Using CoST on self-assessment domain expertise in complex search tasks

Cheyenne Dosso[1], Jose G. Moreno[2], Aline Chevalier[1] and Lynda Tamine[2]

[1]*Université Jean-Jaures, CLLE, Toulouse, France*
[2]*University of Toulouse, IRIT, UMR 5505 CNRS, F-31000, Toulouse, France*

Abstract

While great progress is made in the area of information access, there are still open issues that involve designing intelligent systems supporting task-based search. Despite the importance of task-based search, the information retrieval and information science communities still feel the lack of open-ended and annotated datasets that enable the evaluation of a number of related facets of search tasks in downstream applications. Existing datasets are either sampled from large-scale logs but provide poor annotations, or sampled from lower-scale user studies but focus on ranked list evaluation. In this work, we briefly present *CoST*[1]: a novel richly annotated dataset for evaluating complex search tasks, collaboratively designed by researchers from the computer science and cognitive psychology domains, and intended to answer a wide range of research questions dealing with task-based search. *CoST* collection has been entirely detailed in a previous paper [1]. We report here its main design methodology, characteristics of the data provided and illustrative evaluations showing its importance to the IR community, among which a new evaluation (in comparison to [1]) related to the impact of user's domain expertise on his self-assessment about task complexity.

Keywords

Complex search task, Expertise, User study, Evaluation

1. Introduction

Over the years, users' search activity has been increasingly diversified from solving simple and well-defined tasks such as fact finding, to more complex and knowledge-oriented tasks such as learning and decision-making [2, 3, 4, 5]. These complex tasks generally involve richer search interactions requiring the mobilization of cognitive resources on the part of the user. While search systems are well adapted for simple tasks, they do not suitably assist users to solve complex tasks.

Early attempts to fill this gap is through TREC and CLEF evaluation compaigns such as TREC Interactive tracks (1997–2002) [6], followed later by the TREC Session Tracks (2011–2014) [7]. Recently, the TREC Dynamic Domain Track (2015–2017) [8], TREC Tasks track (2015–2017) [9] and the CLEF Dynamic Search Lab (2017–2018) [10] have also allowed a significant progress

[1]The data collection is available at https://doi.org/10.6084/m9.figshare.15286353 and fully described in [1].

CIRCLE'22: Joint Conference of the Information Retrieval Communities in Europe, July 04–07, 2022, Samatan, Gers, France

cheyenne.dosso@univ-tlse2.fr (C. Dosso); jose.moreno@irit.fr (J. G. Moreno); aline.chevalier@univ-tlse2.fr (. A. Chevalier); tamine@irit.fr (L. Tamine)
0000-0002-8852-5797 (J. G. Moreno)

in this research area. Other attempts are exemplified by collections sampled from the publicly available AOL query log [11] providing human annotation of within-session tasks [12] and cross-session tasks [13, 14]. Recently, Volske et al. [15] published a large-scale search AOL log annotated with cross-user task identifiers, extended with Google and Bing query suggestions.

The AOL logs and TREC tracks have a design focus on shared ranking tasks and does not provide complete and systematic data on search tasks, sessions and users. Through an open design methodology, the CoST collection [1] addresses this limitation by providing rich task and session data. It is worth of mention that the CoST collection provides queries annotated by experts in the domain of cognitive psychology and computer science. Finally, it makes possible the evaluation of agnostic or multilingual models since it is the first collection to have been published in French. The CoST dataset includes 5667 queries recorded from 630 task-based sessions that result from a user study involving 70 participants with varying domains of expertise (computer science, medicine, psychology). Among these tasks, 3 are simple fact-finding tasks (designed for evaluation control) and 12 are complex search tasks that are related to 3 domains of expertise: medicine, psychology and computer science.

2. The CoST Data collection

Table 1
Statistics of the *CoST* data collection.

