Improving expert search effectiveness: Comparing ways to rank and present search results

Thomas Schoegje
tmschoegje@gmail.com
Utrecht university
Utrecht, Netherlands

Lynda Hardman
lynda.hardman@cwi.nl
CWI
Amsterdam, Netherlands

Arjen de Vries
arjen@acm.org
Radboud University Nijmegen
Nijmegen, Netherlands

Toine Pieters
t.pieters@uu.nl
Utrecht University
Utrecht, Netherlands

ABSTRACT

Expert search systems help professionals find colleagues with specific expertise. Expert search results can be presented as a list of documents with their associated experts, or as a list of candidate experts with evidence for their expertise based on documents they authored. The type of result may affect search behaviour, and therefore search task performance. Previous work has not considered such effects from the result presentation, focusing instead on how to rank experts or on ways to interact with the search results.

We compare the task performance of novice users using either a document-centric interface (where each search result is a document and its associated expert) or a candidate-centric interface (where each search result is a candidate expert and their associated documents). We also compare candidate-centric and document-centric ranking functions per interface.

A post-experiment survey indicated that two variables affect which interface participants preferred: the retrieval unit (candidates or documents) and the complexity (number of documents per search result). These variables affected participants' search strategy, and consequently their task performance. A quantitative analysis revealed that 1) using the candidate-centric interface results in a higher rate of correctly completed tasks, as users evaluate candidates more thoroughly, and 2) the document-centric ranking yields faster task completion. Weak evidence of a statistical interaction effect was found that prevents a straightforward combination of the most effective interface type and the most efficient ranking type. Present work resulted in a more effective, albeit less efficient, search engine for expert search at the municipality of Utrecht.

CCS CONCEPTS

• **Information systems** → *Search interfaces*; *Presentation of retrieval results*; **Expert search**.

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0434-5/24/03
https://doi.org/10.1145/3627508.3638296

KEYWORDS

expert search, search interface, retrieval unit, ranking, people search

ACM Reference Format:
Thomas Schoegje, Lynda Hardman, Arjen de Vries, and Toine Pieters. 2024. Improving expert search effectiveness: Comparing ways to rank and present search results. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3627508.3638296

1 INTRODUCTION

Up to 73% of professionals in the public sector often encounter (complex) work tasks for which they seek advice from colleagues [33]. Oftentimes it is unclear for professionals where they can find a colleague with expertise on a given topic, resulting in the need to find the right expert for the right task (e.g. 'who can tell me how the sound leak in concert hall Tivoli was repaired?'). Recent work found that 59.5% of queries are conducted to find a person, based on the enterprise search logs of a large biotech company [25]. Similarly, a study on policy worker search tasks found that half of the tasks were about finding the correct person, rather than finding information directly [37]. This search strategy was employed by policy workers to solve complex search tasks, as it allowed users to acquire the information they need for less effort. Additionally, an expert could help solve one's task and contextualise the available information [37].

Previous works on expert search interfaces have considered what information is required to evaluate whether an expert is relevant (e.g. [14, 15]) and explored different ways for interacting with list of search results (e.g. [9, 11, 12, 22, 42]). However, to the authors' knowledge, no evaluation has directly considered whether expert search results should be presented as documents or as experts. We observed that, during informal think-aloud studies, participants re-framed their original search intents from a people-focused goal to an evidence-centric sub-goal: what type of documents might the person in question write? Users translated their information needs to the functionality shown in the search interface. We hypothesise that the presentation of search results affects the search strategy, and therefore task completion. In this paper, we quantify the influence of presenting search results as either documents or candidates on task performance. This can inform what is otherwise an easily overlooked and unconsidered design decision in practice.

Our scope is on expert search within the context of the municipality of Utrecht, as this lets us evaluate expert search within the enterprise search context of 'find a colleague with expertise'. In this setting, we are interested in reproducing previous findings on how to rank experts (by document or by candidate) and then to consider how to present the search results. The two ranking types and the two interface types investigated are shown in Figure 1. Finally, we are interested in whether there are statistical interaction effects between the type of ranking function and the type of interface, and which has a larger effect size on task performance. These interests result in our research questions:

RQ1 Is a document-centric result ranking or a candidate-centric result ranking preferable for findings experts who work at the municipality of Utrecht?
RQ2 Is a document-centric interface or a candidate-centric interface preferable for finding experts who work at the municipality of Utrecht?
RQ3 Are there interaction effects between the ranking type and the interface type?
RQ4 What are the relative effect sizes of the ranking type and the interface type on task completion?

We relate this study to previous work in section 2. The dataset is characterised in Section 3, and discuss the implementation of the system in Section 4. In section 5 our method is described which encompasses both a qualitative study and a quantitative study. Section 6 details the qualitative analysis, where two factors were found that affect how users engage with the interface: the complexity of the information and the presented retrieval unit. These variables appear to affect task performance. The quantitative results in section 7 indicate that the document-centric ranking type is faster, whereas the candidate-centric interface type leads to more tasks completed correctly. Weak evidence for an interaction effect was observed between the interface type and the ranking type, prohibiting us from combining the best interface tested with the best ranking tested. In the discussion in section 8 we argue that correct task completion is preferable over efficient task completion in this context, as approaching the incorrect expert can incur a social cost and lose time, which is not measured in this study. Based on the experiment with novice users trying to find colleagues at an organisation, the paper concludes that presenting expert search results as overviews of candidates elicits a more thorough assessment of search results, resulting in more effective task completion for novice users.

2 RELATED WORK

Literature on how to rank experts consists of two main approaches [2]: ranking individual documents (document-centric) or creating some model of the candidate's expertise, and ranking these candidates directly (candidate-centric). The search behaviours of professionals can also be characterised as being either document-centric or candidate-centric. For instance, professionals perform a document-centric search strategy when they search for a relevant document and then contact the author [1]. An example where people perform a candidate-centric strategy is when they ask colleagues whether they know experts who can help them solve a task [36].

Given our domain of interest, we assume the authors of documents are experts on the topic and therefore avoid challenges in attributing expertise to the right people [2]. We note that some documents are more informative of their authors' expertise than others [26], which we do not account for in this paper as it does not pertain to our research questions.

Another line of research has investigated why, and how, people search for experts [14, 42]. Such studies informed what contextual information should be included within each search result [15, 16], assisting users in their decision of whom to approach. This decision is based on both the perceived quality of the expert as well as their approachability [32]. Some of these studies note the value of presenting the search results as people as opposed to documents (e.g. [30]) and designing retrieval units suitable for the current work task (e.g. [38]).

