CoST: An annotated Data Collection for Complex Search

Cheyenne Dosso, Jose G. Moreno, Aline Chevalier, Lynda Tamine

CoST: An annotated Data Collection for Complex Search

Cheyenne Dosso
cheyenne.dosso@univ-tlse2.fr
Universite Jean-Jaures, CLLE
Toulouse, France

Jose G. Moreno
jose.moreno@irit.fr
Universite Paul Sabatier, IRIT
Toulouse, France

Aline Chevalier
aline.chevalier@univ-tlse2.fr
Universite Jean-Jaures, CLLE
Toulouse, France

Lynda Tamine
lynda.tamine@irit.fr
Universite Paul Sabatier, IRIT
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 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* includes 5667 queries recorded in 630 task-based sessions that result from a user study involving 70 french native participants who are expert in one among 3 different domains of expertise (computer science, medicine, psychology). Each participant completed 15 tasks with 5 different types of cognitive complexity (fact-finding, exploratory learning, decision-making, problem-solving, multicriteria-inferential). In addition to search data (e.g., queries and clicks), *CoST* provides task and session-related data, task annotations and query annotations. We illustrate possible usages of *CoST* through the evaluation of query classification models and the understanding of the effect of task complexity and domain on user's search behavior.

KEYWORDS

Complex search task, Expertise, User study, Evaluation

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

https://doi.org/10.1145/3459637.3481998

1 INTRODUCTION

Search systems are the main access to the world-scale digital library, allowing people to achieve tasks triggered from problematic real-world situations. As a result, the range and level of complexity of search tasks span from simple ones (e.g., fact finding), to more intensive knowledge oriented tasks, well known today as *complex search tasks* [1, 7, 16, 33] (e.g., decision-making, planning). Such tasks induce multidimensional search interactions and are generally structured in multiple subtasks with many cognitive resources involved. Clearly, to tackle complex tasks by means of search systems, users would require support for task achievement that extends well beyond the list of 'ten blue links'. While research investigation in the area of task-based search gained maturity since the early works rooted in the interactive Information Retrieval (IR) community [39], the topic still attracts an increasing interest by researchers, as acknowledged by the several recent related workshops among which: *Supporting Complex Search Tasks* (2015-2017) [1], *Learning from User Interactions* (2018) [28], *Task Intelligence* (2019) [15] and *Investigating Learning During Web Search* (2020)[2].

The IR community has a strong tradition of making experimental resources available for re-use, more particularly test collections built-up on shared tasks. Early initiatives which specifically focused on the evaluation of user-system interactions while achieving tasks, include TREC Interactive tracks (1997–2002) [31], followed later by the TREC Session Tracks (2011–2014) [6]. Recently, the TREC Dynamic Domain Track (2015–2017) [45], TREC Tasks track (2015–2017) [22] and the CLEF Dynamic Search Lab (2017–2018) [21] have also brought significant benefits to the research progress in this area. However, the information seeking (IS) and IR communities still feel the lack of richly annotated datasets for understanding and evaluating multiple aspects of task-based search. We identify two barriers to the lack of publicly available data collections. The first one is privacy and the second one is the cost underlying user studies and associated human annotations. To address this issue and contribute to a faster research progress in this area, we introduce *CoST*, a data collection built-up from a relatively large-scale user study, particularly designed for the evaluation of complex search tasks. Unlike the previous TREC and CLEF-like resources cited

[2]https://iwilds2020.wordpress.com/

above which are designed considering the traditional shared task structure, *CoST* is open-ended, putting task complexity and task doers (i.e., users) at the center of the experimental design. This type of resource design allows better standardization and an increasing level of collection re-use as highly recommended in the interactive IR community [3]. *CoST* includes 5667 queries recorded in 630 task-based sessions that result from a user study involving 70 participants. Each participant is expert either in the computer science, medicine or psychology domain and completes 15 search tasks. Among those tasks, 3 belong to simple fact-finding search tasks (designed for evaluation control) and the remainder 12 ones are complex search tasks which are related to his/her own domain expertise, as well as out of his/her domain expertise. As a result, each user achieved the same tasks, with the aim of allowing fair comparative evaluation. We consider 5 types of cognitive complexity following three taxonomies [2, 12, 18]: fact-finding task [2], exploratory learning task [12], decision-making, problem solving task [18] and multicriteria-inferential task [2]. In addition to all the data related to search sessions, user expertise, task complexity and participant's final answers at the end of their task completion, we provide human assessments about the queries (exploration, exploitation, narrow exploitation, spelling correction) and expected subtasks for each task w.r.t. the domain-knowledge involved in the task. We also provide all the experimental material including non personal answers to pre-task and post-task questionnaires. We release the complete session and tasks data that fit with all the requirements of privacy policy, the human annotations and the design procedure documentation. Our ultimate goal is to encourage reproducible and replicable experiments which greatly reduce the effects of privacy and cost barriers in task-based search evaluation.

2 RELATED WORK

There have been several attempts at making available Web search query logs splitted into session-based tasks. The first attempts are exemplified by collections sampled from the publicly available AOL query log [32]. Lucheese et al. [25] is a dataset sampled from AOL including 1424 queries submitted by 13 users. The queries were first clustered from 307 time-gap sessions. The tasks were annotated by humans within each session, leading to a total number of 554, with an average of 2.57 queries per task. Task labels across sessions are necessarily different. Each task is defined by a tag and optionally with a longer textual description. Sen et al. released the Extraction (CSTE) dataset [37] which extends the dataset released by Lucheese et al. [25] by a cross-session annotation. Thus the time gap session was not applicable and the annotators were instructed to re-label task identifiers spanning across different query sessions. The Webis Search Mission Corpus 2012 (WSMC12) dataset [14] comprises 8800 queries with associated clicks, submitted by 127 unique users. The queries have been annotated within and across sessions leading to a total of 1378 tasks with an average of 3.45 queries per task. Task annotation is limited to a numeric identifier. However, in all the above datasets the annotations were made per-user even if the underlying search intents of a subset of queries are shared across users. Recently, Volske et al. [41] tackled this limit. The authors released a large-scale web-search log extracted from the WSMC12 and Lucchese et al. [25] datasets with associated task identifiers across users and then extended them with Google and Bing query suggestions resulting in 7202 across-user tasks with an average of 172 queries per-task.

The TREC and CLEF evaluation initiatives attempted to address the need of task-based search datasets allowing the evaluation of shared tasks. TREC Session [6] allows evaluating whether systems can improve their effectiveness for a target query by considering previous user-system interactions. The dataset includes: 80 topics and corresponding 1.282 sessions in total, leading up to the test query for each session; the top ranked list of documents from the ClueWeb12 collection and for each past query, the user's clicking behavior. In TREC 2012-2014, sessions are classified based on two facets of complexity: product and goal. These two facets produce four types of sessions: known-item (factual specific), interpretive (intellectual specific), known-subject (factual amorphous) and exploratory (intellectual amorphous). The graded relevance labels of a document were provided per topic. With a more system-centered approach, the main objective of TREC Dynamic Domain Track [45] and the TREC Tasks track [22] is to optimize user-system interactions involved in complex search tasks. In the TREC Dynamic Domain Track [45], optimal rankings w.r.t. provided relevance labels per topic are expected at early stages of the task. The track provides 15 topics (i.e., tasks) descriptions and associated subtopics (i.e., substasks). The TREC Tasks track [22] was designed to evaluate systems supporting multi-aspect or multi-step tasks. The track provides 50 queries designed in such a way that the underlying tasks can be decomposed in a sequence of subtasks. The latter are provided as ground truth with the aim of evaluating system understanding of the tasks. The dataset also includes usefulness and relevance labels of documents with the aim of evaluating systems' document rankings.

