SciNoon: Exploratory Search System for Scientific Groups

Yaroslav Nedumov
ISP RAS
Moscow, Russia
yaroslav.nedumov@ispras.ru

Anton Babichev
ISP RAS
Moscow, Russia
babichev@ispras.ru

Ivan Mashonsky
ISP RAS
Moscow, Russia
ivan2110@ispras.ru

Natalia Semina
ISP RAS
Moscow, Russia
semina@ispras.ru

ABSTRACT

Exploratory search task poses three challenges to search engines: low specifity of the search goal, long duration of the search and hard to consume search results. Exploratory searches are iterative, multi-tactical and better performed by groups.

We present the demonstration of the first prototype of the exploratory search system for scientific groups. Its main goal is to help with collection of scientific articles related to a scientific group's current project. We tried to meet all exploratory search challenges with focus on support of team work.

SciNoon provides a shared workspace where articles could be collected and annotated. The workspace could be visualized either as an interactive graphical research map or as a table. The research map shows citation relations between articles and could be used for better understanding of the structure of the field. The search progress could be estimated using article coloring by values of their attributes. SciNoon also simplifies keyword search extracting possible keywords from already collected articles and integrating them with existing search engines by the browser plugin.

Using SciNoon the members of a scientific group can search, collect, and process articles and get notifications about each other's progress by the chat bot.

CCS CONCEPTS

• **Information systems** → *Collaborative search*; *Digital libraries and archives*; • **Human-centered computing** → *Computer supported cooperative work*.

KEYWORDS

Exploratory search, collaborative search, academic search engines

ACM Reference Format:
Yaroslav Nedumov, Anton Babichev, Ivan Mashonsky, and Natalia Semina. 2019. SciNoon: Exploratory Search System for Scientific Groups. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019* , 6 pages.

IUI Workshops'19, March 20, 2019, Los Angeles, USA

1 INTRODUCTION

Exploration of a new topic is important task for many people. Students and postgraduates have to learn state-of-the-art while working on their first research project. R&D department researchers need to understand available task definitions and solutions while solving a customer problem. Reviewers often have to review a paper that doesn't perfectly fit their main scope of work and so they need to refresh their understanding of the adjacent field of study.

Exploration of a new topic requires big time investments. In the beginning you often don't fully understand the task and don't know right keywords to use. You have to ask someone for help or use general purpose information sources like Wikipedia [2]. You cannot quickly understand results of a search engine and you have to repeat search while improving your understanding of a domain.

Such search tasks with open-ended, persistent, and multi-faceted problem context and opportunistic, iterative, and multi-tactical search process are called *exploratory search tasks* [12]. Exploratory search tasks are hard and challenging for search engines which are mostly intended for lookup search. Lookup search is focused on high precision but exploratory search needs high recall. Lookup search lasts seconds but exploratory search can take weeks. Results of lookup search are easy to consume but for exploratory search you need time for estimating its relevance. All these is particularly actual for academic search domain.

There are specialized search engines for searching scientific articles such as Google Scholar, Microsoft Academic or Semantic Scholar. They have big databases and good text search engines but their support for exploration is quite limited. While query formulation support is good enough, search results exploration as well as team work and long lasting searches are supported pretty bad.

In our demo system[1] we try to augment existing systems and provide a user with a shared workspace which may be used by a team for collection and exploration of intermediate results.

2 SCINOON

The main use case supported by the current SciNoon prototype is the exploration of a new topic. According to the study [2] and our own examination of existing academic search systems the one of the most missing features is integrated support for collaboration. There are three key features for collaborative search: awareness, division of labor, and persistence [8]. We are providing users with