The importance of designing interfaces has been noted in survey papers on expert search as recent as 2019 [10, 16]. Some different interface designs and functionalities have been proposed. Proposed interfaces often let users interact with the results shown (e.g. [9, 11, 22]), and sometimes deviate entirely from a traditional search engine result page (e.g. [27, 28]). There are studies that investigated exclusively ways to present a document-centric search result (e.g. [44]), candidate-centric result (e.g. [23, 32]) or entity-centric result (e.g. [13]). These are typically not directly compared, and in most works this design decision is made without explicit rationale because the research questions are focused elsewhere. However, result presentation affects how users interact with the system, as there is a relationship between the type of knowledge sought and the ideal modality of search results [31]. Studies have also found that presenting result grids or result lists affects how users examine the results [35, 39]. In addition, it was shown that the user's task affects how users engage with the interface [35], and the present authors are not aware of existing research that focused on whether results should be presented as documents or candidates for expert search tasks. Hence we re-examined this fundamental design decision of expert search interfaces, and measured the impact of this decision on the users' effectiveness, efficiency and user satisfaction while searching.

Note that precision and recall are not suitable for evaluating an interface, and hence our evaluation relies on the observation that the best search system is the one that is most useful for the work tasks of the user [4, 18, 40, 41]. This study considers which interface is most useful for expert search tasks. Usefulness is measured as *useful = usability + utility* [29]. All systems in our study have the same utility (i.e., they can solve the same tasks), and therefore the evaluation focuses only on which system is most usable. Usability consists of three components: the system's effectiveness, efficiency and user satisfaction [19]. This is not a novel approach to evaluating expert search interfaces (see e.g. [22]).

Effectiveness can be measured as the proportion of tasks that were completed correctly. System efficiency is typically measured in task completion time. One approach to measuring user satisfaction is the System Usability Scale (SUS) questionnaire [6]. Although it does not directly measure satisfaction, it is a widely adopted usability metric for test-level satisfaction (i.e., measuring usability for the whole test session as opposed to measuring it every task) [21].

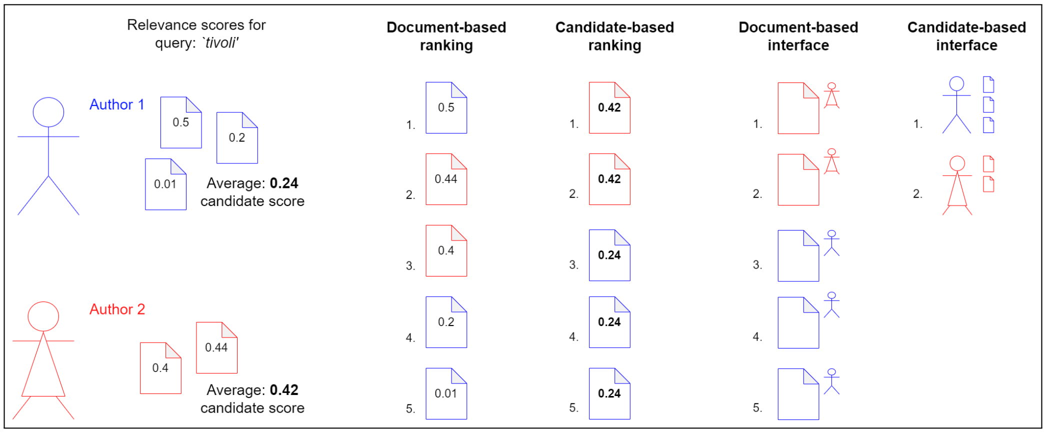


Figure 1: During expert search results both the ranking and interface can focus on the documents or on the candidates. Document-centric ranking sorts by the relevancy of individual documents, whereas the candidate-centric ranking sorts by overall relevancy of the candidate (e.g. the average of their documents).

3 COUNCIL DOCUMENT COLLECTION

In collaboration with one of the country's largest municipalities, we utilized a city council document collection [1]. We opted for this dataset as it is realistic to needs of expert search in an organisation, and it allows us to avoid the problem of linking a candidate expert to the evidence of their expertise. The collection comprises approximately 6000 letters and memos, which were written by around 1600 public servants who directed the documents to council members. The letters are typically two pages long and written to provide information to the city council in preparation for council meetings. Additionally, these letters may include attachments that offer more extensive and detailed information. Memos, on the other hand, are brief updates and are less informative in nature. Each document is associated with a specific sub-domain, such as public health, which users can specify in the document's metadata when uploading it. The collection of sub-domains in which a user possesses expertise is referred to as their portfolio.

The documents were written using standardized templates (created in Microsoft Word), which enables the extraction of author names from the document header with a regular expression (see Appendix A for details). Documents without extractable author names are not included in the indexing process. Documents with multiple authors are also excluded, because these documents could introduce a bias in our experimental setup (see section 5 for more detail). After grouping author aliases 1032 unique authors were found who wrote 4483 documents.

4 IMPLEMENTING EXPERT SEARCH

Although recent approaches to ranking (e.g. [23, 43] and presenting [5] expert search results are sophisticated, our implementation is minimalistic to maintain focus on our research questions. This section describes the design decisions that let us investigate the research questions. Further implementation details on all four combinations of ranking and interface types are in Appendix B, and the code is available at www.github.com/UtrechtUniversity/expertsearch.

Ranking types are implemented by two elasticsearch [2] indexes. In the document-centric index the entries contain the full text of single documents. In the candidate-centric index each entry contains all the text of all the documents that one individual wrote.

Interface designs are implemented by presenting each result with a document panel and an expert panel. The document panel showcases the evidence of expertise, whereas the expert panel displays the candidate's portfolio and contact information. The document-centric interface (shown in Figure 2) emphasises the evidence of expertise, and therefore positions the document panel on the left side. If the result appears relevant, users can then locate the contact information in the expert panel. The candidate-centric interface (depicted in Figure 3) presents an overview of the candidate. Therefore the expert panel is on the left, and multiple pieces of evidence are presented on the right.

Documents are always presented using the document's title and a snippet derived from Elastic's highlight feature, limited to a maximum of 100 characters. Titles are clickable and open the corresponding documents in new tabs, to ensure users do not close the search engine tab. The author panel includes a name, contact details, and portfolio.