# Search tasks	15
Min/Max/Avg/Std Queries per task	1/60/5.39/6.70
# Sessions and human answers	630
# Queries	5667
# Min/Max/Avg/Std Query length	1/36/3.93/2.45 terms
# Min/Max/Avg/Std Clicks per query	1/24/1.7/1.8

Here, we briefly describe *CoST* as full details of the collection can be found in [1].

Participants. Seventy native French-speaking participants took part in the user study. All of them were experts in one domain and non-experts in the other domains: 25 in computer science, 10 in medicine, 35 in psychology. They had at least a bachelor degree and were asked to complete a MCQ in each domain assessing their level of prior knowledge. All participants had to solve 15 tasks, 5 of which were in their domain of knowledge, and 10 out their domain.

Protocol. The user study followed two main steps. First, participants were asked to complete an online pre-questionnaire containing MCQs, free and informed consent, and socio-demographic questions. Second, participants were asked to perform 15 search sessions to solve tasks varying in complexity according to 5 levels. Three fact-finding tasks [16] where the answer is directly accessible on the SERPs [17]. Three multi-criteria inference tasks [16] requiring the production of inferences by the user to clarify the terms of the statement and the integration of different search criteria to access the answer [17]. Three exploratory learning tasks [18] where the objective is to lead users to acquire new knowledge about a topic. Three decision-making tasks where the objective is to compare a set of information in order to make a final decision [19]. Finally, three problem-solving tasks where the objective is to create a new coherent set of

information from the knowledge acquired during IS [19].

Before each task, users had to fill in a pre-questionnaire. This included the expected difficulty assessment proposed by [20], which evaluates it according to 5 items: difficulty in searching for information using a search engine, difficulty in understanding the information found, difficulty in determining the usefulness of the information found, difficulty in integrating all of the information found into the answer, and difficulty in determining when to stop the search. For each item, the response modality was a 4-pts Lickert scale ranging from "not at all difficult" to "very difficult. Then, participants were asked to determine their familiarity level with the tasks' topics.

Along search sessions, participants used a browser developed for the purpose of recording human-system interactions. It was used to generate logs from which we extracted the data of the search sessions published in *CoST*: 1) the keyboard keys; 2) Mouse clicks; 3) SERPs and visited documents; 4) timestamps in milliseconds (e.g., instant of click on a selected SERPs). From these logs, we extracted search sessions data released in *CoST*. More precisely, the *CoST* sessions mainly include: 1) the identifiers (Id) of the search sessions; 2) Id of the search task with the complexity and domain attributes; 3) the anonymous Id of the users about his/her domain of expertise; 4) Query Id and query textual formulation; 5) SERPs' clicks (i.e., page and rank); 6) URLs of visited documents.

After each task, users were asked to complete a post-questionnaire including an evaluation of the difficulty experienced according to the same 5 items as the expected difficulty [20]. In addition, the post-questionnaire contained different questions concerning users' perceptions on: the quality of the answer provided, the thematic relevance of the search engine, the websites and the documents visited, the usefulness and reliability of the information gathered [21] and the general satisfaction regarding the accomplishment of the task. Table 1 shows the statistics of the *CoST* collection and the full set of data retrieved during the user study and integrated into the *CoST* collection.

Query annotation. The *CoST* collection also provides richly annotated queries based on two main query reformulation strategies [22, 23, 24, 25] : exploration vs. exploitation.
The exploration strategy refers to the regulation and adaptation behaviors of the user's information seeking activity. At the task level, the user might dynamically reframe his goal while the search task evolves, by integrating new incoming information from the online visited content. Exploration allows the opening and initiation of new search paths so that the user processes an additional part of the search space (e.g., moving from one subtask to another with a clear cut-off) [26, 23]. At the query formulation level, an exploration strategy results in a large semantic jump between the content of two successive queries.
The exploitation strategy reflects perseverance behaviors in processing similar information needs during the information seeking activity. At the task level, this strategy allows the deep processing of a previously opened search path initiated with the aim of processing a specific part of the search space [24, 22]. At the query formulation level, exploitation corresponds to a narrow semantic jump between the content of two successive queries.
A total of 5667 queries were double annotated by humans using a three-step annotation process.