While AOL-based datasets allowed a significant progress in task-based search, mainly because of their source size, they critically lack of description about the accomplished tasks. The TREC tracks are a significant step forward towards tackling this limit. However, since their design is mostly driven by shared ranking tasks, none of these datasets fully provides data about both the tasks and the search sessions. By adopting an open-ended benchmark design, the built-up *CoST* data collection tackles this limit. *CoST* is obtained through a user study which involved 70 participants (50% female) who are experts in one of 3 targeted knowledge domains (computer science, medicine, psychology). Each participant had to solve a total of 15 search tasks varying in complexity within and outside of her/his domain expertise. Table 1 shows a comparison of different existing datasets released for the purpose of task-based search evaluation. As can be seen, *CoST* provides rich task and session data, in addition to human annotations that are critical for understanding and evaluating a wide range of both interaction and search models. Besides, as a french collection, it is the first one released in another language than English, making it appropriate for the evaluation of language agnostic or multilingual models. Table 2 shows the statistics of *CoST*.

Table 1: Comparison of existing datasets used for task-based search evaluation in terms of data publicly released. Abbreviations used are: L (Language; En: English; Fr: French); TD (Task Description); STD (SubTask Description); TC (Task Complexity); TDK (Task Domain Knowledge); QC (per Query-Task Class); EX (user's EXpertise); CL (CLicks); TO (user's Task Outcome); RL (search result Relevance Label). Data availability is indicated by either Y (Yes) or N (No).

Dataset	L	Tasks and Queries					Sessions			
		TD	STD	*TC*	*TDK*	*QT*	*EX*	*CL*	*RL*	*TO*
Lucchese et al. [25]	En	Y	N	N	N	N	N	Y	N	N
WSMC12 [14]	En	N	N	N	N	N	N	Y	N	N
Lucchese et al. [25]	En	N	N	N	N	N	N	Y	N	N
CSTE [37]	En	N	N	N	N	N	N	Y	N	N
Völske et al. [41]	En	N	N	N	N	N	N	N	N	N
Session TREC [6]	En	Y	N	Y	N	N	N	Y	Y	N
Dynamic Domain TREC [45]	En	Y	Y	N	Y	N	N	N	Y	N
Tasks TREC [22]	En	Y	Y	N	N	N	N	N	Y	N
CoST	*Fr*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*Y*	*N*	*Y*

Table 2: 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

3 THE COST DATA COLLECTION

3.1 Participants

A total of 70 french participants took part in our study, each one being expert in one domain knowledge: 25 experts in computer science, 10 experts in medicine and 35 experts in psychology. All of them have at least bachelor degree in their study domain and 43% among them are PhD students. They have been recruited using a campus-wide opt-in mailing list of students over 3 faculties (computer science, psychology and medicine). To check their knowledge level in the domain, participants had to complete a knowledge test in each domain before starting search sessions. Each test has 10 multiple-choice questions established by senior researchers in the respective domains (30 in total), with 5 candidate answers per question (i.e., 1 correct, 3 incorrect and 1 option "I don't know"). All the participants had to solve the same 15 tasks, described below, 5 from his/her own domain of expertise, the remaining 10 ones out of his/her domain of expertise. In these conditions, all of them were at the same time expert or non-expert according to tasks' topics.
In order to guarantee confidentiality and a strict anonymization process, we carried out several data filtering steps. First, all socio-demographic data concerning the participants were removed. More precisely, *CoST* does not allow inferring the age of the participants, their gender, their level of study, their location, their affiliation, and their Internet use habits. A thorough cleaning of the logs was conducted; For example, if a user visited a website requiring login and password, all of these private data were removed. Thus, from the data included in *CoST*, it is not possible to cross-reference the data in such a way as to be able to trace a particular user. In addition, user ID was anonymous and non-nominative. Only domain expertise data is provided: "Csi" (i.e., computer science); "Med" (i.e., medicine); "Psy" (i.e., psychology).

3.2 Protocol

The user study that allowed the release of *CoST* was conducted in three main stages. First, participants were asked to complete an online pre-questionnaire containing the test that assessed their level of knowledge in each domain (i.e., computer science, medicine, psychology). All the participants also completed a free and informed consent form informing them of the terms of use of the data collected during the study. In addition, before starting each task, users had to complete a pre-questionnaire which contains: a questionnaire of expected difficulty as proposed by Wu et al. [44], a question on familiarity with the task topic (4-pt Lickert scale ranging from "not at all familiar" to "very familiar").

Second, participants were invited to take part in the search sessions, using a browser designed to record their logs. They completed 15 search tasks either at their universities or at home and had the option of doing that in multiple episodes and were autonomous in managing their search time. The order of task presentation was counterbalanced from one user to another one, both in terms of task complexity (Section 3.3), and knowledge domain. The format for presenting the search tasks to be completed was paper and pencil. This choice is motivated by the fact that we wanted the computer to be a tool exclusively dedicated to interactions with the search system during the sessions. All the other activities are therefore carried out directly in writing. A browser has been developed to record the interactions between each participant and the search system during the whole duration of the task completion. This browser was close to a real browser but users couldn't open several tabs. It allowed the generation of logs containing: 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.

Third, immediately after completing each search task, participants were asked to complete a post-questionnaire. The latter contains a questionnaire of experienced difficulty following the form proposed in [44]; a self-assessment of the quality of the answer provided (4-pt Lickert scale ranging from "very bad" to "very good"). As suggested by Jiang et al. [20], we propose 3 questions to evaluate the search engine, websites and visited documents in terms of topical relevance, usefulness of the information gathered and participant's feeling about the reliability of the results.
Finally, participants were asked to complete a scale to evaluate their self-satisfaction with the task (7-pt Lickert scale ranging from "absolutely not satisfied" to "extremely satisfied"). All the answers to both the pre- and post-session questionnaires are provided in *CoST*.

3.3 Task Complexity

3.3.1 Background. It is worth of mention that while a number of previous studies argued that the complexity of tasks may be impacted by both the domain in which they fall and doers' expertise [8, 36, 38], the majority of user studies offer a single task per complexity level and/or type. The main reason underlying this limitation is that the multiplication of tasks requires more effort and commitment from the users. Unlikely, in the user study that allowed the construction of *CoST*, the designed experimental protocol (Section 3.2) overcomes this limitation. As outlined above, all participants completed the same 15 tasks with varying complexity. In this work, we particularly address the *cognitive complexity* which relates to the types of mental processes required to complete the tasks as defined in 3 well established taxonomies [2, 12, 18]. Beyond the types of task complexity, we also consider levels of task complexity as assessed by the authors of each taxonomy as well as the authors of this paper from the cognitive psychology domain. This assessment is based on a comparative analysis of the definitions of task complexity provided by each taxonomy w.r.t. to set of 7 relevant criteria: the goal of the task, the statement, the sub-tasks, the outcomes, the level of cognitive resources required, and the level of prior domain knowledge needed to solve the task. This analysis leads us to fix 5 types of task complexity that we ordered by levels across taxonomies: 1) fact-finding task [2]; 2) exploratory learning task [12]; 3) decision-making task [18]; 4) problem-solving task [18] and 5) multicriteria-inferential task [2].

