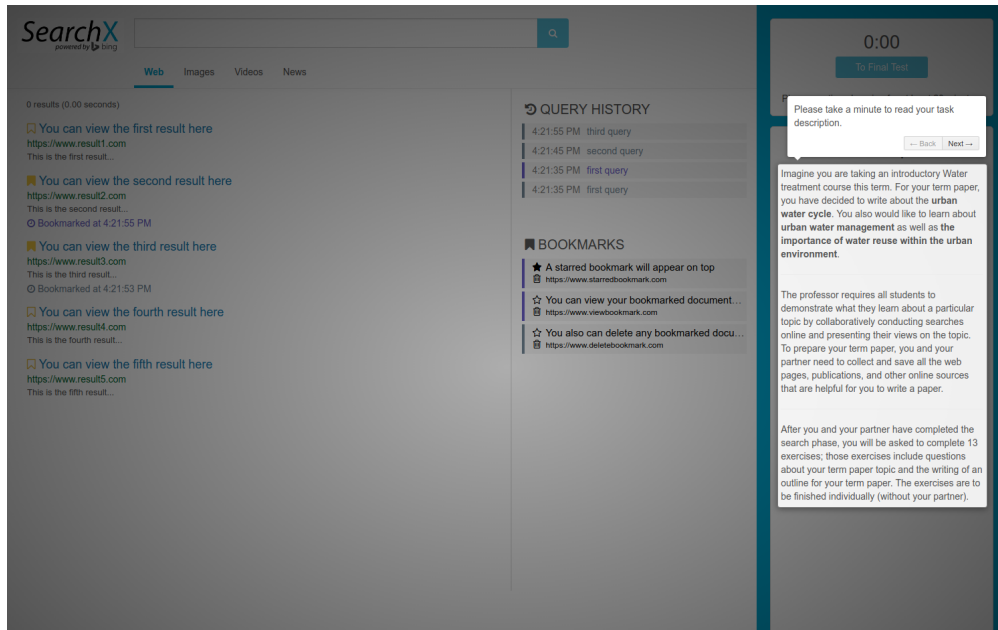


Figure A.7: Interface guide at the beginning of the search session.