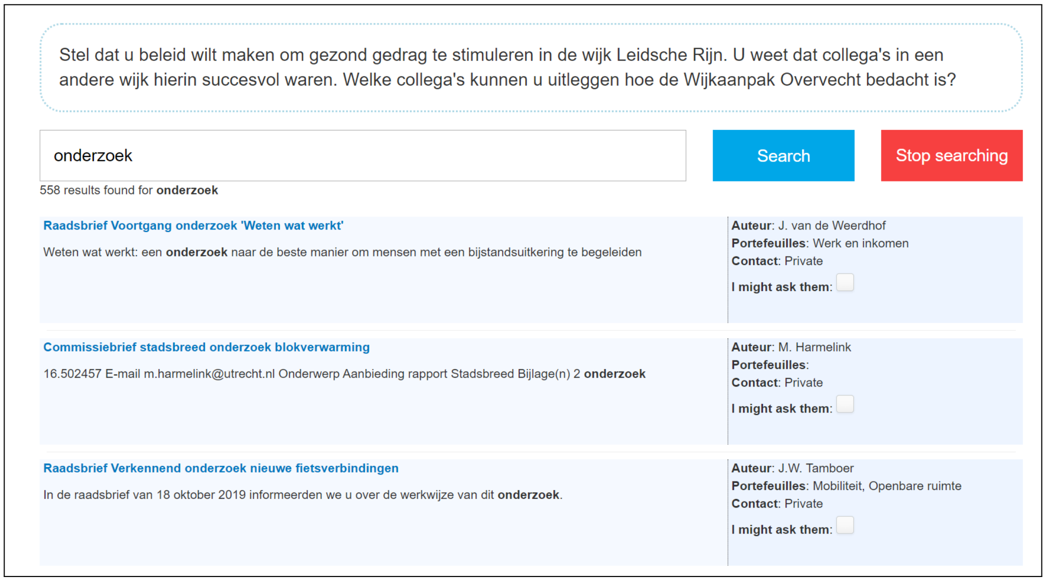
5 METHOD

We investigate the effects of different ranking types and interface types on task completion by having users perform simulated tasks in variations on the same search engine. This experiment was performed in person as 1) search behaviour was logged in the browser's local storage, 2) to ensure participants use the same equipment and environment, and 3) to ensure experiments were performed without distractions.

A power analysis was performed to determine the required sample size for our experiment, based on preliminary findings with the first three participants. A two-factor Analysis of Variance (ANOVA)

[1] zoek.openraadsinformatie.nl - accessed 11-9-2021

[2] elastic.co



Stel dat u beleid wilt maken om gezond gedrag te stimuleren in de wijk Leidsche Rijn. U weet dat collega's in een andere wijk hierin succesvol waren. Welke collega's kunnen u uitleggen hoe de Wijkaanpak Overvecht bedacht is?

onderzoek

558 results found for **onderzoek**

Raadsbrief Voortgang onderzoek 'Weten wat werkt'

Weten wat werkt: een **onderzoek** naar de beste manier om mensen met een bijstandsuitkering te begeleiden

Auteur: J. van de Weerdhof
Portefeuilles: Werk en inkomen
Contact: Private
I might ask them: □

Commissiebrief stadsbreed onderzoek blokverwarming

16.502457 E-mail m.harmelink@utrecht.nl Onderwerp Aanbieding rapport Stadsbreed Bijlage(n) 2 **onderzoek**

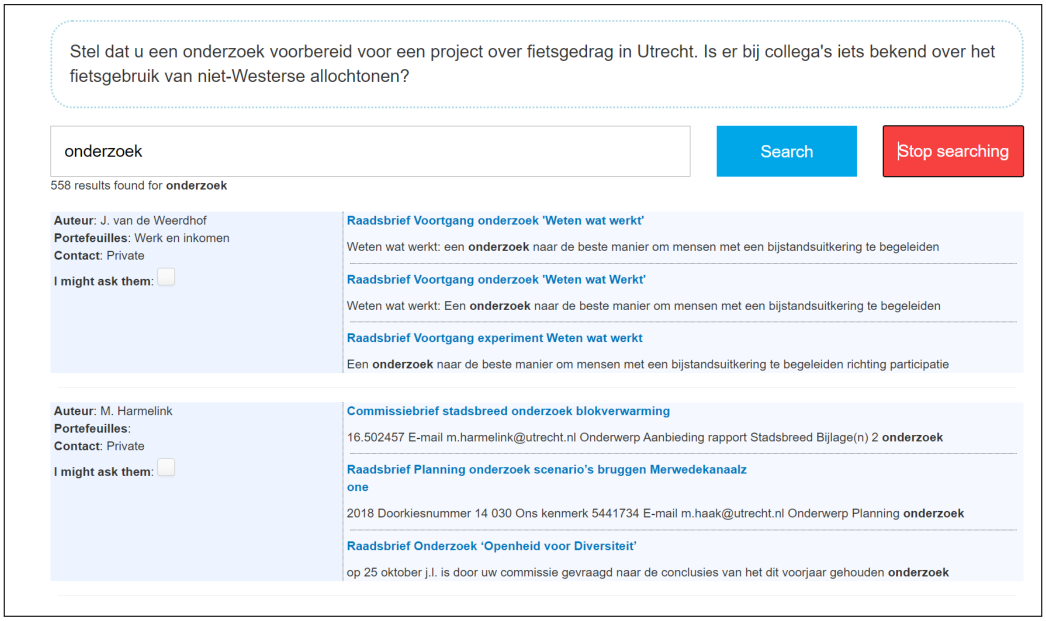
Auteur: M. Harmelink
Portefeuilles:
Contact: Private
I might ask them: □

Raadsbrief Verkennend onderzoek nieuwe fietsverbindingen

In de raadsbrief van 18 oktober 2019 informeerden we u over de werkwijze van dit **onderzoek**.

Auteur: J.W. Tamboer
Portefeuilles: Mobiliteit, Openbare ruimte
Contact: Private
I might ask them: □

Figure 2: The expert search interface with document-centric retrieval units.



Stel dat u een onderzoek voorbereid voor een project over fietsgedrag in Utrecht. Is er bij collega's iets bekend over het fietsgebruik van niet-Westerse allochtonen?

onderzoek

558 results found for **onderzoek**

Auteur: J. van de Weerdhof
Portefeuilles: Werk en inkomen
Contact: Private
I might ask them: □

Raadsbrief Voortgang onderzoek 'Weten wat werkt'

Weten wat werkt: een **onderzoek** naar de beste manier om mensen met een bijstandsuitkering te begeleiden

Raadsbrief Voortgang onderzoek 'Weten wat Werkt'

Weten wat werkt: Een **onderzoek** naar de beste manier om mensen met een bijstandsuitkering te begeleiden

Raadsbrief Voortgang experiment Weten wat werkt

Een **onderzoek** naar de beste manier om mensen met een bijstandsuitkering te begeleiden richting participatie

Auteur: M. Harmelink
Portefeuilles:
Contact: Private
I might ask them: □

Commissiebrief stadsbreed onderzoek blokverwarming

16.502457 E-mail m.harmelink@utrecht.nl Onderwerp Aanbieding rapport Stadsbreed Bijlage(n) 2 **onderzoek**

Raadsbrief Planning onderzoek scenario's bruggen Merwedekanaalz one

2018 Doorkiesnummer 14 030 Ons kenmerk 5441734 E-mail m.haak@utrecht.nl Onderwerp Planning **onderzoek**

Raadsbrief Onderzoek 'Openheid voor Diversiteit'

op 25 oktober j.l. is door uw commissie gevraagd naar de conclusies van het dit voorjaar gehouden **onderzoek**

Figure 3: The expert search interface with candidate-centric retrieval units.

of the task completion time, while using an estimated standard deviation of 0.53 minutes, a detectable contrast of 0.5 minutes, and a desired power level of 0.946. Using these assumptions approximately 40 observations per factor combination are necessary, equivalent to 20 participants performing 8 tasks each.