[1] https://scinoon.com/research/esida-demo

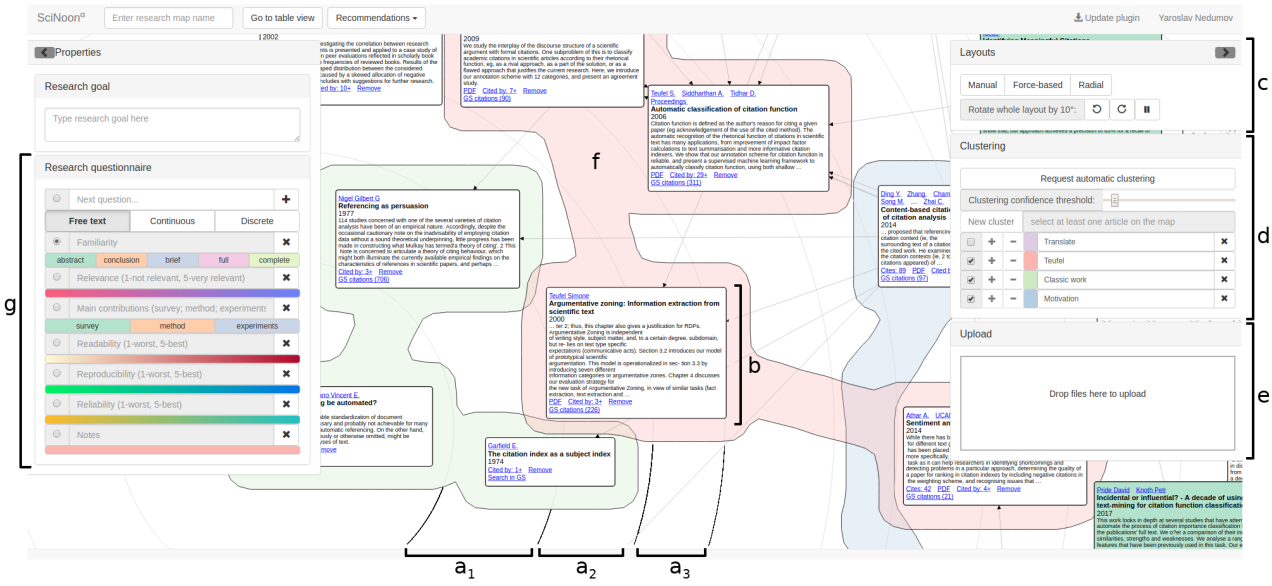

Figure 1: SciNoon user interface elements. a_{1-3} – time-based orbits of radial layout (from the old ones to the new ones), *b* – article's node with navigational controls, *c* – layout managing dialog, *d* – clustering managing dialog, *e* – PDF upload drop zone, *f* – visualization of a cluster, *g* – research questionnaire managing dialog.

the shared workspace (see figure 1) where articles could be collected and processed and a set of corresponding tools. For maintaining awareness we have implemented a chat bot which could be added into research group chat and then will report each team member activity.

Exploration of a new topic is the long-lasting iterative process including several activities: collection of potentially relevant articles, selection and reading of the most interesting ones, summation of read articles according to research-specific aspects.

In the following subsections we will present tools supporting each of these activities implemented in SciNoon.

2.1 Collecting articles

Collecting data about unknown domain is challenging because of quite unspecific search goal. A user doesn't know what to search for and needs help for starting. It is a typical situation for exploratory search task and there are well known partial solutions: query suggestion, dynamic queries, recommendations.

In the case of scientific research one probably already has either a couple of relevant articles obtained from scientific adviser or some keywords to start search from.

If the user already has several relevant articles he or she can upload them to the workspace. SciNoon will parse them, extract metadata and full texts and display them in the workspace. In the background it will extract keywords which may be used later.

If the user doesn't have PDFs he or she probably will use any search engine in order to find articles. SciNoon doesn't maintain its own full text index of articles and instead integrates with Google Scholar. We implemented the plugin for Chromium-based browsers (checked on Chromium and Opera) which is able to grab data automatically from Google Scholar pages for the user. With plugin installed the user will be able to click "Add to research map" button from a search results page and selected articles will be added to the workspace.

Google Scholar provides query suggestions and "Related searches" based on the current query. We augment this functionality by providing research-specific terms. We assume that collected articles, not the current query, explain what a user wants to find. SciNoon extracts terms from the collected articles using ComboBasic algorithm [1]. With SciNoon browser plugin extracted terms are integrated directly into Google Scholar pages and the user can either use them for search alone or append to the current query.

Alternative strategy for collecting articles is snowballing. Using data extracted from uploaded PDFs and cached data from Google Scholar we maintain possibility to "expand" an article node adding either citing articles or cited articles.