For detailed information regarding the annotation process as well as for the confidentiality and anonymization processing procedure, the taxonomies of task complexity, and examples of tasks used, please refer to [1].

Table 2
Summary of ANOVA results for Domain Expertise (DE), and Domain Expertise*Task Complexity (DE*TC).

		Difficulty			Quality			Relevance			Usefulness			Reliability			Satisfaction		
Effects	**F***	***F***	***p***	**N^2p**	***F***	***p***	**N^2p**	***F***	***p***	**N^2p**	***F***	***p***	**N^2p**	***F***	***p***	**N^2p**	***F***	***p***	**N^2p**
DE	F(1,68)	9.77	=.003	0.126	11.1	=.001	0.141	14.2	<.001	0.173	4.75	<.05	0.065	7.35	=.008	0.098	16.6	<.001	0.196
DE*TC	F(4,272)	3.67	=.006	0.051	1.22	n.s		2.47	=.05	0.035	0.58	n.s		1.35	n.s		1.69	n.s	

3. Using the CoST collection to User's Assessments in Complex Search Tasks

In Dosso et al. [1], we presented the performance results of two downstream tasks, namely query task mapping and search strategy identification, using the *CoST* collection. We also studied, using the behavioural data and query annotations provided in the *CoST* collection, the effects of task complexity and domain knowledge of the task on the users' behaviors (ClickSerp, NoClickSerp, TimeSerp, TimeURL, TimeSession) and search strategies (exploration-exploitation). In this paper, we extend this work by analyzing the effects of domain expertise and tasks complexity on: 1) the user's self-assessment of task difficulty, called later perceived difficulty [20] and 2) the users' perceptions [21] (quality of response provided, relevance of information, usefulness of information, reliability of information, and overall satisfaction with task completion). We perform repeated measures of ANOVA on dependent variables cited above link to the pre- and post-questionnaires (See Section 2). We focus on two independent variables. First, the computer science expertise as between-subject factor with two modalities (In domain and Out domain). The experimental group "In domain" includes the 25 computer science students and the group "Out domain" includes the 35 psychology students and the 10 medicine students. Second, we select as within-subject factor the task complexity with 5 modalities (factfinding, exploratory learning, decision-making, problem-solving, multicriteria-inferential). In the case where the ANOVA test is significant, we perform Scheffe's post-hocs. Table 2 presents a summary of the ANOVA results for domain expertise and the interaction between domain expertise and task complexity on users' perceptions. We discuss below the results obtained and the primary findings that emerged from them.

Overall, we can see from Table 2 that the domain expertise in computer science has effects on all the measures (perceived difficulty and users' perceptions). Computer science experts evaluate the tasks as significantly less difficult ($M = 9.9$ $SD = 3.3$) than non-experts do ($M = 11.5$ $SD = 4.8$). Experts rate their final written answer as higher quality ($M = 3$ $SD = 0.7$) than non-experts do ($M = 2.62$ $SD = 0.93$). In addition, domain experts report that the information they accessed during their search sessions is more relevant ($M = 3.04$ $SD = 0.73$), more useful ($M = 4.82$ $SD = 1.52$), and more reliable ($M = 5.44$ $SD = 1$) compared to non-experts in the domain (relevance: $M = 2.71$ $SD = 0.83$; usefulness: $M = 4.43$ $SD = 1.64$; reliability: $M = 4.92$ $SD = 1.3$). Finally, domain experts are more satisfied ($M = 4.8$ $SD = 1.53$) than their counterparts ($M = 3.9$ $SD = 1.9$) with task completion. In summary, domain expertise reduces the perceived difficulty of the tasks regardless of their level of complexity and leads to higher perceived satisfaction with task completion by perceiving: higher quality answers and access to more relevant, useful and reliable information during the search session.