Participants in the study were selected to be novices in the domain, as we observed that they faced the greatest challenges in locating both information and experts. Experienced users already know the most relevant information sources and individuals with expertise. Given that employees such as council members are elected citizens, and that no specialist training is necessary, we assume that citizens exhibit similar information behaviour as new employees.

Due to regulatory restrictions and the unavailability of public servants during the early stages of the COVID-19 pandemic, we conducted the experiment with citizens as participants. In compliance with local regulations at the time of the experiment (restriction contact outside of known social circles), we only recruited acquaintances of the first author. These were unfamiliar with the research goals beyond what was necessary for an informed consent. To ensure the safety of participants, numerous precautions were taken, including maintaining social distancing, conducting repeated self-tests, and regularly disinfecting the hardware and equipment using alcohol wipes.

Twenty participants took part in the experiment, all of whom were native speakers. Half the participants identified as women. Most participants were aged 25 to 35, with four outliers being older than 40. No participant had professional work experience in a similar domain, and none reported having any domain-specific knowledge.

Tasks were adapted from tasks policy workers reported performing at the municipality. Each of the eight simulated task starts of a work task description (i.e., the end goal of the user), which is a textual description of one or two sentences. This is followed by the search task description (i.e., information need) described in a sentence.

Ground truth data was constructed based on the assumption that experts on a relevant sub-domain would know the answer, or would know the person to contact instead. An experienced policy worker from the municipality was available to determine which sub-domains were relevant for each task. They were not able to assess the relevance of individual experts, as they do not know the expertise of all individuals employed at the organisation.

The experimental design took into account that participants should be able to distinguish between the systems in the post-experiment questionnaire, and hence each participant tested two systems with different interfaces. The presentation order of interface types and ranking types are counter-balanced, and the task order is randomised.

If a highly relevant document was marked as relevant in the document-centric interface, all of its authors are marked as relevant. In the candidate-centric interface, the user might mark one candidate as relevant based on this highly relevant document, but not the other. To avoid this asymmetry in relevance assessments, we exclude documents authored by multiple people from the dataset.

Procedure for the experiment was to present participants with one of the interfaces and a brief introduction. After given informed consent and familiarising with the system they were instructed to imagine themselves as new employees at the municipality, tasked with assignments that required input from their colleagues. They were asked to identify and mark the candidate expert(s) whom they would consider approaching for assistance, if any. Then they performed the tasks without time limit. Each task was started and ended by pressing a button. During a task, a description is displayed and participants can check the boxes of experts they would approach. Users completed four tasks in this first system, and proceeded to complete a questionnaire to evaluate the system. Next, participants familiarised themselves with to the second system and performed an additional four tasks. After a questionnaire about this system they were presented a questionnaire that compared the two systems, and asked open-ended questions about how they search for expertise. The list of questions is published alongside the code at github.com/UtrechtUniversity/expertsearch.

Analysis of the qualitative investigated user preferences by manually clustering and interpreting users' responses from the questionnaires. This was followed by a quantitative analysis that measures the effect of the ranking type and interface type on aspects task performance (as introduced in section 2). Systems were compared in terms of effectiveness (rate of successful task completion), efficiency (time to task completion) and user satisfaction (measured using the System Usability Scale (SUS) questionnaire [6]).

6 QUALITATIVE ANALYSIS

Users' preferences for the user interface were divided, with half the users preferring one system in the questionnaire and the other half preferring the other system. We investigate the reasons for this, and whether to account for this in our quantitative analysis. One participant's data was excluded from both the qualitative and quantitative analyses as they misunderstood the instructions, and performed several tasks without issuing queries (and therefore rated the same set of results for each task).

6.1 User preferences

A total of 30 reasons were given by participants to support why they preferred one system over the other. After grouping similar reasons, as shown in Table 1, we found nearly all reasons pertained to the interface type. Exceptions are marked with *, but even then these were only given when participants compared two systems where only the interface (and tasks) changed.

User preferences indicated two main factors: retrieval unit complexity and perceived retrieval unit. Both interfaces represent opposites in terms of these factors, and users disagree on what is preferable. The retrieval unit complexity refers to the level of complexity involved in retrieving information, while the perceived retrieval unit relates to users' perception of the granularity and relevance of the retrieved information. Understanding these factors is crucial for designing interfaces that cater to diverse user preferences and enhance usability in expert search systems. Two main factors emerge that explain preference to one system or the other: the retrieval unit and the retrieval unit complexity.

The most reported factor to prefer the document-centric interface is that it shows less information (D1), which makes it easier to evaluate a search result (D2). This contrasts against the primary reason to prefer the candidate-centric interface: this interface gives a better overview of what a candidate expert does (C1). Participants disagree on the trade off between the amount of information needed to be confident enough of a candidate's expertise. The second factor that participants prefer the document-centric interface is that it allows them to first evaluate the document, and then use the author characteristics (i.e., their portfolio) as further evidence (D3). This contrasts with the second main reason to prefer the candidate-centric interface: these users prefer first evaluating author characteristics and using the written documents as further evidence (C2 and C3). This second factor shows how the interfaces

Table 1: Reasons for preferring one search system over the other (as reported in the questionnaire). A few reasons (marked with *) are not directly about the interface, but were found when only the interface and tasks changed. These reasons could be correlated to interface type. This variable was measured between individuals, but notably none of the reasons mention different interfaces were better for different tasks.

ID	Interface	n	Reason
D1	Document	6	Simpler / not too much information
D2	Document	3	Easier to evaluate a document
D3	Document	3	I first want to evaluate the document, then the author
D4	Document	2	Less irrelevant information is combined
D5	Document	1	Focus on what one does, rather than user characteristics
D6	Document	1	More intuitive
D7	Document	1	The tasks were easier*
C1	Candidate	7	Better idea of what a user does
C2	Candidate	2	Focus on user characteristics rather than writing
C3	Candidate	1	I first want to evaluate the author, then the documents
C4	Candidate	1	Have to be less good at selecting keywords*
C5	Candidate	1	Didn't feel like I found who I wanted in the other*
C6	Candidate	1	Results were more relevant*

represent two different search strategies, where one's mental model is either document- or candidate-centric.