3.3.2 Task design. To allow fair comparison of tasks in terms of cognitive complexity, we indicate below the definitions used in 3 well established taxonomies and according to the 7 criteria cited above.

- *Search tasks adapted from the taxonomy of Bell et al. [2].* Using this taxonomy, we designed *fact-finding* and *multicriteria-inferential* tasks. In *fact-finding* tasks, complexity is manipulated at the level of the task's statement. According to [2], this type of complexity involves clear statements including relevant keywords. The task goal is well-defined and there are few subtasks which are easy to perform. Multiple paths can be led to achieve the goal and the answer is directly accessible from SERPs [35]. It is a closed task where only one answer is right. In addition, a fact-finding task requires neither a high knowledge level in the related domain, nor to mobilize a high number of cognitive resources. In our study, the fact-finding task is designed for evaluation control.
  Unlikely, a *multicriteria-inferential* task has an unclear statement. Users need to have prior domain knowledge about the task topic to achieve the ill-defined goal [2]. The keywords are search criteria to be integrated into the query to allow reaching an answer. In addition, some of those criteria may require mental inferences because they are too fuzzy, ambiguous and unclear [2]. Few paths are possible to reach a unique expected answer. Users have generally to perform several subtasks: 1) select key concepts to be integrated into a query gathering all search criterion; 2) infer new concepts that might replace irrelevant ones; 3) explore different search paths; 4) find information matching to search criterion and aggregate them to identify those that fulfill them [35]. To perform these subtasks, users have to mobilize a high number of cognitive resources allowing to reach their goal with the SERP but also to identify new and more relevant query reformulations.

- *Search tasks adapted from the taxonomy of Marchionini [12].* The third type of task complexity we use is the *exploratory learning task* [12]. The latter is a scenario in which the main goal is to lead users to gain new knowledge about a topic. The task statement is fuzzy and unclear [13] because it does not naturally lead users to perform clear subtasks. Several paths can be proceeded to solve subtasks and several outcomes are acceptable. This task type does not require to have a high level of prior domain knowledge and to mobilize many cognitive resources. For instance, experts can seek specific information about the topic whereas non-expert users can collect basic knowledge. The general statement is *"you want to learn more about..."*. The navigation paths have to be inferred by users such as three subtasks can be completed: 1) understand the main topic; 2) identify various aspects related to the main topic; 3) discover/find out more about various aspects of the main task topic.

- *Search tasks adapted from the taxonomy of Campbell [18].* The fourth and fifth types of task complexity are both work tasks proposed in [18]: *decision-making* tasks and *problem solving tasks*. For *decision-making* tasks, the work goal is to make a decision by evaluating new information collected. The statement is clear but several subtasks have to be achieved: 1) understand the main task topic; 2) understand various aspects of the main task topic; 3) identify the advantages and disadvantages of each aspect; 4) analyze and differentiate the collected arguments; 5) judge arguments/information according to criteria to be established by the user. To perform those subtasks, users can follow several paths but the outcomes heavily depend on criteria inferred by users themselves during the search. The comparison and evaluation work consists in selecting the best solution among several candidate ones to make a decision. This task calls the users to mobilize many cognitive resources and to have a high level of prior domain knowledge to compare information against several criteria inferred during the search.
  Finally, we also designed *problem-solving* tasks [18]. The work task goal is to elaborate and create a new consistent set of information from new knowledge acquired. In other terms, users have to apply information collected from web content to achieve the work goal. The statement is clear but several subtasks have to be achieved: 1) understand the main topic; 2) understand various aspects of the main topic; 3) explore typology and characteristics of various aspects of the main topic; 4) analyze the previously collected information and envision their usage to achieve the task; 5) create a new set of consistent information. Several paths can be followed but users should find the best one to reach the target outcome. In addition, links between paths and outcomes are uncertain even if several outcomes are possible. This task

requires a high number of cognitive resources to perform associated subtasks and to collect relevant information allowing to build a good answer. The level of required prior domain knowledge is high because users cannot directly access the answer.

The different task characteristics tend to impact the level of their complexity [2, 18]. For instance, if the task goal is well-defined, a task will be easier to perform [2]. Also, as the number of paths to the desired outcome increases, the complexity of a task decreases [18]. From the elements presented above, we can establish for each task which characteristics tend to make it more or less complex. To this extent, we can make a hypothesis about the level of complexity of each task. We postulate that the first level would be the fact-finding task, followed by the exploratory learning task, then the decision making task, the problem solving task and finally the inferential multi-criteria task which would be the fifth and final level of complexity.

Examples of tasks in the medicine domain and corresponding subtasks, sessions and answers are presented in Table 3.

3.4 Query Annotation

3.4.1 Background. A number of previous user studies highlighted the impact of different task characteristics on users' query reformulations [24, 34, 38, 43]. The results mainly revealed that: 1) the domain knowledge of the task doer significantly impacts query term changes; 2) the cognitive complexity of the task (e.g., simple, hierarchical, parallel) has a significant effect on users' query reformulation behavior. This motivates us to provide in *CoST*, query annotations that could reveal two main user's search strategies well-known in the IS and IR communities: *exploration* vs. *exploitation* [23, 34] also called generalization vs. specialization [17, 19]. 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) [23, 43]. 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 [19, 34]. At the query formulation level, exploitation corresponds to a narrow semantic jump between the content of two successive queries.

3.4.2 Query annotation process. CoST benefits from a double manual annotation of the 5667 queries by two distinct human annotators. All the queries meet the confidentiality criteria presented in Section 3.2. Both annotators are experts in one among the 3 domains. Specifically, one annotator is an expert in computer science and the other annotator is an expert in psychology. These two annotators also annotated the medicine queries. Under these conditions, the annotations of these queries might be of lower quality than those related to computer science and psychology fields. We describe below the different steps of the query annotation process.

- *Step 1: Presentation of the coding grid and training.* The two expert human annotators (i.e., from computer science and psychology domains) were first trained to use the coding grid. The annotators were instructed to follow the guideline below:
  (1) For each session of n queries q_i, $\{q_1, q_2 \ldots, q_n\}$, observe the first query q_1 and, if any, pairs of successive queries (q_i and q_{i+1}), $i = 1 \ldots n-1$.
  (2) Code q_1 as "1" to indicate *exploration*. Indeed, the user starts the search by exploration.
  (3) Code q_{i+1} as "0" to indicate a reformulation of q_i only to fix spelling errors.
  (4) Code q_{i+1} as "1" to indicate a reformulation for *exploration* in the case where the semantic jump from q_i is qualified as large [17, 19, 34].
  (5) Otherwise, code q_{i+1} as "2" to indicate *exploitation* as an intermediate semantic jump from query q_i but keeping the same search path, without a clear break.
  (6) Otherwise, code q_{i+1} as "3" to indicate *narrow exploitation* in the case of narrow semantic gap observed by the use of lexically similar or semantically close terms (e.g., use of synonyms) in comparison to q_i.