The last tool for collecting articles is content-based recommendations. They require some amount of already collected articles. There are four types currently implemented:

(1) **Most cited locally**. Recommends articles which haven't been added yet but are highly cited by the articles from the map. It could be hard to spot such articles manually but it is a trivial task for the system.
(2) **Cutting edge**. Aggregated "cited by" for all novel articles in the workspace. Could be used for finding novel research.
(3) **Old surveys**. Recommends survey articles cited by the articles from the research map. Could be used as a general domain knowledge source.
(4) **New surveys**. Recommends survey articles citing articles on the map.

After the user has collected enough articles he or she will need to further process them. The first task is selecting the article to

start from and here we provide the user with interactive graphical interface called *research map*.

2.2 Selecting articles

The home page of research is called a research map view and displays already collected articles with citation links between them – subgraph of the citation graph.

There are three possible layouts: manual layout, force-based layout and radial layout. Using manual layout the user can place articles as he or she wishes. Force based layout takes into account links between articles and moves connected articles close to each other.

Our novel radial layout is similar to the time layout by Chen [4], but we used polar coordinates in order to better deal with the larger amount of newer articles. So the oldest articles are placed into the center and the newer ones are placed into concentric orbites depending on the publication year and citations. The position of the articles inside the orbit are determined by citation links as in the force layout, so a research field tends to form a sector.

Each article is represented by a rounded rectangle with several text compartments depending on the zoom level. There are four zoom levels and corresponding views:

(1) **10000ft view** could be used for understanding the general structure of the field and processing progress. Each article is represented as rounded corner square without text and with size depending on amount of citations. The whole research graph could be seen at once.
(2) At **1000ft view** (Figure 2) there is the single line text compartment with first author name and year of publication of the article. Couple of dozens of articles could be seen at once.
(3) At **100ft view** there are compartments for full list of authors (up to 10), article title and controls for graph expansion and search in Google Scholar. So only a couple of articles could be seen at once.
(4) At **10ft view** there is also the compartment for the article abstract. We can see mostly the single-article with only parts of neighbour articles.

Each particular node at any zoom level could be manually expanded by double-clicking.

Using the questionnaire described in the next section the user can color article's node depending on the answer to the selected question and then easily find articles with particular answer on the research map. Using coloring by "familiarity" question the whole research progress could be estimated.

2.3 Processing articles

Collection of articles assumes some further summation. In SciNoon we are providing the users with customizable questionnaire which could be answered by each research group member for each article in the research. By default there are seven questions:

(1) **Familiarity**. Assumes the following reading order: abstract, conclusion, brief reading, full text reading, complete understanding of the whole article text.
(2) **Relevance**. Ranges from 1 (not relevant) to 5 (very relevant). Unrelevant articles probably should be deleted from the map, but could be saved in order to save the other group members time when the article looks relevant until full text reading.
(3) **Main contribution**. Highlights most important for the research contribution: survey, original method or experiments. Most probably this question should be adapted to the field of study.
(4) **Readability**. Ranges from 1 (hard to read) to 5 (easy to read). Particularly useful for recommendation articles for students.
(5) **Reproducibility**. Subjective estimation of possibility to reproduce research, ranges from 1 (hard to reproduce) to 5 (easy to reproduce).
(6) **Reliability**. Subjective estimation of reliability of the article results.
(7) **Notes**. Free text notes.

The user is free to use or drop them and can add additional questions if needed.

Users can collect, select and process articles iteratively, populating the common workspace. They can work simultaneously and independently, but can maintain awareness of each other's work using SciNoon's Telegram chat-bot. SciNoon's chat bot (@ScigraphLoggerBot) allows subscription to various events from the research: adding new article, answering a question and so on. In the next section we provide examples how all this features could be used together.

3 USAGE SCENARIOS EXAMPLES

In this section we are going to show how different SciNoon features could be used together in order to deal with two important tasks: directing student's work and writing a review.

3.1 Maintaining students work

Typical situation for exploratory search is giving a research task to a student that is interesting both to the student and his scientific adviser. The field of study is completely new to the student and the scientific adviser is not so familiar with the given problem either. Moreover he is pretty busy to dive into details and do the research together with the student, so he gives him the basic understanding of the task and two-three articles to start from. The adviser prepares new research map in SciNoon, creates initial list of questions for questionnaire and setups group chat with the student and SciNoon chat bot for getting updates.

