International Journal of Digital Libraries manuscript No.
(will be inserted by the editor)

Task-oriented Search for Evidence-based Medicine

Bevan Koopman · Jack Russell · Guido Zuccon

Received: date / Accepted: date

Abstract

Purpose: Research on how clinicians search shows that they pose queries according to three common clinical tasks: searching for diagnoses, searching for treatments and searching for tests. We hypothesise, therefore, that structuring an information retrieval system around these three tasks would be beneficial when searching for evidence-based medicine (EBM) resources in medical digital libraries.

Methods: Task-oriented (diagnosis, test and treatment) information was extracted from free-text medical articles using a natural language processing pipeline. This information was integrated into a retrieval and visualisation system for EBM search that allowed searchers to interact with the system via task-oriented filters. The effectiveness of the system was empirically evaluated using TREC CDS — a gold standard of medical articles and queries designed for EBM search.

Results: Task-oriented information was successfully extracted from 733,138 articles taken from a medical digital library. Task-oriented search led to improvements in the quality of search results and savings in searcher workload. An analysis of how different tasks affected retrieval showed that searching for treatments was the most challenging and that the task-oriented approach improved search for treatments. The most savings in terms of workload was observed when searching for treatments and tests.

Conclusions: Overall, taking into account different clinical tasks can improve search according to these tasks. Each task displayed different results, making systems that are more adaptive to the clinical task type desirable. A future user study would help quantify the actual cost saving estimates.

Keywords Information retrieval · Evidence-based medicine · Task-oriented search · Clinical decision support

B. Koopman & J. Russell
Australian e-Health Research Centre, CSIRO
Royal Brisbane Hospital, Brisbane, Queensland, Australia
Tel.: +61-7-32533635
E-mail: bevan.koopman@csiro.au

G. Zuccon
School of Electrical Engineering and Computer Science, Queensland University of Technology
E-mail: g.zuccon@qut.edu.au

1 Introduction

Evidence-based medicine (EBM) is the practice of making clinical decisions based on rigorous scientific evidence. EBM relies on effective access to peer-reviewed literature found in medical digital libraries [8]. However, this is hampered by both the exponential growth of medical literature in digital libraries [18] (Figure 1) and a lack of efficient and effective means of searching and visualising this literature.

While there are mature resources for searching medical literature (the PubMed digital library being a widely used example), these are primarily focused on retrieving literature for research purposes, not for clinical decision support. This paper investigates information retrieval (IR) and natural language processing methods for accessing EBM resources in digital libraries specifically for clinical decision support. These methods are based on research on how clinicians (doctors, nurses or other health professionals) search in a clinical decision support setting [9]. Specifically, that clinicians pose

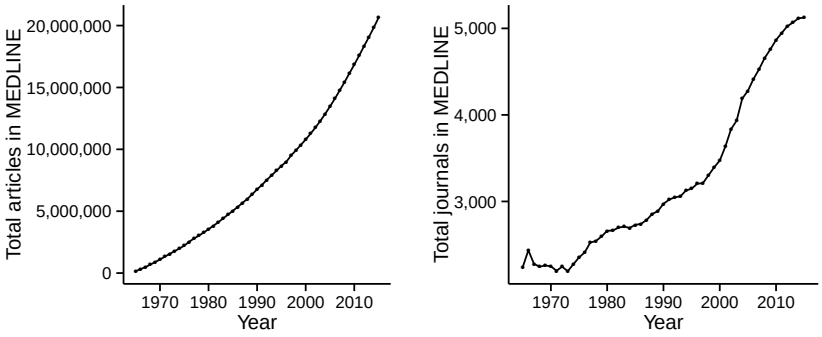

Fig. 1 Growth in articles contained in MEDLINE — 1966–2015. Statistics taken from National Library of Medicine [22].

queries within three common clinical tasks: i) searching for *diagnoses* given a list of symptoms; ii) searching for relevant *tests* given a patient's situation; and iii) searching for the most effective *treatments* given a particular condition. These three tasks are at the core of both our proposed methods for retrieving EBM resources, how these resources are presented back to clinicians and our analysis of how these three tasks influenced retrieval effectiveness. Specifically, we investigate both the representation of medical articles and the retrieval and display of search results around these three clinical tasks — hence a task-oriented search approach. This paper investigates the following research questions:

1. How can task-oriented information be extracted from free-text medical articles?
2. How can a task-oriented approach be integrated into a retrieval and visualisation method for EBM search and, importantly, does this improve retrieval effectiveness?
3. How do the different task types affect retrieval effectiveness?