  Following the presentation of the grid, the two annotators practiced coding queries on a representative sample of all the designed tasks and domains. Then, they were invited to discuss their disagreement during meetings, fix the underlying reasons and then homogenize their feelings about the coding.

- *Step 2: Full annotation.* During this step, the annotators had to annotate all the queries produced by the participants in the framework of our study. Each query received a specific code: "0" (i.e., spelling correction), "1" (i.e., exploration), "2" (i.e., exploitation), "3" (i.e., narrow exploitation). Throughout their annotations, both annotators had access to the coding grid, detailed definitions and query examples for each code.

- *Step 3: Annotation validation.* The two annotators coded a total of 5667 queries based on the theoretical and practical elements presented above. The overall agreement of Cohen's Kappa coefficient between annotators is 95% which is considered excellent.

Table 4 provides a summary of the data provided in the full *CoST* data collection release[3].

4 TASK-BASED SEARCH EVALUATION EXPERIMENTS

In this section, we study two main tasks that may be addressed using the human labels provided in *CoST*. First, query-task mapping which is a task recently studied [26, 42] that consists in assigning to a given issued query, the task that it more likely belongs to, among

[3]The CoST dataset is available at https://doi.org/10.6084/m9.figshare.15286353

Table 3: Sample of tasks and subtasks in the medicine domain with their respective complexity type, session, and answer. The subtasks (numerated list in second column) were not accessible to the participants during the study. Full descriptive data is provided in the CoST data collection. The '→' symbol is used to separate queries within the session. Only the answer is automatically translated to English.

Task Type	Task Examples and associated subtasks (from Medicine domain)	Session	Answer
Decision-making (TDMed)	*An 83-year-old woman had a non-sequelae stroke 5 months ago.* *At the stroke assessment, atrial fibrillation was discovered.* *She had dropped 3 times in the last 2 months.* *Should anticoagulant therapy be initiated?* *After having evaluated the benefit-risk ratio of the initiation or not of an anticoagulant treatment, select the management that seems best to you and justify your choices.* 1. Understanding non-sequential stroke 2. Understanding atrial fibrillation 3. Understanding the risk of falls 4. Identify the advantages and disadvantages of anticoagulant therapy 5. Judge this information according to criteria to be established (benefit/risk assessment)	**Non-expert**: avc non séquellaire → fibrillation auriculaire → fibrillation auriculaire traitemebnt → anti coagulant avc → anticoagulant avc → AVC ischémique aigu → anticoagulant avc → scholar google → Anticoagulants dans l'accident vasculaire cérébral (AVC) ischémique aigu → scholar google → anticoagulant avc → avk → anticoagulant avc	The stroke appears to be due to a blood clot created by the atrial fibrillation. An anticoagulant treatment seems to be prescribed to avoid a recurrence. Depending on the type of anticoagulant, special monitoring will be required.
Problem-solving (TRPMed)	*A 47-year-old man presented to the emergency room with left hypochondrium pain that had been evolving for 24 hours and was not relieved by level 1 analgesics.* *His history included cutaneous lupus and polycythemia.* *The biological workup was normal.* *The abdomino-pelvic scanner found two splenic hypodensities.* *With the information collected on the Internet, propose your diagnosis and etiological hypotheses.* 1. Understanding the left hypochondrium 2. Understanding of level 1 analgesics 3. Understanding cutaneous lupus and polycythemia 4. Understanding splenic hypodensity 5. Analyze the information previously collected and propose a diagnosis with etiological hypotheses	**Expert**: hypochondre gauche hypodensité splénique → hypodensité splénique → hypodensité splénique polyglobulie → maladie de vaquez → maladie de vaquez hypodensité → maladie de vaquez "hypodensité" → lupus cuta → lupus cutané polyglobulie → hypodensité splénique cause → hypodensité splénique kupus → hypodensité splénique lupus → syndrome de felty → scanner hypodensité → scanner hypodensité splénique → infractus s → infarctus splénique lupus → infarctus splénique → scanner hypodensité splénique → splénomégalie → splénomégalie scnanner → splénomégalie lupus → evolution ;lupus cutannée en systématique → evolution lupus cutannée en systématique	It could be splenomegaly which can lead to splenic infarction (by overload). The spleen is located in the left hypochondrium and the polycythemia may be due to a myeloproliferative syndrome. Splenic hypodensity suggests splenomegaly. Splenomegaly may be a (rare) sign of systematic lupus manifestation (digestive manifestation).

Table 4: Summary of the *CoST* resources

Filename.	Description	Data in each record
CoSTQueries.tsv	All the task-based retrieval data	QueryId, IdS, Query, Task, QueryActivity, QueryTime
CoSTSessionAnn.tsv	The session data and annotations about user's expertise, user's answer	IdS, Exp, Task, Answer
CoSTClicks.tsv	The click data of each query	QueryId, URL, URLTime
CoSTTasks.tsv	The task data and human-assessed sub-tasks	Task, Task description, expert expected sub-tasks

a set of candidate ones. Second, we also evaluate search strategy identification which consists in indicating if an issued query in context (based on previous queries in the session) mainly indicates either an exploration or an exploitation strategy. Note that both tasks are multi-class problems, but the numbers of labels vary from 15 for the former to 4 for the latter (Section 3.4). In the following sections we describe the query representation models and baselines used, followed by an analysis of the results obtained in each task.

4.1 Query Representation, Metrics, and Baselines

For query representation, we benefit from the power of recent contextualized embeddings, namely transformers [9]. In particular, we use three BERT-based architectures adapted for French: 1) the original BERT multilingual [9] that covers more than 100 languages and developed jointly with the BERT model; 2) LaBSE [11], a sentence enriched language-agnostic model, and 3) a language specific model for French, namely CamemBERT [27]. To the sake of simplicity, we use base models with the standard configuration for each architecture and opt for an all-frozen layers configuration. At the top of the architecture, we attach the two classification algorithms that use the CLS token representation.

- *K-Nearest Neighbors (k-NN)*: This is a classical classification algorithm that relies on the neighbors distribution to assign labels. In this case, we explore multiple neighbors configurations ($k = 1, ..., 50$) and present results of the best parameters.
- *AdaBoost*: We opt for an ensemble algorithm based on three classifiers. In this case we focus on the optimization of the number of estimators ($n = 100, ..., 1100$) used for the ensemble strategy. Results are reported for the best configuration.

For both algorithms, we present Accuracy, F-measure, Precision and Recall measures over a five cross-validation experiment with a grid search strategy for parameter optimization and a fixed random seed[4]. Accuracy is used to select the best model but all metrics are presented for each selected model. No special pre-processing was performed on the text queries.

4.2 Query-Task Mapping

The query task mapping problem has attracted recent attention as it may help understanding the user search intent thus, enhancing for instance document ranking and query suggestion models [26, 42]. In our context, the labels are known *a priori* as interactions were recorded under controlled search tasks and are provided in *CoST*. To facilitate future comparisons, the multiple baselines for *CoST* are presented in Table 5. Note that surprisingly the k-NN algorithm outperforms the strong AdaBoost classifier regardless of the contextualized representations. This could be explained by the fact that as tasks are similar across-users, queries likely overlap or are likely similar helping the k-NN algorithm to assign the correct label while AdaBoost is not able to correctly identify the pattern. Moreover, it is also surprising that the multilingual model and the language-agnostic model outperform a language specific model, but this difference is small when comparing multilingual BERT and CamemBERT models.