The student starts with studying the articles that his scientific adviser gave him and adds them to the research map. While reading articles he answers questions from the questionnaire. For example, "familiarity" question shows his progress in studying the articles, "relevance" question represents articles relevance to the given task, "notes" – is where the student puts his thoughts and summary about research. The adviser is notified by chat bot about each such update and is able to correct his student when needed.

As the student continues his research he needs some more articles. SciNoon helps him with search queries in Google Scholar by showing keywords extracted from his research map and personal recommendations. It also provides him with forward citation chaining either from SciNoon internal database or from Google Scholar for recent advances in the field along with backwards citation chaining for better understanding the field roots. For articles

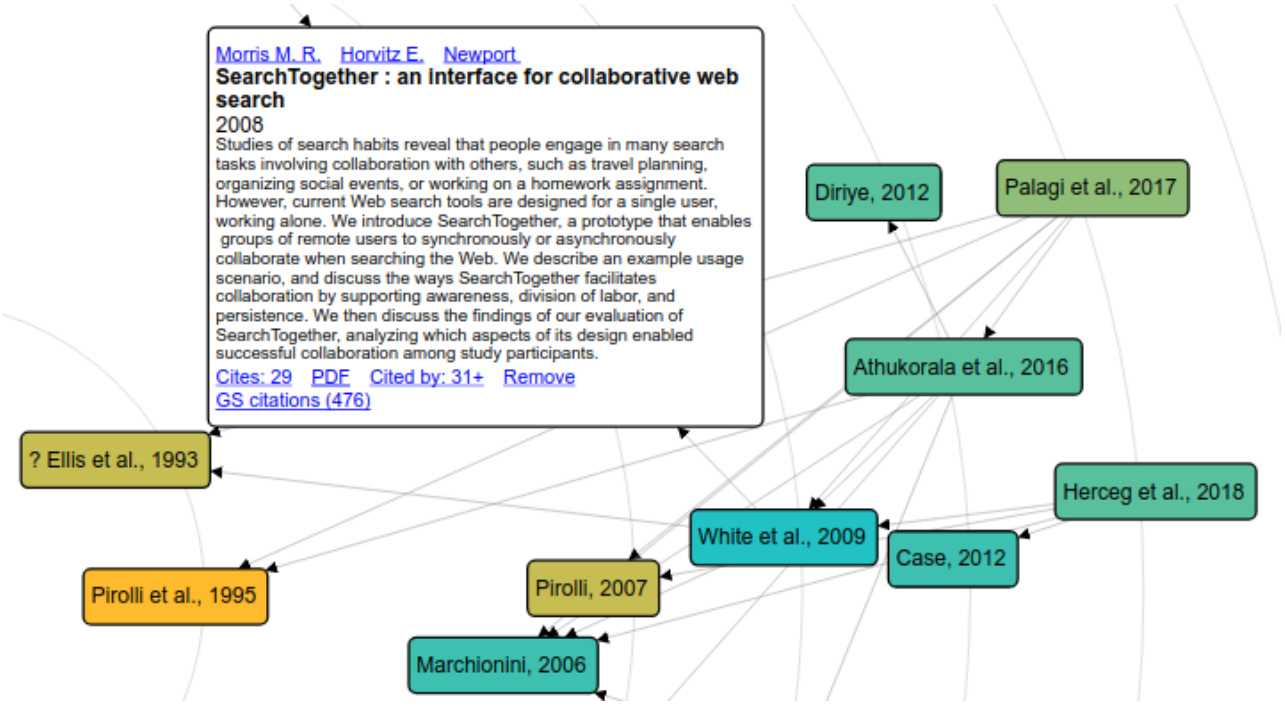

Figure 2: A fragment of research map at 1000ft view with one expanded article node and several unexpanded connected by citation links. Articles are colored according to their relevance from yellow to blue. Radial layout is used.

with full text available backwards citation chaining is also available. The other features for article collection also could be used.

Since the adviser is notified about his student progress he can easily correct student's article selection mistakes. So, after working together in such manner some time the team will have a list of relevant articles and answers for questions important for their research. As the last step all work could be exported to the csv file.

3.2 Writing review

The second frequent case where fast exploration of a new topic could be needed is writing a critical review. In this case preliminary familiarity with the topic is much higher, but anyway you need to recall the exact topic of the article and freshen your knowledge regarding recent advances in the field. SciNoon probably can help.