2 Related Work

2.1 Task-oriented clinical questions

Research on how clinicians search has indicated that clinical questions fall within a number of common categories; these have been organised into a standard taxonomy of clinical questions [9]. The three commonest question categories were: 1) "What is the treatment of choice for condition x?"; 2) "What is the cause of symptom x?"; and 3) "What test is indicated in situation x?". These three questions can be expressed as searching for treatments, searching for diagnoses and searching for tests, respectively. Most of the remaining question categories [9] could be expressed as specialisations of one of these three. The fact that clinicians express clinical queries according to these categories is motivation for structuring information retrieval (IR) systems around these three, both in terms of how the system retrieves articles and how the clinicians interacts with those results. This is the motivation and the methodology adopted by this study.

2.2 Interacting with search systems

Structuring information retrieval systems around different categories of informations (diagnoses, tests and treatments in our case) is a common approach in IR. The categories are generally referred to as facets and the approach as faceted retrieval [10]. Search results in faceted retrieval are presented to the user organised around the various facets. E-commerce sites such as eBay and Amazon are typical example where search results are organised around product facets. Searchers engaging in complex search tasks have been shown to benefit from the faceted approach [33]. The benefits come from organising the search results around facets, thus providing an easy overview of the results and via the ability to filter the search results according to results of interest. Faceted retrieval reduces mental workload by promoting recognition over recall and by suggesting logical yet unexpected navigation paths to the user [33]. Meaningfull facets have been found to support learning, reflection, discover and information finding [16,27,33]. EBM-based search can be viewed as a complex search task [13]: clinical have complex information needs and are often time pressured. Thus, an IR approach such as faceted retrieval, which reduces mental overhead, is desirable. In this paper, we test the hypothesis that faceted retrieval, which has shown benefits in general web search, can improve search for EBM.

The importance of access to biomedical literature has resulted in many retrieval systems specific for searching this type of content. Hersh [11] provides an extensive overview of many of these, including a section

specifically dedicated to search interfaces [11, ch5]. Many of these interfaces were concerned with searching MEDLINE — the same source of medical journal articles used in this study. It is worth noting that mention was made of different types of clinical queries: therapy, diagnosis, harm and prognosis. These have parallels with the diagnosis, test, treatment tasks identified by [9] and used in our study. Although query categories (which were akin to tasks) were identified, they were not explicitly integrated into the retrieval method and the way the searcher was presented with and interacts with the search results. Our study uses the clinical tasks as the bases for both retrieval and clinician interaction. Finally, most methods for searching EBM resources were for research purposes, rather than clinical decision support. As such, recall was an important factor (i.e., finding all the relevant articles for a particular information need). In contract, for clinical decision support, precision can be more important (i.e., finding the article that helps with the clinical task without reading many irrelevant articles). Our study bases the design of the IR system around improving precision.

2.3 Clinical retrieval methods

There have been a number of retrieval methods that attempt to exploit task-specific information to improve retrieval effectiveness.

One method to improve retrieval effectiveness in EBM is to extract as much structured information from free-text medical articles. EBM itself advocates a more structured approach with the four key elements making up a well-built clinical question [23] being: Population, Intervention, Comparison and Outcome (PICO). Based on this a number of retrieval approaches have been develop that attempt to explicitly model medical queries and articles according to the PICO categories [6, 7]. Based on PICO structured information, Demner-Fushman & and Lin [7] developed a clinical question answering system that re-ranked PubMed articles according to a criteria specific for EBM. Users of system had to enter a structured query according the four PICO categories. Considerable effort was devoted to the development of classification and extraction methods for PICO categories. For article ranking, different levels of evidence were considered whether the article comprised a rigorous random controlled trial or meta-analysis vs. limited quality patient oriented evidence [6]. The number of citations to articles was also used as an indication of relevance. PICO categories can be mapped to diagnosis, test, treatments tasks, which are the basis of this study. However, in this study we do not focus heavily on extracting PICO structured information from articles and do not require that the clinician enter their query according to the PICO structure. We take a more lightweight approach to extracting just three types of task (diagnosis, test and treatment). Furthermore, we focus more on how the clinician may interact with the retrieval system (via a user interface) and how that interaction affected the retrieval effectiveness.