In a follow-up questionnaire, seven participants reported consciously modifying their search strategies. Five of these indicated they changed whether they evaluated documents or candidates first. One participant mentioned that in the candidate-centric interface, they searched for topics, whereas in the document-centric interface, they were uncertain of what to search for, and attempted searching by function titles instead. More experience with the system might have affected their search behaviour. The final participant mentioned that the candidate-centric system required them to open more documents before they were certain an author was relevant. Participant p17 succinctly remarked that "in the [candidate-centric] interface, you find experts, and in the [document-centric] interface, you find documents". The order of the document and candidate panels in the interface influences how users evaluate search results, as the perceived retrieval unit changes.

An interesting side-note is that users who preferred simple information did not like when the candidate-centric interface presented irrelevant documents (D4). This occurred when an author had one highly relevant document and a number of tangentially related documents. Although this signals one's limited expertise, these users would prefer just not seeing it.

7 QUANTITATIVE STUDY

The effectiveness, efficiency, and user satisfaction achieved with both search systems were analysed as dependent variables. The independent variables were the interface type, ranking type, and also the interface preference. This was included as the qualitative analysis indicated this is an important variable. Models were constructed in the form of *dependent_variable ∼ interface_type * ranking_type * interface_preference*, with the dependent variable being the task completion rate, time spent, or SUS score.

During six tasks the participants started the timer and then delayed starting the task, as they had a question to the observer. This inflated the starting time between the starting the task and the first query. To correct for this, these false starting times were substituted with the participant's average starting time.

7.1 Effectiveness

An overview of how many of the tasks had correct results are shown in the violin plot in Figure 4, with supplementary effectiveness metrics available in appendix D. A logistic regression tested for significant differences in the task completion rate. Interface type had a significant effect on the task completion rate (p = .044, log odds ratio = −2.25), as the comprehensive overviews in the candidate-centric interface lead to better task completion rates.

There was weak evidence of an interaction effect between the interface type and ranking type (p = .068, log odds ratio = −2.73). The candidate-centric interface performed well when there were many relevant documents per candidate (i.e., with the candidate-centric ranking) but worse when a search result included one highly relevant document as well as slightly relevant or irrelevant documents (as produced by the document-centric ranking). Participants may have ignored relevant authors when they also saw irrelevant documents.

Showing multiple documents per candidate reduces the variance in correct task completion (the distributions in Figure 4 are less tall). The candidate-centric interface appears to provide a more stable signal of a candidate's expertise, although this does not necessarily translate to more correctly completed tasks. This may be because some tasks may require a person with in-depth expertise on a topic (as evidenced by many relevant documents), whereas others require someone with experience in a very specific project (as recorded in specific documents). Further work is necessary to understand when and how conflicting information should be shown.

No significant effects were observed for the ranking type (p = .17, log odds ratio = −1.67) or other factors. The log odds ratio is greater for the interface type than for the ranking type, indicating that the interface type plays a more significant role in task completion.

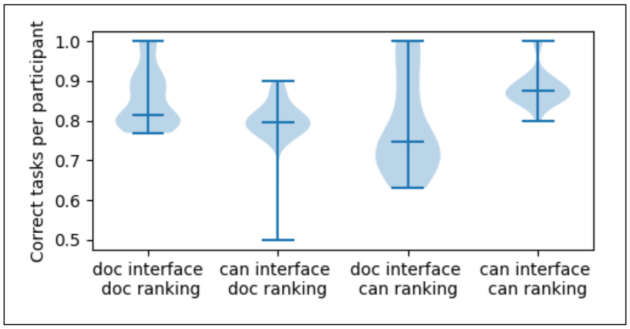


Figure 4: Violin plot showing the distribution of the correctly completed tasks. The ratio (i.e. correct tasks / total tasks) is shown per participant. For each violin plot the outliers and the median are marked.

7.2 Efficiency

An overview of how quickly tasks were performed is shown in the violin plot Figure 5, with additional efficiency metrics available in appendix D. An ANOVA tested for significant differences in the time to task completion. To prepare the data for analysis, we addressed a positive skew in the model residuals. The task completion times were transformed using the function $log|10(time)$. Afterwards, the task completion times no longer violated the normality and variance assumptions of the ANOVA test. The normality assumption was tested using Shapiro's test, which yielded a non-significant result ($F(3, 140) = 0.99, p = .32$). The variance assumption was assessed using Levene's test on the task completion times, which also resulted in a non-significant finding ($F(3, 140) = .19, p = .31$).

The ANOVA indicated a significant effect of the ranking type on task completion times ($F(3, 140) = 4.63, p = .033, \eta_p^2 = .035$), as the document-centric ranking lead to faster completion times. This could be attributed to finding the most relevant pieces of information, leading to more confidence during relevance assessments. This is consistent with previous research showing that document-centric rankings tend to produce more optimal rankings [20], because if the top results include more promising candidates, it can lead to quicker task completion. There is weak evidence indicating that the candidate-centric interface type slows down task completion ($F(3, 140) = 2.80, p = .096, \eta_p^2 = .022$). This could be attributed to the presence of more information that users need to parse and evaluate, potentially leading to longer completion times.

There is also a significant interaction effect between the interface type, ranking type and user's interface preference ($F(3, 140) = 9.08, p = .0031, \eta_p^2 = .06$). This shows that users who prefer different interfaces also need different systems to search as quickly as they can. It might be that interface preference indicates which interface aligns with a user's search strategy, although future work is necessary to understand why. For example, users who want an overview of a candidate might be slower in the document-centric interface if they are looking for multiple documents by the same author. Another explanation for the same finding could be that users lose time if users lose time translating the problem 'who do I need' to the problem 'what kind of documents would this person write'. With more exposure to the system, this individual effect may reduce as users learn to employ the most effective search strategies. Hence we also consider the (non-significant) completion times of different interface and ranking types.

7.3 User satisfaction

The average user satisfaction as measured by the SUS was similar for all systems. An ANOVA found no significant difference on the user responses based on the interface type ($F(3, 15) = 0.047, p = .82, \eta_p^2 = .00037$), and no evidence for an effect of the ranking type ($F(3, 15) = 1.02, p = .31, \eta_p^2 = .0080$). Descriptive statistics per system are available in appendix D.

Users who prefer the candidate-centric interface did provide significantly higher SUS scores ($F(3, 15) = 17.7, p = .000048, \eta_p^2 = .12$). The reason for this finding remains unclear. One possible explanation is that users who favour a comprehensive overview might feel more at ease with tasks that involve assessing the overall relevance of a candidate in general.

Additionally, there was a significant interaction effect between the user's interface preference and the ranking type on the SUS scores ($F(3, 15) = 13.16, p = .00041, \eta_p^2 = .094$). Users who favoured the candidate-centric interface provided the highest scores for systems with the document-centric ranking, although the reason behind this remains unclear. This may be because it finds the most relevant pieces of evidence (documents). No further significant effects were found.