A reviewer can add article's PDF to the fresh research map and then do one backwards citation chaining step using "Cites" button on the added article. Following this by "Cutting edge" recommendation all competing articles could be easily found.

In the next section we will briefly explain SciNoon's internals.

4 SYSTEM DESIGN

SciNoon is the client-server web-based application. The server part has modular design and is based on Play framework[2] and Akka[3]. The client part comprised in rich web application and browser plugin. Client and server are communicating via HTTP API.

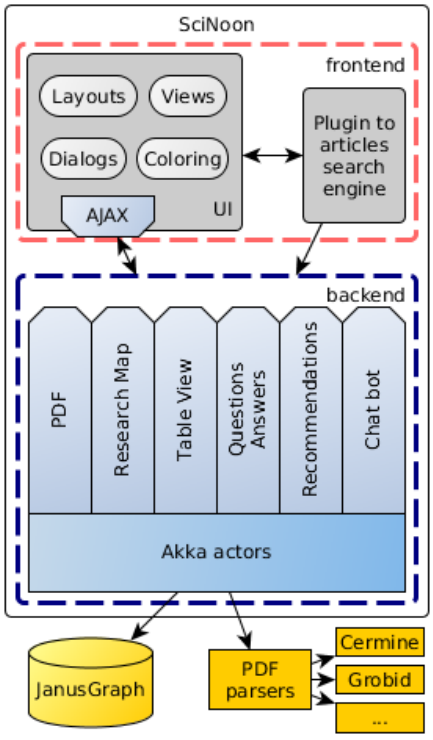

Figure 3: Architecture of SciNoon

[2]https://www.playframework.com/
[3]https://akka.io/

4.1 SciNoon server

Play framework assumes MVC architecture. Since we use rich client the view component is reduced and consists of trivial HTML template and JavaScript code of the client building the main part of HTML in browser.

We use graph-based datamodel so all data is represented either as nodes or as edges. Main types of nodes are *articles*, *scientists*, *answers* and *researches*. Nodes have properties depending on their type. For example *article* node has title, *scientist* node has first name and so on. There are several types of edges: *cites* directed edge starts from citing article node and ends in cited article node, *author* edge connects article with its author and so on. We use JanusGraph[4] graph database with Cassandra backend for persistence.

HTTP API is divided into several controllers (look at Figure 3) doing data conversion and passing data for processing into Akka actor system for asynchronous processing.

We use external systems for extracting metadata and bibliography from PDFs: CERMINE [11] and GROBID [6].

4.2 SciNoon client

SciNoon's client part is written in Typescript language which is translated into JavaScript. We use d3js and Bootstrap frameworks. There are several interconnected modules for getting data from the server, drawing research map, processing user's input and so on.

SciNoon's browser plugin is written in JavaScript using Web Extensions API. It integrates with SciNoon and Google Scholar sites using content scripts.

5 RELATED WORK

There are several well-known search engines for scientists such as Google Scholar, Microsoft Academic, Semantic Scholar, aMiner, CiteSeerX, PubMed and the others. There are specialized social networks like ResearchGate or Academia.edu also providing some search possibilities. At last many digital libraries provide search tools on their sites. But their support for exploratory search task is quite limited.

However there are several systems and approaches less known but better suited for exploratory search (in general or for scientific articles). In this section we will describe some of them.

SearchTogether [8] was aimed at the task very similar to ours: small-group collaborative searching. This system was general purpose and was build on top of classic search engines integrating them together with instant messaging and recommendations, and providing a common workspace. The authors showed that searching with SearchTogether is more efficient than searching without it. Unfortunately the project didn't evolve and after some time with the rise of social networks was claimed by the authors to be outdated [7]. Despite this fact we believe that its focus on awareness, division of labor and persistence will be much more useful for professional users such as scientists and R&D specialists.

Interesting approach for organization of a scientist's workspace was proposed by Beel et al [3]. Their tool, Docear, was developed as "Microsoft Office for scientists". It contains several modules: digital library module providing access to research articles, research module providing keyword search, PDF viewer for reading and annotating articles, mind mapping module for managing all information, word processing module for writing articles and reference manager for managing bibliography. Docear is an amazing demo, but it is not built for search and this is single-user application missing any collaboration features which makes it difficult to use it in research groups.