4.3 Search Strategy Identification

Here, our objective is to automatically identify the behavioral search strategy adopted by the user based on the observation of an issued

[4] set to 42.

Table 5: Query-task mapping performances using three different BERT-based architectures. Values correspond to average and standard deviation over a five cross-validation setup with parameters optimization of the best model. Power value corresponds to the standard deviation.

		Accuracy	F1	Precision	Recall
k-NN	CamemBERT	$0,89^{0,01}$	$0,89^{0,01}$	$0,90^{0,01}$	$0,89^{0,01}$
	LaBSE	$0,95^{0,01}$	$0,95^{0,01}$	$0,96^{0,01}$	$0,95^{0,01}$
	BERT-multi	$0,90^{0,01}$	$0,90^{0,01}$	$0,91^{0,01}$	$0,90^{0,01}$
Adaboost	CamemBERT	$0,31^{0,03}$	$0,30^{0,03}$	$0,35^{0,03}$	$0,31^{0,03}$
	LaBSE	$0,47^{0,03}$	$0,44^{0,04}$	$0,48^{0,05}$	$0,47^{0,03}$
	BERT-multi	$0,30^{0,02}$	$0,28^{0,03}$	$0,31^{0,03}$	$0,30^{0,02}$

Table 6: Search strategy identification performances using three different BERT-based architectures. Queries are encoded in a Single and Session-aware fashion. Values correspond to average and standard deviation over a five cross-validation setup with parameters optimization of the best model. Power value corresponds to the standard deviation.

			Accuracy	F1	Precision	Recall
k-NN	CamemBERT	Single query	$0,58^{0,02}$	$0,50^{0,02}$	$0,53^{0,04}$	$0,58^{0,02}$
		Session-aware	$0,57^{0,01}$	$0,51^{0,02}$	$0,50^{0,03}$	$0,57^{0,02}$
	LaBSE	Single query	$0,57^{0,02}$	$0,51^{0,02}$	$0,52^{0,04}$	$0,57^{0,02}$
		Session-aware	$0,58^{0,01}$	$0,53^{0,01}$	$0,53^{0,02}$	$0,58^{0,01}$
	BERT-multi	Single query	$0,57^{0,01}$	$0,51^{0,01}$	$0,52^{0,02}$	$0,57^{0,01}$
		Session-aware	$0,59^{0,01}$	$0,54^{0,01}$	$0,54^{0,02}$	$0,59^{0,01}$
Adaboost	CamemBERT	Single query	$0,54^{0,02}$	$0,50^{0,01}$	$0,49^{0,02}$	$0,54^{0,02}$
		Session-aware	$0,55^{0,01}$	$0,52^{0,01}$	$0,50^{0,01}$	$0,55^{0,01}$
	LaBSE	Single query	$0,57^{0,01}$	$0,54^{0,02}$	$0,53^{0,02}$	$0,56^{0,01}$
		Session-aware	$0,58^{0,01}$	$0,56^{0,01}$	$0,55^{0,01}$	$0,58^{0,01}$
	BERT-multi	Single query	$0,56^{0,01}$	$0,53^{0,01}$	$0,52^{0,01}$	$0,56^{0,01}$
		Session-aware	$0,58^{0,01}$	$0,56^{0,01}$	$0,55^{0,01}$	$0,58^{0,01}$

query. Similarly to query-task mapping, we address search strategy identification as a classification problem. Unlikely, the labels here are obtained in the post-experiment stage (Section 3.4.1): "0" (i.e., spelling correction), "1" (i.e., exploration), "2" (i.e., exploitation), "3" (i.e., narrow exploitation). Moreover, it is likely that query context (e.g., previous query, session) may be useful for identifying the user's search strategy. Thus, we use the same classification algorithms as those used for query-task mapping but extend the query representation with a session-aware representation. The latter consists in concatenating previous queries to the current one and has been shown to be effective on session retrieval [40]. Results are presented in Table 6. Note that, as can be expected, most of the session-aware representations outperform their counterparts and achieve best performances for all metrics. Differently to the query-task mapping results, the Adaboost algorithm outperforms k-NN in terms of F1 and Precision. Overall, this problem seems to be more challenging than the query-task mapping problem as accuracy ranges are 0.30 absolute points lower. However, as suggested by our results, feature engineering at session-level may help on this problem[5].

[5]This exploration is left as future work

5 ANALYZING THE EFFECTS OF COMPLEXITY AND DOMAIN KNOWLEDGE OF THE TASKS ON USER'S SEARCH BEHAVIOR

In this section, we show the usage of the CoST collection in understanding the user's search behavior. More precisely, our objective here is to examine the effects of task complexity and domain knowledge of the task on the users' behavior based on seven quantitative behavioral features, among which: 1) five ones directly observed from the browsing behavior: total number of clicks on SERPs (ClickSerp), total number of SERPs viewed that did not lead to a click (NoClickSerp), total time spent on SERPs (TimeSerp), total time spent on web pages (TimeURL) and total time to complete a search session (TimeSession) ; 2) two additional features related to search strategies inferred from the query annotations (Section 3.4): total number of exploration queries (Exploration) and total number of exploitation queries (Exploitation). Note that exploitation queries include queries with code "2" and those with code "3" (Section 3.4). In this experiment, we perform repeated measures of ANOVA on the seven dependent variables cited above. We select two independent variables: 1) Task complexity as within-subject factor (fact-finding, exploratory learning, decision-making, problem-solving, multicriteria-inferential); 2) Task domain knowledge as within-subject factor (computer science, medicine, psychology). In the case where the ANOVA test is significant, we perform Scheffe's post-hocs. To this extent, all comparisons presented in the results analysis are significant. Table 7 presents a summary of the ANOVA results and Table 8 details means and standard deviations of the behavioral feature values. We discuss below the results obtained and the primary findings that emerged from them.

5.1 Analyzing the Effects of Complexity

Overall, we can clearly see from Table 7, that all the behavioral features are significant for explaining user's search behavior. Let us for instance have a first close look to NoClickSerp and TimeURL features regarding the browsing behavior. As can be seen from Table 8, looking particularly at NoClickSerp feature, the fact-finding task ($M = 1.94$, $SD = 2$) and the exploratory learning task ($M = 2.2$, $SD = 3.72$) lead to significantly fewer unlinked SERPs than the other types of task complexity. Indeed, for the fact-finding task, the answer is directly accessible on SERPs, with the first SERP therefore relevant to finding the answer [10, 29, 35]. Therefore, we do not observe an increase in the number of unlinked SERPs for this low-level task complexity. For the exploratory learning task, the goal is to collect information to gain new knowledge [12] and is therefore similar to a multiple fact search. Several paths can be pursued and depend directly on the sub-goals that the user sets for himself. It is therefore not difficult to access relevant documents as long as they allow the user to learn more about the subject of the task. Another trend that emerges from the results is the fact that the multicriteria-inferential task ($M = 12.93$, $SD = 11.73$) leads to the most unclicked SERPs compared to all the other types of task complexity. This is in line with previous findings indicating that if users fail to infer new additional keywords than those provided in the statement which are likely to be ambiguous, they cannot access

Table 7: Summary of ANOVA results for Task Complexity (TC), Task Domain (TD), and Task Complexity*Task Domain (TC*TD).