Diagnosis, test and treatment information have been successfully integrated into a number of information retrieval models. One approach to do this is to map all clinical queries and clinical documents being searched to medical concepts according to an external domain knowledge resource (e.g., the UMLS medical thesaurus); matching is then done at the concept level, comparing a query concept with a document concept [30]. Improvements in retrieval effectiveness were obtained when this concept-based approach was restricted according to the clinical tasks symptom, test, diagnosis and treatment [21, 20]. This shows the benefits of focusing retrieval around these three clinical tasks. Although concept retrieval using tasks has proved effective, the tasks were simply used as features within the retrieval model and never exposed to the clinician [21,20]. In this study, we attempt to make the task-based information explicit in the way the clinician interacts with the system, as well as the basis for the underlying retrieval model. We also aim to analyse interactions with the system to better understand how different task types affect retrieval effectiveness?

To empirically evaluate the methods proposed in this paper we make use of an existing test collection for clinical search, namely the Text Retrieval Conference Clinical Decision Support (TREC CDS) challenge [25, 24]. TREC CDS was an international shared task aimed at evaluating information retrieval systems in searching PubMed articles in a clinical decision support setting. The goal of the task was, given a description of a patient, to retrieve relevant articles that help a clinician in diagnosing, testing and treating that patient. A number of teams developed systems for TREC CDS. Most teams developed retrieval methods designed to improve precision and recall by making use of features of the patent query or PubMed article [24]. As this was a TREC batch evaluation, there was no interactive systems or considerations of search engine interfaces. Instead, in this study, we consider the interaction the searcher would have with the search system and how a user interface would facilitate such interactions.

In summary, while other studies attempt to extract detailed, structured information from medical articles, we adopt a lightweight approach by considering only diagnoses, tests and treatments. These three tasks were

```
51-year-old smoker with hypertension and diabetes,
in menopause, needs recommendations for preventing
osteoporosis.
```

Fig. 2 Sample topic (topic# 2014-29) representing a single patient. The topic type for this topic was 'treatment', indicating that relevant articles about how best to treat this patient are sought.

treated in a facet-based approach, which has proved effective in improving search interactions in other domains. The tasks-oriented information is used not only as a feature in retrieval but also as a means improving and better understanding the way clinician might interact with the system.

3 Methods

3.1 Task and data

The TREC CDS test collection was used for empirical evaluation of the proposed methods. The test collection contained:

Documents A collection of 733,138 medical articles from PubMed. These were full-text, open access articles.

Queries/topics Sixty search queries (called topics in TREC CDS).[1] Each topic represented a patient case report that detailed the conditions and history of a particular patient. A sample of one such topic is provide in Figure 2. Furthermore, each topic was assigned one of three topic types — diagnosis, test or treatment. The topic type represented the particular clinical task required for the patient: i) searching for *diagnoses* given a list of symptoms; ii) searching for relevant *tests* given a patient's situation; and iii) searching for the most effective *treatments* given a particular condition.

Relevance assessments The TREC CDS organisers employed clinician assessors to review a selection of articles for each topic. The assessors were instructed to judge articles as either "definitely relevant", "not relevant", or "possibly relevant". These relevance assessments represent the gold standard against which different information retrieval systems can be evaluated.

Further details on the methodology and participation in TREC CDS were provided by the organisers [25, 24].

[1] Thirty topics from TREC 2014 and thirty topics from TREC 2015.

3.2 Overview of the task-oriented approach

We provide an overview of the task-oriented approach before detailing the individual components in the subsequent sections. Figure 3 shows the overall architecture. In the indexing phase, medical articles were fed to the task extraction process which automatically annotated mentions of diagnoses, tests and treatments from free-text. The resulting annotated articles were indexed into an information retrieval inverted index with separate fields for diagnoses, tests and treatments. In the retrieval phase, a clinician interacted with the system via a web-based user interface. The interface allowed for free-text querying, with results displayed in task-oriented focus, with respect to diagnoses, tests and treatments.

3.3 Task extraction from free-text

The task-oriented approach to EBM search required that the different task types — diagnoses, tests and treatments — were extracted from the free-text medical articles. To achieve this we developed a natural language processing pipeline involving a number of steps:

1. First, we applied an information extraction system that identified mentions of medical concepts from free-text. We used QuickUMLS [28], a system that maps free-text to UMLS medical concepts. The Unified Medical Language System (UMLS) is a compendium of many controlled vocabularies in the biomedical sciences. QuickUMLS was applied to the title and abstract of each article.[2]
2. Within UMLS, each medical concept has an overarching semantic type (e.g., the concept "Headache" belongs to the semantic type "Sign or Symptom"). We mapped each concept from Step 1 to its corresponding semantic type.
3. Each semantic type could then be mapped to the clinical tasks diagnosis, treatment or test by consulting the i2b2 challenge guidelines [32] which defined a mapping between UMLS semantic types and clinical tasks.
4. Once the task was identified the original span of text from the article was annotated with details of the task type. A sample text, with annotated spans, is shown in Figure 4.