8 DISCUSSION

Generalisability of the best ranking type and interface type can be expected at other organisations where 1) novice colleagues need the expertise of colleagues, 2) those expert colleagues document (the outcome of) their work in a shared content system, and 3) the users seek a similar type of expertise as the policy workers at the municipality of Utrecht. Our search tasks (as listed in Appendix C) seek expertise from someone with declarative knowledge ('what is it'), whereas other search tasks might require procedural knowledge ('how to do it') [17]. Future work should investigate different types of expertise, and whether searching for different types of expertise requires different types of support.

It is unclear whether current findings with novice users generalise situations where expert users need to find other experts. Although experienced employees are more likely to already 'know their way around' and find experts through e.g. their network [37],

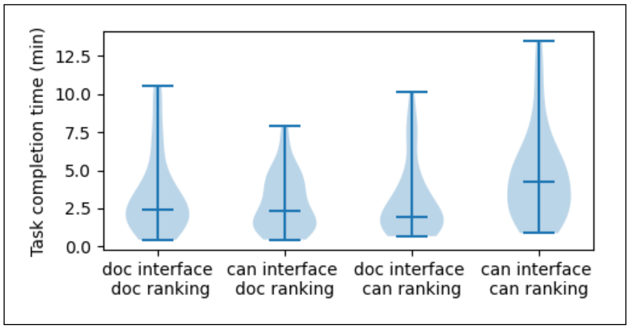


Figure 5: Violin plot showing the distribution of how many minutes each task took. For each violin plot the outliers and the median are marked.

it would be interesting to see when they do struggle with expert search tasks, and whether they execute these differently. We also note that the novice users in our experiment did not make mention of whether candidate experts were still experts, or perhaps were experts in the past. We expect that experienced users would place more emphasis on finding people with recent expertise. Further work could also investigate the effect of how much experience the users have in searching (for expertise) on the preferable interface and ranking.

Limitations of analysis include that our study did not control for the type of retrieval unit and the complexity of retrieval units separately. Future work could change not only the order of the two panels, but also the number of documents shown per search result. Additionally, a post-experiment power analysis revealed that the ANOVA for efficiency was underpowered. Consequently, the analysis might have missed significant effects, and the effect sizes in statistical tests could be overestimated. This occurred because the initial findings showed a considerably lower variance in task completion time than the full dataset (0.53 minutes instead of 2.02 minutes). To achieve a power of 0.95 for interaction effects with variance we find now, the study would require a sample size of 290 participants.

User interface preferences of individuals were associated with two factors, although it is unclear what causes these preferences. This may be due to differences in cognitive styles that affect the processing of information [34] and search strategies [3, 24]. Holistic individuals, for example, tend to focus on the big picture and may be more inclined to prefer the comprehensive overview provided by the candidate-centric interface. Serialistic individuals, on the other hand, tend to approach tasks analytically in individual steps. In this study we measured the overall preferences of individuals, but we found no evidence that users preferred different interfaces for different tasks (it would be found in Table 1). The preferable user interface likely depends on the task performed, and for our set of expert search tasks user preferences appear to be stable. Future work could investigate how a task needs to change to affect user interface preferences.

Combining the optimal ranking and interface types may be impossible, as we found weak evidence of an interaction effects for effectiveness. In the tested systems we need to choose between the (more effective) candidate-centric interface or the (more efficient) document-centric ranking. We argue that, when searching for an internal colleague, approaching the correct candidate is more important, as approaching a person without expertise will lose time and potentially incur a social cost in wasting someone's time (RQ4). This argument will not hold in other expert search contexts, as others settings include hiring or selling to candidates. In these cases, the user might be more concerned with identifying true positives regardless of whether their candidate wants to be approached. This interaction effect likely followed from presenting search results that included conflicting information (one highly relevant document and additional slightly relevant documents). It may be possible to design an interface that does has no interaction effect with the ranking function, allowing for a combination of the strengths of finding highly relevant evidence (in the ranking) while also concisely displaying an overview of the author (in the interface). A step in this direction could be to show each document's relevance in the interface.

Further observations include that there are less authors than documents (as illustrated in Figure 1), and that presenting less search results in total might be preferential, especially during high-recall tasks. A final consideration is that expert search tasks can be directed at different types of expertise, such as procedural knowledge (e.g. 'how to do this') or declarative knowledge (e.g. 'what is this') (see e.g. [18]). The current study focuses on the latter, and found that a candidate-centric interface is better for finding declarative expertise. Procedural tasks tend to have different relevance criteria that can be included in the interface, such as first-hand experience [7, 8].

9 CONCLUSION

Presenting search results in interface as documents or as experts with an overall expertise affects search behaviour. Similarly, the ranking of experts by individual documents or their overall expertise affects whether the order of search results is more effective. Our study compared two types of interfaces and two types of ranking functions. The four combinations were evaluated using simulated tasks, based on task performance and questionnaires.

The document-centric interface presented documents as the retrieval unit, with the author characterised in a secondary panel. This was compared to a candidate-centric search result presentation where the retrieval unit presented was a candidate, and where the secondary panel presented up to three documents written by this author. A document-centric and a candidate-centric ranking function were implemented by indexing and searching for results at either document-level or candidate-level (the latter by appending all of an author's documents as a string).

A qualitative analysis found that users disagreed on which interface was preferable. Two variables affected this preference: the perceived retrieval unit and the complexity of the retrieval unit. Changing the retrieval unit affected how participants searched, as they first assessed the relevance of the retrieval unit in the left-hand panel and then used the right-hand panel as further evidence of a result. Whereas some users preferred the simplicity of the a single document per search result, others appreciated the overview given by the more complex candidate-centric retrieval units. Although users may prefer different retrieval units based on the users' characteristics, we suggest to instead design retrieval units that elicit desired search behaviour. The quantitative analysis investigates which interface results in successful search behaviour.

As interface preference was related to the users' search strategies it was included in the quantitative analysis. 144 tasks were analyzed, performed by eighteen participants, resulting in three main findings. The candidate-centric interface leads to higher rates of correct task completion ($p = .044$, log odds ratio = -2.25). The document-centric ranking leads to faster task completion ($F(3, 140) = 4.63, p = .33, \eta_p^2 = .035$). Finally, there are significant interaction effects between the type of interface, type of ranking and the user's interface preference.