		ClickSerp			NoClickSerp			TimeSerp			TimeURL			TimeSession			Exploration			Exploitation		
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**	***F***	***p***	**N^2p**
TC	F(4,276)	50.2	<.001	0.421	94.2	<.001	0.577	46.34	<.001	0.402	74.06	<.001	0.518	76.93	<.001	0.527	60.31	<.001	0.466	75.49	<.001	0.522
TD	F(2,138)	10	<.001	0.127	4	<.05	0.055	3.5	<.05	0.048	0.2	n.s		0.98	n.s		4.51	<.05	0.061	9.43	<.001	0.12
TC*TD	F(8,552)	32.4	<.001	0.32	49.34	<.001	0.417	14.9	<.001	0.178	1.4	n.s		6.32	<.001	0.084	38.56	<.001	0.358	38.8	<.001	0.36

Table 8: Means and (standard deviations) of Analysis for the Psychology Task (PT), Computer Science Task (CST), Medicine Task (MT), and Total Task (TT) complexity.

	ClickSerp				NoClickSerp				TimeSerp				TimeURL				TimeSession				Exploration				Exploitation			
	PT	**CST**	**MT**	***TT***	**PT**	**CST**	**MT**	**TT**	**PT**	**CST**	**MT**	**TT**	**PT**	**CST**	**MT**	**TT**	**PT**	**CST**	**MT**	**TT**	**PT**	**CST**	**MT**	**TT**	**PT**	**CST**	**MT**	**TT**
Fact-finding	1.4 (1.6)	1.64 (1.6)	2.3 (1.84)	1.72 (1.7)	1.93 (1.9)	1.9 (2.1)	1.9 (2)	1.94 (2)	60 (79.3)	49.42 (69.3)	52.33 (62.71)	57.33 (70.6)	54.33 (80.1)	60.23 (85.1)	113.4 (103.4)	72.33 (93.5)	114.3 (132.41)	109.7 (116.1)	165.7 (114.9)	129.7 (123.5)	1.41 (1.04)	1.5 (1.1)	1.2 (.61)	1.4 (0.94)	.66 (1.12)	.5 (.8)	.6 (.8)	.6 (.91)
Exploratory learning	4.7 (4.2)	2.3 (1.6)	1.84 (1.2)	3.2 (2.93)	3.8 (5.4)	1.73 (2.62)	.84 (1.21)	2.2 (3.72)	168.9 (273.2)	68.21 (68.12)	42.52 (66.23)	97.5 (174.94)	440.5 (380.72)	436.13 (341.62)	449.82 (344.52)	474.7 (354.41)	609.4 (477.02)	504.45 (347)	492.4 (355.73)	572.24 (399.3)	2.33 (3.24)	1.5 (1.2)	1.3 (.63)	1.74 (2.1)	1 (1.5)	.24 (.91)	.2 (.5)	.5 (1.1)
Decision-making	2.3 (1.71)	5.43 (4.1)	3.9 (3.4)	3.9 (3.43)	1.8 (3.24)	8.7 (7.23)	5.31 (4.33)	5.24 (5.91)	83.6 (155.7)	209.13 (204.3)	198.7 (301.2)	156.8 (234.5)	623.4 (446.2)	664.8 (710.42)	579 (493.2)	649.9 (560.3)	706.9 (488.2)	873.9 (790.6)	777.6 (580.14)	806.5 (633.12)	1.6 (1.2)	4.23 (3.24)	3 (2.3)	2.83 (2.6)	.6 (1.6)	2.9 (3.01)	1.9 (2.23)	1.7 (2.6)
Problem solving	4.6 (4.13)	3.7 (3.8)	10.14 (8.5)	6.1 (6.5)	6.3 (6.8)	6.04 (5.6)	16.41 (13.64)	9.6 (10.54)	227.4 (250.71)	206.9 (268.74)	343.6 (256.63)	253.6 (264.5)	593.24 (550)	543.23 (421.7)	659.7 (637.8)	601.33 (543.24)	820.7 (597.8)	749.1 (511.1)	983.32 (819.3)	838.62 (659.9)	2.8 (2)	2.7 (2.23)	7.43 (5.9)	4.2 (4.4)	2.34 (3.6)	2.11 (2.5)	5.11 (5.4)	3.1 (4.21)
Multicriteria-inferential	8.6 (6.71)	4.6 (3.3)	4.81 (3.8)	5.81 (5.14)	21.14 (13.64)	10 (7.5)	8.04 (6.03)	12.93 (11.73)	377.8 (294.7)	208.2 (165.2)	148.11 (92.2)	237.3 (223.6)	302.34 (276.14)	239.6 (265.52)	231.43 (185.3)	263.8 (246.5)	680.1 (490.01)	447.82 (379)	379.6 (237.91)	501.1 (402.5)	7.1 (5.12)	3.4 (2.4)	3.44 (2.2)	4.6 (3.9)	9.2 (7.3)	3.4 (3.4)	2.8 (3.03)	5 (5.7)
Total Task domain	4.3 (4.8)	3.7 (4.8)	4.5 (4.8)	4.14 (4.61)	6.83 (9.1)	5.9 (9.3)	6.41 (9.3)	6.4 (8.9)	185.9 (230.4)	145.1 (232.21)	150.5 (230.3)	163 (220.21)	418.2 (471.22)	418.33 (467.74)	400.7 (461.5)	399.4 (246.5)	604 (570.5)	563 (564.7)	541.91 (553.7)	561 (546.6)	3.1 (3.4)	2.7 (3.5)	3.1 (3.5)	3 (3.3)	2.7 (3.92)	1.8 (4.01)	2.02 (4.1)	2.23 (3.9)

to relevant links [35]. Finally, there is also a significant difference between the decision-making task ($M = 5.24$, $SD = 5.91$) and the problem-solving task ($M = 9.6$, $SD = 10.5$), which both lead to more unclicked SERPs. As argued in previous work [18], the problem-solving task implies the need to investigate several search paths in order to discard those that are not relevant to the goal. To this extent, some irrelevant paths are directly discarded during the browsing activity, which explains why the problem-solving task leads to more SERPs offering no relevant links to the user. On the contrary, for the decision-making task, several possibilities of relevant paths can be taken [18] which could lead to fewer failed SERPs than the problem-solving task.

When looking at the TimeURL feature, the significance of the ANOVA results (see Table 7) coupled with the statistics in Table 8, highlight several interesting trends. First, the fact-finding task ($M = 72.33$, $SD = 93.5$) requires the least amount of time spent on URLs compared to all the other types of task complexity. This could be explained by the fact that in such a task, the user does not need to access documents to locate the target information [34, 35]. The completion of this task does not induce an in-depth processing of web page content. The other interesting observation that emerges is that the exploratory learning task ($M = 474.7$, $SD = 354.41$) results in less time spent on URLs than the decision-making task ($M = 649.9$, $SD = 560.3$) and the problem-solving task ($M = 601.33$, $SD = 543.24$). This is because the multiple information gathering objective [12] does not require the direct application of the newly acquired knowledge as it is the case for the decision-making task and the problem-solving task. The latter requires more time spent on the URLs (i.e., to process in depth the content of the web pages). This processing activity underlies the mobilization of a significant amount of cognitive resources [34], since the completion of those tasks, require from users to investigate beyond simple fact-finding [5]. For the decision-making task, the objective is to analyze web pages, to locate comparison criteria and arguments, to conduct evaluation work to make the best decision among several possibilities [18]. Regarding the problem-solving task, the challenge is to understand the new incoming information in order to be able to apply it and produce a new coherent set of information [18]. Finally, the last result worth highlighting regarding the TimeURL indicator is the one concerning the multicriteria-inferential task ($M = 263.8$, $SD = 246.5$) which requires less time spent on URLs than the problem solving, decision-making and exploratory learning task. According to [34], ill-defined goals, as those targeted in multicriteria-inferential tasks, likely lead users to search failure and thus, they spend little time processing URLs thoroughly.