To go further in our analysis, we aim to determine to what extent the interaction of expertise and task complexity have effects on perceived difficulty and user perceptions. For difficulty, we note that it is specifically the problem-solving task and the multicriteria-inferential task that are perceived differently between experts and non-experts. Specifically, computer science perceive these two tasks as less difficult (problem solving task: $M = 12.13$ $SD = 3.11$; multicriteria-inferential task: $M = 10.7$ $SD = 3.4$) than non-experts ($M = 14.7$ $SD = 4.1$ and $M = 13.1$ $SD = 4$ respectively). As for the users' perceptions, only the perception variable related to the relevance of the information retrieved during the search session is significant. Experts in the computer science domain judge that they had access to more relevant information for the problem solving task ($M = 3$ $SD = 0.7$) than non-experts ($M = 2.23$ $SD = 0.9$). For [19], it is important to consider both the objective complexity of the tasks and the subjective complexity which refers to the difficulty of the tasks perceived by the users to understand the dynamics of solving complex tasks. Here, we find that two of the most complex tasks in the *CoST* collection (i.e. problem solving [19] and multicriteria-inferential [16, 17]) are perceived as more difficult by users without domain expertise in computer science. Going further, we find that for the problem solving task the non-experts felt they had access to less relevant information than the domain experts. By coupling the results of difficulty and perceived relevance, we can argue that experts were able to access more relevant information for the task thanks to their previous knowledge of the domain and thus felt that they could solve it with more ease than non-experts of the domain.

4. Conclusions and future work

In this paper, we briefly presented the *CoST* collection, specifically useful for task-based search evaluation. The CoST collection includes a wide range of task-related and user-related data and annotations including particularly the cognitive complexity of tasks, the domain expertise of participants and query type annotations. These critical attributes allow a wide range of experiments for researchers from different fields including but not limited to IR, IS and psychology.

Acknowledgements

This work was supported by the Agence National de la Recherche (ANR), through project CoST (https://www.irit.fr/COST/), code ANR-18-CE23-0016.

We thank Claire Ibarboure for applying the double annotation of the queries.

References

[1] C. Dosso, J. G. Moreno, A. Chevalier, L. Tamine, Cost: An annotated data collection for complex search, in: Proceedings of the 30th ACM International on Conference on Information and Knowledge Management, CIKM '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 4455–4464. URL: https://doi.org/10.1145/3459637.3481998. doi:10.1145/3459637.3481998.

[2] M.-A. Cartright, R. W. White, E. Horvitz, Intentions and attention in exploratory health search, SIGIR '11, 2011, pp. 65–74.
[3] A. Hassan Awadallah, R. W. White, P. Pantel, S. T. Dumais, Y.-M. Wang, Supporting complex search tasks, CIKM '14, 2014, pp. 829–838.
[4] S. Y. Rieh, K. Collins-Thompson, P. Hansen, H.-J. Lee, Towards searching as a learning process: A review of current perspectives and future directions, Journal of Information Science 42 (2016) 19–34.
[5] N. Belkin, T. Bogers, J. Kamps, D. Kelly, M. Koolen, E. Yilmaz, Second workshop on supporting complex search tasks, CHIIR '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 433–435. URL: https://doi.org/10.1145/3020165.3022163.
[6] P. Over, The trec interactive track: an annotated bibliography, Information Processing & Management 37 (2001) 369–381. URL: https://www.sciencedirect.com/science/article/pii/S0306457300000534. doi:https://doi.org/10.1016/S0306-4573(00)00053-4, interactivity at the Text Retrieval Conference (TREC).
[7] B. Carterette, P. Clough, M. Hall, E. Kanoulas, M. Sanderson, Evaluating retrieval over sessions: The trec session track 2011-2014, SIGIR '16, 2016, pp. 685–688.
[8] G. H. Yang, I. Soboroff, TREC 2016 dynamic domain track overview, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016, volume 500-321 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2016.
[9] E. Kanoulas, E. Yilmaz, R. Mehrotra, B. Carterette, N. Craswell, P. Bailey, TREC 2017 tasks track overview, in: Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017, volume 500-324 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2017.
[10] E. Kanoulas, L. Azzopardi, G. H. Yang, Overview of the clef dynamic search evaluation lab 2018, in: P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2018, pp. 362–371.
[11] G. Pass, A. Chowdhury, C. Torgeson, A picture of search, InfoScale '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 1–es.
[12] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, G. Tolomei, Identifying task-based sessions in search engine query logs, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11, 2011, p. 277–286.
[13] P. Sen, D. Ganguly, G. Jones, Tempo-lexical context driven word embedding for cross-session search task extraction, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 283–292.
[14] M. Hagen, J. Gomoll, A. Beyer, B. Stein, From search session detection to search mission detection, in: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, OAIR '13, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013, p. 85–92.
[15] M. Völske, E. Fatehifar, B. Stein, M. Hagen, Query-task mapping, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, 2019, p. 969–972.