[2] The full body was not included as it contained large amounts of HTML formatting that QuickUMLS could not interpret.

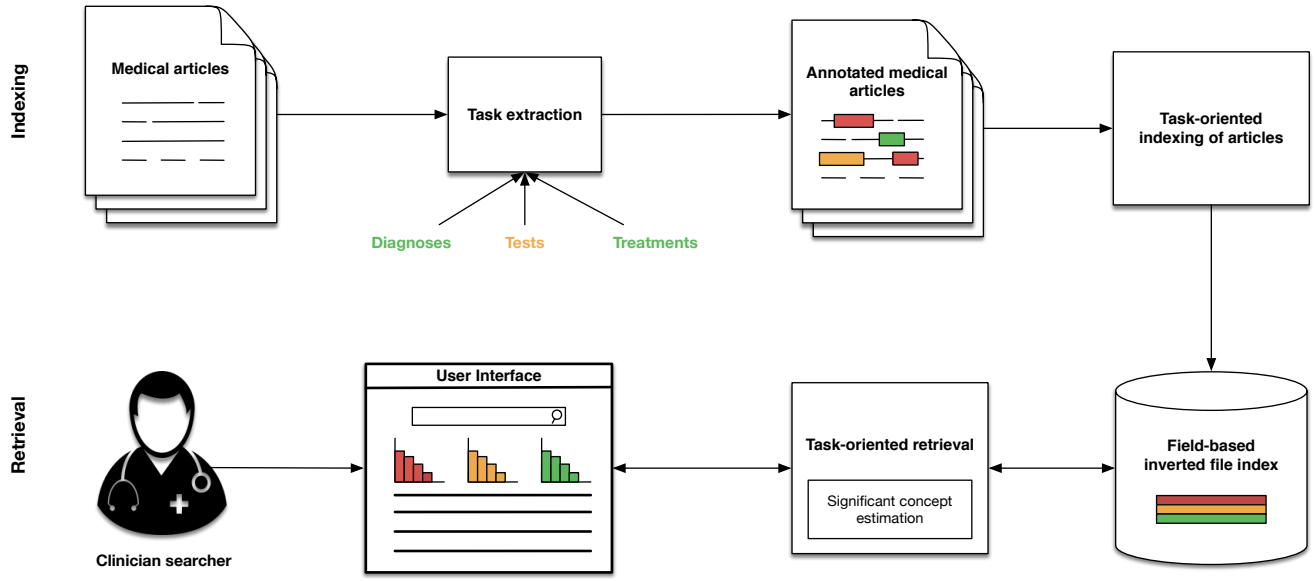

Fig. 3 Overview of the task-oriented search approach.

```
Patients with a <test UMLDid="C2238079" title="blood smear">blood smear</test> found to be positive for
<diagnosis UMLDid="C0024530" title="malaria [disease/finding]">malaria</diagnosis> were often administered
<treatment UMLDid="C0034414" title="quinidine [chemical/ingredient]">quinidine</treatment>.
```

Fig. 4 An excert of a PubMed article containing three task annotations: a test ("blood smear"), a diagnosis ("malaria") and a treatment ("quinidine").

3.4 Task-oriented indexing of articles

After all articles were annotated with relevant tasks, the articles were indexed. Indexing was performed using the Elasticsearch information retrieval library.[3]

Each article originally contained different fields: title, abstract, journal, publication date and article body. We added to these the three fields for diagnosis, test and treatment that were identified as part of the task extraction process.

All fields were indexed as separate fields in Elasticsearch. Separate fields allowed retrieval and ranking to be specific to particular field; for example, it allowed searching only on information related to treatments.

3.5 Task-oriented retrieval

When a user posed a clinical query, they would typically be provided with a long list of search results. In a task-oriented approach to EBM, it was desirable to provide the clinician with a summary of the significant diagnoses, tests and treatments. This allowed them to quickly gain an understanding of what they might have expected to find when examining the search results. In addition, if these summaries were interactive (e.g., the searcher could drill-down on specific tests or treatments) then they were provided with an easy mechanism to navigate the information space. Thus, given a set of search results, we aimed to estimate significant diagnoses, tests and treatments. (Recall that each diagnosis, test or treatment was in fact a UMLS concept, which may have actually been comprised of one or more terms.)