The document-centric ranking is faster (RQ1), which may be because the top results contained more relevant candidates. The candidate-centric interface is more effective (RQ2), probably because it provides a more comprehensive overview. There was weak evidence of an interaction effect between the document type and ranking type for effectiveness, implying that it is not possible to simply match the best interface type with the best ranking type (RQ3). Instead there is a need to combine the strengths of both approaches. This would be a system that retrieves evidence with a high precision (document-centric ranking) and displays an overview of the expert (candidate-centric interface) that is not not too complex. When choosing between effectiveness and efficiency, we argue that approaching appropriate candidates is arguably more important than finding a candidate expert quickly. When working with internal colleagues both saves time for the user and avoids a potential social cost (RQ4).

The implications of this study for designing (expert) search systems are 1) the perceived retrieval unit and its complexity should be appropriate for the current task and user, which for expert search means that 2) users can be nodded towards a search strategy where they thoroughly evaluate candidates by presenting thorough overviews of experts, and 3) the presentation of search results appears more important than the order of search results in terms of task completion. In conclusion, our study resulted in a more effective, albeit less efficient, expert search system for the municipality of Utrecht. Similar design choices may be expected to yield similar results at other organisations where domain novice users search for colleagues with expertise.

REFERENCES

[1] Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. 2006. Formal models for expert finding in enterprise corpora. In *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 6-11, 2006*, Efthimis N. Efthimiadis, Susan T. Dumais, David Hawking, and Kalervo Järvelin (Eds.). ACM, Seattle, Washington, USA, 43–50. https://doi.org/10.1145/1148170.1148181

[2] Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, and Luo Si. 2012. Expertise Retrieval. *Found. Trends Inf. Retr.* 6, 2-3 (2012), 127–256. https://doi.org/10.1561/1500000024

[3] David Bawden and Lyn Robinson. 2011. Individual differences in information-related behaviour: what do we know about information styles? *New directions in information behaviour* 1 (2011), 127–158.

[4] Nicholas J. Belkin. 2015. People, Interacting with Information. *SIGIR Forum* 49, 2 (2015), 13–27. https://doi.org/10.1145/2888422.2888424

[5] Mark Berger, Jakub Zavrel, and Paul Groth. 2020. Effective distributed representations for academic expert search. In *Proceedings of the First Workshop on Scholarly Document Processing, SDP@EMNLP 2020, November 19, 2020*, Muthu Kumar Chandrasekaran, Anita de Waard, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Eduard H. Hovy, Petr Knoth, David Konopnicki, Philipp Mayr, Robert M. Patton, and Michal Shmueli-Scheuer (Eds.). Association for Computational Linguistics, Online, 56–71. https://doi.org/10.18653/v1/2020.sdp-1.7

[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[7] Bogeum Choi, Jaime Arguello, and Robert Capra. 2023. Understanding Procedural Search Tasks "in the Wild". In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR 2023, March 19-23, 2023*, Jacek Gwizdka and Soo Young Rieh (Eds.). ACM, Austin, TX, USA, 24–33. https://doi.org/10.1145/3576840.3578302

[8] Bogeum Choi, Sarah Casteel, Robert Capra, and Jaime Arguello. 2022. Procedural Knowledge Search by Intelligence Analysts. In *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, March 14 - 18, 2022*, David Elsweiler (Ed.). ACM, Regensburg, Germany, 169–179. https://doi.org/10.1145/3498366.3505810

[9] Sujatha Das Gollapalli, Prasenjit Mitra, and C. Lee Giles. 2012. Similar researcher search in academic environments. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, June 10-14, 2012*, Karim B. Boughida, Barrie Howard, Michael L. Nelson, Herbert Van de Sompel, and Ingeborg Sølvberg (Eds.). ACM, Washington, DC, USA, 167–170. https://doi.org/10.1145/2232817.2232849

[10] Rodrigo Gonçalves and Carina Friedrich Dorneles. 2019. Automated Expertise Retrieval: A Taxonomy-Based Survey and Open Issues. *ACM Comput. Surv.* 52, 5 (2019), 96:1–96:30. https://doi.org/10.1145/3331000

[11] Shuguang Han, Daqing He, Jiepu Jiang, and Zhen Yue. 2013. Supporting exploratory people search: a study of factor transparency and user control. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, October 27 - November 1, 2013*, Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi (Eds.). ACM, San Francisco, CA, USA, 449–458. https://doi.org/10.1145/2505515.2505684

[12] Shuguang Han, Danchen Zhang, Daqing He, and Qikai Cheng. 2016. User exploration of slider facets in interactive people search system. *IConference 2016 Proceedings* 1, 1 (2016), 5 pages.

[13] Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2017. Dynamic Factual Summaries for Entity Cards. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, Shinjuku, Tokyo, Japan, 773–782. https://doi.org/10.1145/3077136.3080810

[14] Morten Hertzum. 2014. Expertise seeking: A review. *Inf. Process. Manag.* 50, 5 (2014), 775–795. https://doi.org/10.1016/j.ipm.2014.04.003

[15] Katja Hofmann, Krisztian Balog, Toine Bogers, and Maarten de Rijke. 2010. Contextual factors for finding similar experts. *J. Assoc. Inf. Sci. Technol.* 61, 5 (2010), 994–1014. https://doi.org/10.1002/asi.21292

[16] Omayma Husain, Naomie Salim, Rose Alinda Alias, Samah Abdelsalam, and Alzubair Hassan. 2019. Expert finding systems: A systematic review. *Applied Sciences* 9, 20 (2019), 4250.

[17] Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn - Integration of Information Seeking and Retrieval in Context.* The Kluwer International Series on Information Retrieval, Vol. 18. Springer, Dordrecht, Netherlands. https://doi.org/10.1007/1-4020-3851-8

[18] Peter Ingwersen and Kalervo Järvelin. 2006. *The turn: Integration of information seeking and retrieval in context.* Vol. 18. Springer Science & Business Media, Dordrecht, Netherlands.

[19] W Iso. 1998. 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs). *The international organization for standardization* 45, 9 (1998), 22 pages.

[20] Udo Kruschwitz and Charlie Hull. 2017. Searching the Enterprise. *Found. Trends Inf. Retr.* 11, 1 (2017), 1–142. https://doi.org/10.1561/1500000053

[21] James R. Lewis. 2018. The System Usability Scale: Past, Present, and Future. *Int. J. Hum. Comput. Interact.* 34, 7 (2018), 577–590. https://doi.org/10.1080/10447318.2018.1455307

[22] Ruud Liebregts and Toine Bogers. 2009. Design and Evaluation of a University-Wide Expert Search Engine. In *Advances in Information Retrieval, 31th European Conference on IR Research, ECIR 2009, April 6-9, 2009. Proceedings (Lecture Notes in Computer Science, Vol. 5478)*, Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soulé-Dupuy (Eds.). Springer, Toulouse, France, 587–594. https://doi.org/10.1007/978-3-642-00958-7_54

[23] Rennan C. Lima and Rodrygo L. T. Santos. 2022. On Extractive Summarization for Profile-centric Neural Expert Search in Academia. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, Madrid, Spain, 2331–2335. https://doi.org/10.1145/3477495.3531713

[24] Han-Chin Liu. 2018. Investigating the impact of cognitive style on multimedia learners' understanding and visual search patterns: an eye-tracking approach. *Journal of Educational Computing Research* 55, 8 (2018), 1053–1068.