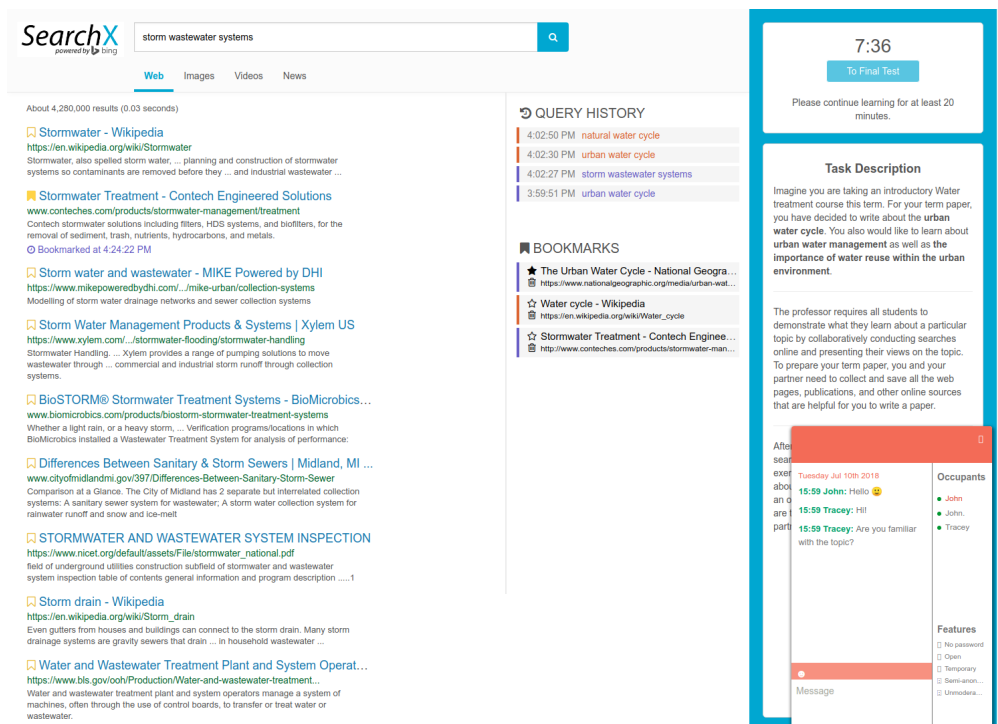
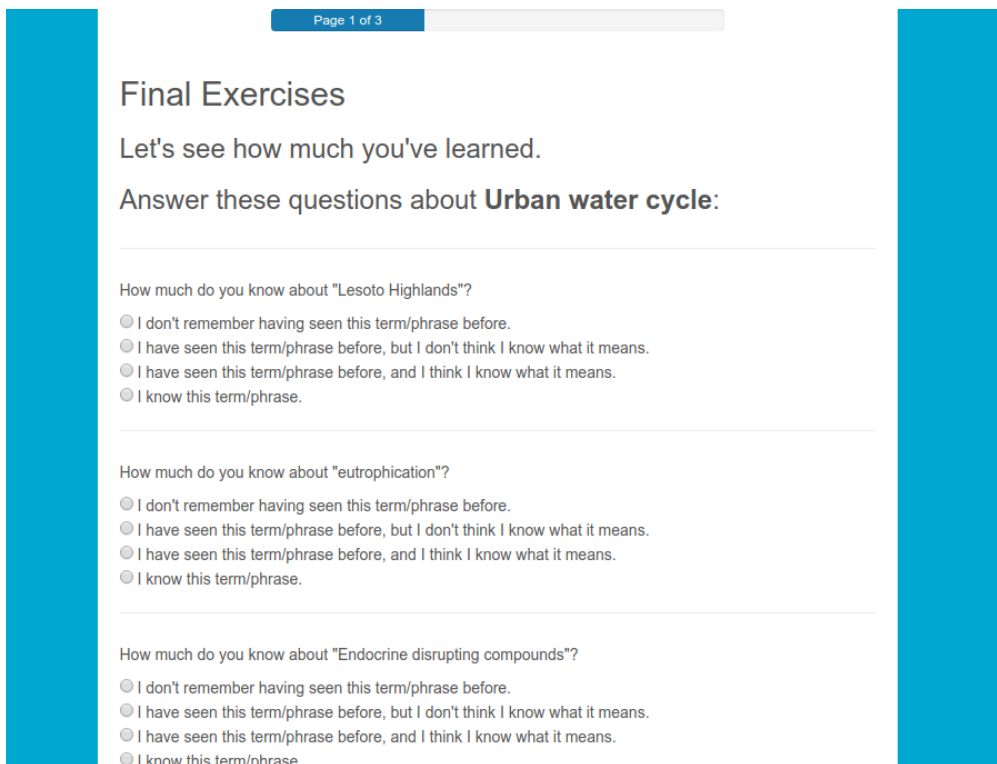


Figure A.8: The deployed search interface for the CSE condition. The SE condition has a similar interface but without colour coding, group chat and the query history.

A.5 Post-test



Page 1 of 3

Final Exercises

Let's see how much you've learned.

Answer these questions about **Urban water cycle**:

How much do you know about "Lesoto Highlands"?

- I don't remember having seen this term/phrase before.
- I have seen this term/phrase before, but I don't think I know what it means.
- I have seen this term/phrase before, and I think I know what it means.
- I know this term/phrase.