Each diagnosis, test and treatment concept was scored according to the frequency it appeared within a set of search results (foreground probability) vs. the frequency it appeared within the collection as a whole (background probability). A concept was considered significant if there was a noticeable difference in the foreground and background estimates. Formally, we derived a significance estimate below.

Let C be the set of articles in the entire collection of medical articles. Let $C_t \subseteq C$ be the subset of articles that contained task concept t. Let S_Q be the set of articles returned for a particular query, Q. The foreground probability of t was calculated as:

$$P_f(t|Q) \propto \frac{C_t \cap S_Q}{S_Q}. \quad (1)$$

The background probability of t was calculated as:

$$P_b(t) \propto \frac{C_t}{C}. \quad (2)$$

[3] Elasticsearch version 2.2.0: https://www.elastic.co/downloads/past-releases/elasticsearch-2-2-0.

The significance of t, given query Q was calculated as:

$$\text{sig}(t, Q) = \Big(P_f(t|Q) - P_b(t)\Big)\frac{P_f(t|Q)}{P_b(t)}. \quad (3)$$

For a given clinician's query, a set of diagnoses, tests and treatments could then be ranked in descending order of significance and provided back to the searcher (along with the regular search results for that query).

Elasticsearch provides a number of standard retrieval functions to score articles according to their relevance to a query. We adopted the default Elasticsearch retrieval model which is a Vector Space Model with TF/IDF weighting.

3.6 Visualisation of search results

A web-based user interface was developed to provide a clinician with a means to search and interact with the search results. The interface provided a single input box where clinicians could provide a free-text, keyword query. Retrieval results were displayed as a ranked list in decreasing order of relevance score to the query. Each result was comprised of the article title, journal title, publication date and a 'snippet' — that is, the portion of the article that matched the query keywords.[4]

Three barplots plots were also generated to present the top-five most significant diagnoses, tests and treatments, respectively. These plots were interactive: clinicians could click on a particular diagnosis, for example, and the set of search results would be filtered to include only articles mentioning that diagnosis. Multiple filters could be applied. The purpose of this was to allow the clinicians to, firstly, easily get an overview of the search results by inspecting the plots and, secondly, easily navigate the set of search results by applying various filters.

3.7 Task-oriented retrieval vs. topic type

To understand how different tasks affected retrieval, we first need to clarify the difference between topic type and task-oriented filter.

The topic type was stipulated in the TREC CDS shared task and represented the particular search task being performed for each topic: searching for diagnoses, searching for tests and searching for treatments.

In contrast, task-oriented filtering was the process of filtering search results based on the significant diagnoses, tests or treatments calculated by own method. As such, clinicians may not have used the same task-oriented filter as the one stipulated in the topic type. For example, a topic type that required searching for relevant *tests* to identify malaria may have led a clinician to actually filter via a *diagnosis* of Plasmodium (malaria parasite) so that they may have viewed the most significant tests for malaria.

We make this distinction between topic type and task-oriented filter so that we could consider both as part of our analysis on how different task affected retrieval.

3.8 Evaluation methodology

The relevance assessments from the TREC CDS shared task provided the gold standard against which retrieval systems could be empirically evaluated. In this study, we aimed to evaluate the retrieval effectiveness of the task-oriented approach. To this aim, we conducted two separate experiments outlined in next subsections.

3.8.1 Retrieval effectiveness evaluation

The first evaluation experiment resembled a standard TREC setting, where we compared the retrieval effectiveness of the system without filtering with that of the system with a specific task filter. This setting simulated the situation in which a clinician searcher either interacted with a standard search interface (no filtering) or with our system by selecting one of the specific task filters (facets).

As an evaluation measure we adopted two evaluation measures precision @ 10 and mean reciprocal rank. Precision @ 10 is the portion of relevant articles returned in the top 10 results retrieved.[5] Precision @ 10 was chosen because it captured the behaviour of a clinician performing a search and reviewing only the top 10 results returned. Precision @ 10 was also an official evaluation measure for the TREC CDS shared task. Mean reciprocal rank is the multiplicative inverse of the rank position of the first relevant article.[6] Reciprocal rank was chosen because it captures the behaviour of a clinician performing a search and looking through the rank list for the first correct article to their query.

Both precision @ 10 and reciprocal rank are precision based. Precision was favoured because clinicians, who are often time pressured, typically focus on finding a small set of high quality articles that allows them to perform their task (diagnosing, testing or treating).