[25] Marianne Lykke, Ann Bygholm, Louise Bak Søndergaard, and Katriina Byström. 2022. The role of historical and contextual knowledge in enterprise search. *J. Documentation* 78, 5 (2022), 1053–1074. https://doi.org/10.1108/JD-08-2021-0170

[26] Vítor Mangaravite and Rodrygo L. T. Santos. 2016. On Information-Theoretic Document-Person Associations for Expert Search in Academia. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, July 17-21, 2016*, Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel (Eds.). ACM, Pisa,Italy, 925–928. https://doi.org/10.1145/2911451.2914751

[27] Junichiro Mori, Nathalie Basselin, Alexander Kröner, and Anthony Jameson. 2008. Find me if you can: designing interfaces for people search. In *Proceedings of the 13th International Conference on Intelligent User Interfaces, IUI 2008, January 13-16, 2008*, Jeffrey M. Bradshaw, Henry Lieberman, and Steffen Staab (Eds.). ACM, Gran Canaria, Canary Islands, Spain, 377–380. https://doi.org/10.1145/1378773.1378834

[28] Harumi Murakami, Hiroshi Ueda, Shin'ichi Kataoka, Yuya Takamori, and Shoji Tatsumi. 2010. Summarizing and Visualizing Web People Search Results. In *ICAART 2010 - Proceedings of the International Conference on Agents and Artificial Intelligence, Volume 1 - Artificial Intelligence, January 22-24, 2010*, Joaquim Filipe, Ana L. N. Fred, and Bernadette Sharp (Eds.). INSTICC Press, Valencia, Spain, 640–643.

[29] Jakob Nielsen. 2012. Usability 101: Introduction to usability. https://www.nngroup.com/articles/usability-101-introduction-to-usability/

[30] Anne Oeldorf-Hirsch, Brent J. Hecht, Meredith Ringel Morris, Jaime Teevan, and Darren Gergle. 2014. To search or to ask: the routing of information needs between traditional search engines and social networks. In *Computer Supported Cooperative Work, CSCW '14, February 15-19, 2014*, Susan R. Fussell, Wayne G. Lutters, Meredith Ringel Morris, and Madhu C. Reddy (Eds.). ACM, Baltimore, MD, USA, 16–27. https://doi.org/10.1145/2531602.2531706

[31] Georg Pardi, Steffen Gottschling, Peter Gerjets, and Yvonne Kammerer. 2023. The moderating effect of knowledge type on search result modality preferences in web search scenarios. *Computers and Education Open* 4 (2023), 100126.

[32] Sharoda A. Paul. 2016. Find an Expert: Designing Expert Selection Interfaces for Formal Help-Giving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, May 7-12, 2016*, Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade (Eds.). ACM, San Jose, CA, USA, 3038–3048. https://doi.org/10.1145/2858036.2858131

[33] Marisa Peacock. 2009. The search for expert knowledge continues. https://www.cmswire.com/cms/enterprise-cms/the-search-for-expert-knowledge-continues-004594.php

[34] Richard Riding and Indra Cheema. 1991. Cognitive Styles—an overview and integration. *Educational Psychology* 11, 3-4 (1991), 193–215. https://doi.org/10.1080/0144341910110301 arXiv:https://doi.org/10.1080/0144341910110301

[35] Nirmal Roy, David Maxwell, and Claudia Hauff. 2022. Users and Contemporary SERPs: A (Re-)Investigation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, Madrid, Spain, 2765–2775. https://doi.org/10.1145/3477495.3531719

[36] Miamaria Saastamoinen and Sanna Kumpulainen. 2014. Expected and materialised information source use by municipal officials: intertwining with task complexity. *Inf. Res.* 19, 4 (2014). http://www.informationr.net/ir/19-4/paper646.html

[37] Thomas Schoegje, Arjen de Vries, Lynda Hardman, and Toine Pieters. 2023. Improving the Effectiveness and Efficiency of Web-Based Search Tasks for Policy Workers. *Information* 14, 7 (2023), 371.

[38] Thomas Schoegje, Arjen P. de Vries, and Toine Pieters. 2022. Adapting a Faceted Search Task Model for the Development of a Domain-Specific Council Information Search Engine. In *Electronic Government - 21st IFIP WG 8.5 International Conference, EGOV 2022, September 6-8, 2022, Proceedings (Lecture Notes in Computer Science, Vol. 13391)*, Marijn Janssen, Csaba Csáki, Ida Lindgren, Euripides N. Loukis, Ulf Melin, Gabriela Viale Pereira, Manuel Pedro Rodríguez Bolívar, and Efthimios Tambouris (Eds.). Springer, Linköping, Sweden, 402–418. https://doi.org/10.1007/978-3-031-15086-9_26

[39] Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2022. From linear to non-linear: investigating the effects of right-rail results on complex SERPs. *Advances in Computational Intelligence* 2, 1 (2022), 14.

[40] Pertti Vakkari. 2020. The Usefulness of Search Results: A Systematization of Types and Predictors. In *CHIIR '20: Conference on Human Information Interaction and Retrieval, March 14-18, 2020*, Heather L. O'Brien, Luanne Freund, Ioannis Arapakis, Orland Hoeber, and Irene Lopatovska (Eds.). ACM, Vancouver, BC, Canada, 243–252. https://doi.org/10.1145/3343413.3377955

[41] Pertti Vakkari, Michael Völske, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Modeling the usefulness of search results as measured by information use. *Inf. Process. Manag.* 56, 3 (2019), 879–894. https://doi.org/10.1016/j.ipm.2019.02.001

[42] Dan Wu, Shu Fan, and Fang Yuan. 2021. Research on pathways of expert finding on academic social networking sites. *Inf. Process. Manag.* 58, 2 (2021), 102475. https://doi.org/10.1016/j.ipm.2020.102475

[43] Zimeng Yang, Song Yan, Abhimanyu Lad, Xiaowei Liu, and Weiwei Guo. 2021. Cascaded Deep Neural Ranking Models in LinkedIn People Search. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, November 1 - 5, 2021*, Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, Virtual Event, Queensland, Australia, 4312–4320. https://doi.org/10.1145/3459637.3481899

[44] Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowl. Inf. Syst.* 53, 2 (2017), 297–336. https://doi.org/10.1007/s10115-017-1042-4

Received 19 October 2023; accepted 12 December 2023; revised 12 January 2024