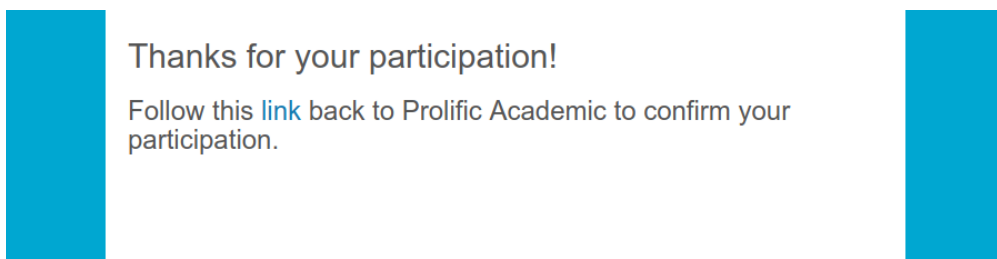
How much do you know about "eutrophication"?

- I don't remember having seen this term/phrase before.
- I have seen this term/phrase before, but I don't think I know what it means.
- I have seen this term/phrase before, and I think I know what it means.
- I know this term/phrase.

How much do you know about "Endocrine disrupting compounds"?

- I don't remember having seen this term/phrase before.
- I have seen this term/phrase before, but I don't think I know what it means.
- I have seen this term/phrase before, and I think I know what it means.
- I know this term/phrase.

Figure A.9: Vocabulary assessment for the post-test. Only the first three questions (out of ten) are shown.



Thanks for your participation!

Follow this [link](#) back to Prolific Academic to confirm your participation.

Figure A.10: Study completion message. Participants can click on the link to confirm their completion and receive their payment.

Page 2 of 3

Based on what you have learned from the learning session, please write an outline for your paper.
Tip: An outline is an organizational plan to help you draft a paper. Here is a simple template example:

1. Introduction
 - 1.1. Main argument: ...
 - 1.2 Purpose of the paper: ...
2. Body
 - 2.1 Argument 1:
 - 2.2 Argument 2:
3. Conclusions

Summary:

Write your outline here:

Please write what you have learned about this topic from the learning session. Use at least 50 words.

During your searches did you have difficulties finding information about something? If so, describe briefly what you were looking for.

Do you have any additional comments regarding the learning phase?

[Previous](#) [Next](#)

Figure A.11: Written assessment and study feedback for the post-test.

Page 3 of 3

We would also like you to describe your experience in collaborating with your partner.

Did you find the collaborative features useful?

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Shared Query History	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shared Bookmarks	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Group Chat	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

How did you make use of the collaborative features and the information from your partner to help in doing your task?

Do you have any additional comments regarding collaborating with your partner?

[Previous](#) [Complete](#)

Figure A.12: Feedback on the collaboration experience for the CSE condition.

Appendix B

Essay Annotation

In order to assign a score for the written assessment, we rely on crowd-workers to annotate each essay according to our criteria.

B.1 Quality of Facts (*D-Qual*)

DATA | {{topic}} {{summary}} {{outline}}

Topic: Anaesthesia

Summary
I learn a more detailed explanation of what anesthesia is and how it is given to the person. I found out that there are different types of anesthesia procedures that can be performed and the different affects it can have on the body. I found out some of the drugs that are given to you when you receive anesthesia.

Essay Outline
1: Brief Introduction of what anesthesia is 2 How it works 3 Different types 4 Effects on the body 5 Conclusion

QUESTION | pulldown menu

Choose the appropriate grade for the results:

0 - Irrelevant ▼

Figure B.1: Question structure for scoring the quality of facts.