[16] D. J. Bell, I. Ruthven, Searcher's assessments of task complexity for web searching, Lecture Notes in Computer Science (2004) 57–71. doi:https://doi.org/10.1007/978-3-540-24752-4_5.

[17] M. Sanchiz, A. Chevalier, F. Amadieu, How do older and young adults start searching for information? impact of age, domain knowledge and problem complexity on the different steps of information searching, Computers in Human Behavior 72 (2017) 67–78. doi:https://doi.org/10.1016/j.chb.2017.02.038.

[18] M. Gary, Exploratory search: from finding to understanding, Communications of the ACM 49 (2006) 41–46. doi:https://doi.org/10.1145/1121949.1121979.

[19] C. D. J., Task complexity: A review and analysis, Academy of Management Review 13 (1988) 40–52. doi:https://doi.org/10.5465/amr.1988.430677.

[20] W.-C. Wu, D. Kelly, A. Edwards, J. Arguello, Grannies, tanning beds, tattoos and nascar : Evaluation of search tasks with varying levels of cognitive complexity, in: Proceedings of the 2012 Information Interaction in Context, IIiX '12, Association for Cumputing Machinery, New York, NY, USA, 2012, pp. 254–257. doi:https://doi.org/10.1145/2362724.2362768.

[21] J. Jiang, D. He, D. Kelly, J. Allan, Understanding ephemeral state of relevance, in: Proceedings of the 2017 Conference on Human Information Interaction & Retrieval, CHIIR '17, Association for Cumputing Machinery, New York, NY, USA, 2017, pp. 137–146. doi:https://doi.org/10.1145/3020165.3020176.

[22] M. Sanchiz, F. Amadieu, A. Chevalier, An evolving perspective to capture individual differences related to fluid and crystallized abilities in information searching with a search engine, in: W. T. Fu, H. van Oostendorp (Eds.), Understanding and Improving Information Search: A Cognitive Approach, Human-Computer Interaction, 1 ed., Springer, Cham, Switzerland, 2020, pp. 71–96.

[23] J. Liu, S. Sarkar, C. Shah, Identifying and predicting the states of complex search tasks, in: Proceedings of the 2020 Conference on Human Information Interaction & Retrieval, CHIIR '20, Association for Cumputing Machinery, New York, NY, USA, 2020, pp. 193–202. doi:https://doi.org/10.1145/3343413.3377976.

[24] B. J. Jansen, D. L. Booth, A. Spink, Patterns of query reformulation during web searching, J. Am. Soc. Inf. Sci. Technol. 60 (2009) 1358–1371.

[25] Y. He, J. Tang, H. Ouyang, C. Kang, D. Yin, Y. Chang, Learning to rewrite queries, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, 2016, pp. 1443–1452.

[26] B. M. Wildemuth, D. Kelly, E. Boettcher, E. Moore, G. Dimitrova, Examining the impact of domain and cognitive complexity on query formulation and reformulation, Information Processing & Management 54 (2018) 433–450.