Now, turning our search attention to strategy-based features (Exploration, Exploitation), we note that overall, the fact-finding task and exploratory learning task require the least number of exploration and exploitation queries compared to the other tasks. As previously shown, the fact-finding task does not require initiating new search leads since the first lead provides direct access to the desired result [35]. The other interesting result is that the multicriteria-inferential task ($M = 5, SD = 5.7$) requires, at a highest extent compared to the other tasks, the exploitation strategy. This result is consistent with previous work indicating that ill-defined tasks lead to abusive exploitation behaviors in processing similar information despite repeated failures [34]. At the query level, the literature indicates that narrow semantic changes in two successive queries (i.e., exploitation) repeated multiple times, are a clear signal that users struggle to find relevant information while solving the task [4, 30].

5.2 Analyzing the Effects of Task Domain and its Interaction with Task Complexity

Here, we focus on the effect of task domain and the joint effect of task complexity and task domain on browsing behavior and search strategies. From Table 7 and Table 8, we can observe several significant effects. Regarding the browsing behavior, let us focus for instance on ClickSerp and NoClickSerp features. From the task domain perspective, we can see that the computer science domain significantly impacts ClickSerps feature. Tasks in the computer science domain lead to significantly fewer clicks from SERPs ($M = 3.7$, $SD = 4.8$) than those in the psychology domain ($M = 4.3$, $SD = 4.8$)

and medicine domain ($M = 4.5$, $SD = 4.8$). Tasks in the psychology domain ($M = 6.83$, $SD = 9.1$) induce significantly more SERPs that do not lead to a click compared to tasks from the computer domain ($M = 5.9$, $SD = 9.3$).

Regarding the search strategies (i.e., exploration vs. exploitation), we can observe that tasks in the computer science domain ($M = 2.7$, $SD = 3.5$) lead to significantly less exploration in comparison to tasks in the medicine domain ($M = 3.1$, $SD = 3.5$). Finally, tasks in the psychology domain ($M = 2.7$, $SD = 3.92$) induce significantly more exploitation behavior than in comparison to tasks in the computer science domain ($M = 1.8$, $SD = 4.01$) and medicine domain ($M = 2.02$, $SD = 4.1$).

When we examine the interaction between task complexity and task domain, we can observe from Table 7 and Table 8 clear patterns. For the decision-making task, the computer science domain leads to significantly more clicks from SERPs ($p < .001$), more SERPs visited that do not lead to a click ($p < .001$), and more exploration ($p < .001$) than the decision-making tasks in the psychology domain. For the problem-solving task, the medicine-related tasks lead to significantly more clicks from SERPs ($p < .001$), more SERPs visited that do not lead to a click ($p < .001$), and more exploration ($p < .001$) than those in the psychology and computer science problem-solving tasks. Also, the problem-solving task in medicine leads to significantly more exploitation compared to their counterparts in the psychology ($p = .002$) and computer science ($p < .001$) domains. Regarding the multi-criteria inferential task, the tasks in the psychology domain stand out from the others. Indeed, it leads to significantly more clicks from SERPs ($p < .001$), more unclicked SERPs ($p < .001$) and more exploration ($p < .001$) and exploitation ($p < .001$) than its counterparts in the medicine and computer science domains.

In the light of all the above results, we can globally postulate that in addition to the objective complexity of tasks [2, 12, 18], the knowledge domain in which these tasks fall also has a significant and joint effect on user's browsing and search behavior. This study has been made possible thanks to *CoST* which has been designed in such a way that each participant has expertise in one among the 3 domains and has completed the same 15 search tasks corresponding to one task per complexity (×5) and per domain (×3). Thus, *CoST* setting fills in the critical gap identified by Choi et al. [8] in the majority of previous users studies which only design a single task per complexity.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we presented the *CoST* collection which provides the research community with a richly grounded resource that can be used for the evaluation of diverse task-based search settings, from both the information access system and the user perspectives. Specifically, the *CoST* collection is collaboratively designed by researchers from the computer science and cognitive psychology domains. The collection releases critical attributes of tasks and their doers including mainly the cognitive complexity of tasks and domain expertise of participants, thus enabling researchers to investigate a number of search and interaction models by relating these attributes. We showcased the usage of *CoST* for query mapping and search strategy classification tasks as well as for studying user search behavior w.r.t. tasks of varying cognitive complexity. In the future, we plan to use the *CoST* collection in more in depth comparative analysis with additional applications, baselines and benchmarks.

ACKNOWLEDGMENTS

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.

REFERENCES

[1] Nicholas Belkin, Toine Bogers, Jaap Kamps, Diane Kelly, Marijn Koolen, and Emine Yilmaz. 2017. Second Workshop on Supporting Complex Search Tasks *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 433–435. https://doi.org/10.1145/3020165.3022163

[2] David J. Bell and Ian Ruthven. 2004. Searcher's Assessments of Task Complexity for Web Searching. *Lecture Notes in Computer Science* (2004), 57–71. https://doi.org/10.1007/978-3-540-24752-4_5

[3] Toine Bogers, Samuel Dodson, Luanne Sinnamon, Maria Gäde, Mark Hall, Marijn Koolen, Vivien Petras, Nils Pharo, and Mette Skov. 2019. Workshop on Barriers to Interactive IR Resources Re-use (BIIRRR 2019). *CHIIR '19: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 389–392. https://doi.org/10.1145/3295750.3298965

[4] Pia Borlund and Sabine Dreier. 2014. An investigation of the search behavior associated with Ingwersen's three types of information needs. *Information Processing & Management* 50, 4 (2014), 493–507. https://doi.org/10.1016/j.ipm.2014.03.001

[5] Katriina Byström and Preben Hansen. 2005. Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology* 56, 10 (2005), 1050–1061. https://doi.org/10.1002/asi.20197

[6] Ben Carterette, Paul Clough, Mark Hall, Evangelos Kanoulas, and Mark Sanderson. 2016. Evaluating Retrieval over Sessions: The TREC Session Track 2011-2014 *(SIGIR '16)*. 685–688.

[7] Marc-Allen Cartright, Ryen W. White, and Eric Horvitz. 2011. Intentions and Attention in Exploratory Health Search *(SIGIR '11)*. 65–74.