<p>Overview</p> <p>Previously, we have asked crowd-workers to study a topic (that they were unfamiliar with) for 20 minutes on the Web, and then express what they learned in the form of a Topic Summary and an Essay Outline. In this job, we will have you judge the quality of their reported results.</p> <hr/> <p>Steps</p> <ol style="list-style-type: none"> 1. Read the Summary 2. Examine the Essay Outline 3. Give a score from 0 to 3 according to the Grading Criteria <hr/> <p>Grading Criteria</p> <p>0 (Zero) - Irrelevant</p> <ul style="list-style-type: none"> • The results are mostly irrelevant to the topic. • Facts are irrelevant to the subject; facts hold no useful information or advice. <p>1 (One) - Lacking</p> <ul style="list-style-type: none"> • The learner knows the topic but did not demonstrate an understanding of the topic. • Facts are generalised to the overall subject matter; facts hold little useful information or advice. <p>2 (Two) - Sufficient</p> <ul style="list-style-type: none"> • The learner demonstrates a basic understanding of the topic. • Facts fulfil the required information need (introducing the subject) and are useful. <p>3 (Three) - Elaborate</p> <ul style="list-style-type: none"> • The learner demonstrates a deep understanding of the topic. • A level of technical detail is given via multiple key terms associated with the technology of the subject. <hr/> <p>Tips</p> <ul style="list-style-type: none"> • The learners were only given 20 minutes to learn a topic they were unfamiliar with. Don't expect them to produce a comprehensive summary. • The learners are allowed to focus on only specific aspects of the topic as long as it is elaborate. • For the outline, look for informative contents explaining the topic (e.g. structure, hierarchy, explanations). Standard chapter sections (e.g. introduction, conclusion) should be ignored.
--

Figure B.2: Instructions for scoring the quality of facts.

B.2 Fact and Statement Counting (F-Fact and F-State)

To make counting easier for crowd-workers, we combined the content of the outline with the summary. We removed outline chapters, and copied over complete sentences to the end of the summary. Additionally, we told workers to find sentences instead of statements for clarity, as in most cases, the number of statements refer directly to the number of sentences [105].

DATA | {{topic}} {{summary}}

Topic: Anesthesia

Summary
 I learn a more detailed explanation of what anesthesia is and how it is given to the person. I found out that there are different types of anesthesia procedures.that can be performed and the different affects it can have on the body. I found out some of the drugs that are given to you when you receive anesthesia.

QUESTION | text box (single line)

Number of Sentences

Some sentences might not be separated by a full stop (.). In those cases, please use your judgment to either count it as one sentence or two separate sentences.

QUESTION | text box (single line)

Number of Facts

Facts refers to individual pieces of information (as opposed to opinions) either explicitly listed or contained within statements. One sentence can contain multiple facts. Opinions or claims are not considered facts.

Figure B.3: Question structure for counting the number of facts and statements.

<p>Overview</p> <p>Previously, we have asked crowd-workers to study a topic (that they were unfamiliar with) for 20 minutes on the Web, and then express what they learned in the form of a Topic Summary. In this job, we will have you help us grade the results by counting the number of sentences and facts within each summary.</p> <hr/> <p>Steps</p> <ol style="list-style-type: none"> 1. Read the Summary 2. Determine the Number of Sentences in the summary. 3. Determine the Number of Facts in the summary. <hr/> <p>Example</p> <p>Topic: Depression</p> <p>I have learned of how events in peoples lives can cause a person to be depressed and how a person’s body reacts on a cellular level, Certain neurotransmitters such as serotonin, dopamine, and norepinephrine may change levels in a person’s brain, resulting in depression. It can however be treated by counseling, solving the event which caused the depression, or medication (antidepressants). I also learned what is meant by post-partum depression.</p> <ul style="list-style-type: none"> • Number of Sentences: 4 Explanation: The first two sentences were accidentally separated using a coma instead of a full stop (.). In this case, you should count the two sentences as separate since they are not cohesive (i.e. they don’t discuss the same topic). • Number of Facts: 3 Explanation: Only the first three sentences contained factual information: a person’s body reacts to depression on a cellular level; certain neurotransmitters may change levels to cause depression; depression can be treated by [treatment]. The last sentence is only a claim of knowing and does not contain much information, therefore it should not be counted as a fact. <hr/> <p>Tips</p> <ul style="list-style-type: none"> • The learners were only given 20 minutes to learn a topic they were unfamiliar with. Don’t expect them to produce a comprehensive summary. • Claims of knowledge are not counted as facts (i.e. I have learned the meaning of [something]) unless they elaborate more on it. • If you notice some minor inaccuracies in the facts (e.g. mixing up the meaning of cellular respiration and glycolysis), you should still count it as a fact.

Figure B.4: Instructions for counting the number of facts and statements.