In the works [9, 10] the authors describe the IntentRadar search system specifically designed for exploratory search for scientific articles. The main idea is to model user's search intents by keywords and interactively evolve them getting relevance information from the specially designed user interface. The authors proves efficiency of the proposed technique in the series of experiments. The developed interactive user interface covers both query formulation and search results exploration tasks and so could be very helpful for the exploration of a new topic. However, as well as Docear, IntentRadar is a single-user application.

There is some research regarding exploration of particular fields of study: science mapping. One of the most known science mapping tool is CiteSpace [5]. This is standalone Java application implementing methods for co-citation analysis enabling the users with the possibility to reveal the structure of the field and emerging trends. CiteSpace doesn't maintain its own database of articles and should be provided with data exported from Web of Science. This complicates usage of CiteSpace for exploratory search but we are going to implement some methods of co-citation analysis in the future versions of SciNoon in order to make them available for interactive use by a team of scientists.

6 CONCLUSION

Exploratory search is a complex task challenging search systems in many ways.

We developed SciNoon – exploratory search system for scientific groups providing unique combination of tools helping exploring new domains.

To collect articles, we provide the user with three tools: navigation on citation graph enabling possibility to do snowballing, integration with the Google Scholar search engine (with custom keywords generation) and recommendations, based on already collected articles.

To help the user to select the most interesting articles we implemented interactive graphical interface visualizing the citation graph of the collected articles. It supports several graph layouts and different level of details. We also proposed the new *radial* layout for easier overview of a research.

The user can create questionnaire specific for his or her research and then fill it in for each collected article. Article nodes will be colored depending on the given answers.

All research team members could work simultaneously and independently and they will be notified of each other's activity by SciNoon's chat bot.

All this makes our system a helpful companion to existing full text search academic search engines enabling possibility to do lasting research by the several team members together.

[4] http://janusgraph.org/

7 ACKNOWLEDGMENTS

Denis Turdakov, Nikita Astrakhantsev and Anna Loik for reading early versions of the paper. The reported study was partially funded by RFBR according to the research project 17-07-00978 A.

REFERENCES

[1] Nikita Astrakhantsev. 2015. *Methods and software for terminology extraction from domain-specific text collection.* Ph.D. Dissertation. Ph. D. thesis, Institute for System Programming of Russian Academy of Sciences.

[2] Kumaripaba Athukorala, Eve Hoggan, Anu Lehtiö, Tuukka Ruotsalo, and Giulio Jacucci. 2013. Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. *Proceedings of the Association for Information Science and Technology* 50, 1 (2013), 1–11.

[3] Joeran Beel, Bela Gipp, Stefan Langer, and Marcel Genzmehr. 2011. Docear: An Academic Literature Suite for Searching, Organizing and Creating Academic Literature. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*. ACM, New York, NY, USA, 465–466. https://doi.org/10.1145/1998076.1998188

[4] Chaomei Chen. 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the Association for Information Science and Technology* 57, 3 (2006), 359–377.

[5] Chaomei Chen, Fidelia Ibekwe-SanJuan, and Jianhua Hou. 2010. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology* 61, 7 (2010), 1386–1409. https://doi.org/10.1002/asi.21309 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21309

[6] Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Research and Advanced Technology for Digital Libraries*, Maristella Agosti, José Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 473–474.

[7] Meredith Ringel Morris. 2013. Collaborative Search Revisited. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1181–1192. https://doi.org/10.1145/2441776.2441910

[8] Meredith Ringel Morris and Eric Horvitz. 2007. SearchTogether: An Interface for Collaborative Web Search. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (UIST '07)*. ACM, New York, NY, USA, 3–12. https://doi.org/10.1145/1294211.1294215

[9] Tuukka Ruotsalo, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. 2015. Interactive intent modeling: Information discovery beyond search. *Commun. ACM* 58, 1 (2015), 86–92.

[10] Tuukka Ruotsalo, Jaakko Peltonen, Manuel JA Eugster, Dorota Głowacka, Patrik Floréen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. 2018. Interactive Intent Modeling for Exploratory Search. *ACM Transactions on Information Systems (TOIS)* 36, 4 (2018), 44.

[11] Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. 2015. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)* 18, 4 (01 Dec 2015), 317–335. https://doi.org/10.1007/s10032-015-0249-8

[12] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.