[8] Bogeum Choi, Ward Austin, Li Yuan, Arguello Jaime, and Capra Robert. 2019. The Effects of Task Complexity on the Use of Different Types of Information in a Search Assistance Tool. *ACM Transactions on Information Systems* 38, 1 (2019), 1–28. https://doi.org/10.1145/3371707

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[10] Aurélie Dommes, Aline Chevalier, and Sarah Lia. 2011. The role of cognitive flexibility and vocabulary abilities of younger and older users in searching for information on the web. *Applied Cognitive Psychology* 25, 5 (2011), 717–726. https://doi.org/10.1002/acp.1743

[11] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT Sentence Embedding. arXiv:2007.01852 [cs.CL]

[12] Marchionini Gary. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46. https://doi.org/10.1145/1121949.1121979

[13] Souvick Ghosh, Manasa Rath, and Chirag Shah. 2018. Searching as Learning: Exploring Search Behavior and Learning Outcomes in Learning-Related Tasks. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Cumputing Machinery, New York, NY, USA, 22–31. https://doi.org/10.1145/3176349.3176386

[14] Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. 2013. From Search Session Detection to Search Mission Detection. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval* (Lisbon, Portugal) *(OAIR '13)*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 85–92.

[15] Ahmed Hassan Awadallah, Cathal Gurrin, Mark Sanderson, and Ryen W. White. 2019. Task Intelligence Workshop @ WSDM 2019. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. 848–849.

[16] Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. Supporting Complex Search Tasks *(CIKM '14)*. 829–838.

[17] Yunlong He, Jiliang Tang, Hua Ouyang, Changsung Kang, Dawei Yin, and Yi Chang. 2016. Learning to Rewrite Queries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis,

Indiana, USA) *(CIKM '16)*. 1443–1452.

[18] Campbell Donald J. 1988. Task Complexity: A Review and Analysis. *Academy of Management Review* 13, 1 (1988), 40–52. https://doi.org/10.5465/amr.1988.430677

[19] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2009. Patterns of Query Reformulation During Web Searching. *J. Am. Soc. Inf. Sci. Technol.* 60, 7 (July 2009), 1358–1371.

[20] Jiepu Jiang, Daqing He, Diane Kelly, and James Allan. 2017. Understanding Ephemeral State of Relevance. In *Proceedings of the 2017 Conference on Human Information Interaction & Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Cumputing Machinery, New York, NY, USA, 137–146. https://doi.org/10.1145/3020165.3020176

[21] Evangelos Kanoulas, Leif Azzopardi, and Grace Hui Yang. 2018. Overview of the CLEF Dynamic Search Evaluation Lab 2018. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 362–371.

[22] Evangelos Kanoulas, Emine Yilmaz, Rishabh Mehrotra, Ben Carterette, Nick Craswell, and Peter Bailey. 2017. TREC 2017 Tasks Track Overview. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017 (NIST Special Publication, Vol. 500-324)*. National Institute of Standards and Technology (NIST).

[23] Jiqun Liu, Shawon Sarkar, and Chirag Shah. 2020. Identifying and predicting the states of complex search tasks. In *Proceedings of the 2020 Conference on Human Information Interaction & Retrieval* (Vancouver, BC, Canada) *(CHIIR '20)*. Association for Cumputing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/3343413.3377976

[24] Kun Lu, Soohyung Joo, Taehun Lee, and Rong Hu. 2017. Factors That Influence Query Reformulations and Search Performance in Health Information Retrieval: A Multilevel Modeling Approach. *J. Assoc. Inf. Sci. Technol.* 68, 8 (Aug. 2017), 1886–1898.

[25] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. 2011. Identifying Task-Based Sessions in Search Engine Query Logs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China) *(WSDM '11)*. 277–286.

[26] Luis Lugo, Jose G. Moreno, and Gilles Hubert. 2021. Extracting Search Tasks from Query Logs Using a Recurrent Deep Clustering Architecture. In *Advances in Information Retrieval*. 391–404.

[27] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[28] R Mehrotra, AH Awadallah, and E. Yilmaz. 2018. Report on the WSDM 2018 Workshop on Learning from User Interactions. *SIGIR Forum* 52, 1 (Aug. 2018), 797–798.

[29] Sophie Monchaux, Franck Amadieu, Aline Chevalier, and Claudette Mariné. 2015. Query strategies during information searching: Effects of prior domain knowledge and complexity of the information problems to be solved. *Information Processing & Management* 51, 5 (2015), 557–569. https://doi.org/10.1016/j.ipm.2015.05.004

[30] Daan Odijk, Ryen W. White, Ahmed Hassan Awadallah, and Susan T. Dumais. 2015. Struggling and Success in Web Search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1551–1560. https://doi.org/10.1145/2806416.2806488

[31] Paul Over. 2001. The TREC interactive track: an annotated bibliography. *Information Processing & Management* 37, 3 (2001), 369–381. https://doi.org/10.1016/S0306-4573(00)00053-4 Interactivity at the Text Retrieval Conference (TREC).

[32] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A Picture of Search *(InfoScale '06)*. Association for Computing Machinery, New York, NY, USA, 1–es.

[33] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.

[34] Mylene Sanchiz, Franck Amadieu, and Aline Chevalier. 2020. An Evolving Perspective to Capture Individual Differences Related to Fluid and Crystallized Abilities in Information Searching with a Search Engine. In *Understanding and Improving Information Search: A Cognitive Approach* (1 ed.), Wai Tat Fu and Herre van Oostendorp (Eds.). Springer, Cham, Switzerland, Chapter 5, 71–96.

[35] Mylene Sanchiz, Aline Chevalier, and Franck Amadieu. 2017. 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. https://doi.org/10.1016/j.chb.2017.02.038

[36] Mylène Sanchiz, Jessie Chin, Aline Chevalier, Wai-Tat Fu, Franck Amadieu, and J. He. 2017. Searching for information on the web: Impact of cognitive aging, prior domain knowledge and complexity of the search problems. *Information Processing & Management* 53, 1 (2017), 281–294. https://doi.org/10.1016/j.ipm.2016.09.003

[37] Procheta Sen, Debasis Ganguly, and Gareth Jones. 2018. 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)*. 283–292.

[38] Lynda Tamine and Cecile Chouquet. 2017. On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. *Information Processing & Management* 53, 2 (2017), 332–350. https://doi.org/10.1016/j.ipm.2016.11.004

[39] Pertti Vakkari. 1999. Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information Processing & Management* 35, 6 (1999), 819–837.

[40] Christophe Van Gysel, Evangelos Kanoulas, and Maarten de Rijke. 2016. Lexical Query Modeling in Session Search. In *ICTIR*, Vol. 2016. ACM.

[41] Michael Völske, Ehsan Fatehifar, Benno Stein, and Matthias Hagen. 2019. Query-Task Mapping. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, 969–972.

[42] Michael Völske, Ehsan Fatehifar, Benno Stein, and Matthias Hagen. 2019. Query-Task Mapping. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 969–972. https://doi.org/10.1145/3331184.3331286

[43] Barbara M. Wildemuth, Diane Kelly, Emma Boettcher, Erin Moore, and Gergana Dimitrova. 2018. Examining the Impact of Domain and Cognitive Complexity on Query Formulation and Reformulation. *Information Processing & Management* 54, 3 (May 2018), 433–450.

[44] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. 2012. 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* (Nijmegen, The Netherlands) *(IIiX '12)*. Association for Cumputing Machinery, New York, NY, USA, 254–257. https://doi.org/10.1145/2362724.2362768

[45] Grace Hui Yang and Ian Soboroff. 2016. TREC 2016 Dynamic Domain Track Overview. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016 (NIST Special Publication, Vol. 500-321)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST), 22 pages.