[4] We used the default snippet generation provided by Elasticsearch.

[5] Formally, precision@$n = \frac{|\text{Rel} \cap \text{Ret}_n|}{|\text{Ret}_n|}$, where Rel is the set of relevant documents and Ret_n is the set of top n retrieved documents.

[6] Formally, recip. rank $= \frac{1}{\text{rank}}$, where rank is the rank position of the first correct result in a ranked list of results.

They would likely only review a limited set of search results [19]. In addition, TREC CDS used precision @ 10 as an official measure [25,24]. The pooling methodology — how documents were selected for judging — was done by selecting the top 20 results of the various teams that participated in TREC CDS. Thus the gold standard relevance assessments that we used were constructed in way that focused on early precision.

To evaluate the effectiveness of task-based filtering we conducted the following experiment. First, we issued each query topic to the retrieval system and, with no filtering, evaluated the corresponding precision @ 10 and mean reciprocal rank. We then simulated the clinician interacting with the results by selecting individual diagnoses, tests and treatments as filters. Specifically, we filtered the search results, one at a time, by each of the top-five diagnoses, tests and treatments; for example, filter with only the first treatment and evaluate the results, then filter with only the second treatment and evaluate the results, etc. Precision @ 10 and mean reciprocal rank were calculated after each filter had been applied. Thus, the change in effectiveness between the first ('No filter') search and each of the subsequent task-oriented searches could be calculated. The retrieval effectiveness of the three different task types could be compared and contrasted.

3.8.2 Cost model evaluation

The second evaluation experiment involved a cost-benefit analysis of interacting with the system. We considered the same type of user interactions (i.e., 'No filter' vs. selecting a single filter among diagnoses, tests and treatments) but compared the difference in gains and costs associated with these two different types of interactions. To model gains and costs we developed a simple cost model of the interaction the clinical searcher had with the system following the work of Azzopardi and Zuccon [3,4,2].

The cost model was used to determine whether, given the same amount of gain (thus, gain is constant), the interface with task-oriented filtering provided a lower interaction cost than the interface with no filtering. Specifically, we assumed that gain was associated with finding a relevant article and that the gain was constant for each of the relevant articles (the models can be extended to consider graded relevance/gain). Gain was defined as the clinician finding n relevant articles for a topic. Cost was defined as the number of articles that the clinician had to view before reaching the gain of n relevant articles; in addition, other costs of interaction were considered, e.g., posing a query or selecting a filter. The costs for filtering vs. not filtering could then be calculated and compared to determine the most economic system.

To develop the cost model, we assumed the following actions take place. The clinical searcher issued a query q; this action had a cost of C_q, which was experienced irrespective of the interface used. In the interface with no filtering (NF) option, the clinician would have to assess $N_{NF}(n)$ articles in order to have found exactly n relevant articles. The examination and assessment of an article (whether relevant or not) cost C_d. Thus, the cost of interacting with the interface without filtering to retrieve n articles was:

$$\mathcal{C}_{NF}(n) = C_q + N_{NF}(n) \cdot C_d \tag{4}$$

A similar cost model could be developed for the interface with task-oriented filtering; as with the previous evaluation study, we only considered a single task filter being applied. With task-oriented filtering, the action of selecting a specific filter cost C_f: this cost was made up of the clinician having to weigh up the filtering options that were displayed and then clicking on the chosen option. Using a task-oriented filter, the clinician would have to assess $N_F(n)$ articles in order to have found exactly n relevant articles. As for the previous 'No filter' setting, the examination and assessment of a article cost C_d. Thus, the cost of interacting with the interface by selecting a task-oriented filter to retrieve n articles was:

$$\mathcal{C}_{F}(n) = C_q + C_f + N_{F}(n) \cdot C_d \tag{5}$$

To compare whether the clinical searcher was better off using the interface with task-oriented filtering or not, we compared the associated costs at a fixed gain level ($n = k$). The clinician was better off using task-oriented filtering if $\mathcal{C}_F(n = k) < \mathcal{C}_{NF}(n = k)$. With some algebraic operations this condition becomes:

$$\begin{aligned} &C_f < C_d \cdot [N_{NF}(n = k) - N_F(n = k)] \implies \\ &\frac{C_f}{C_d} < N_{NF}(n = k) - N_F(n = k) \end{aligned} \tag{6}$$

Note that costs were expressed as positive real numbers and thus the ratio $\frac{C_f}{C_d}$ had a lower bound of zero (i.e., $\frac{C_f}{C_d} > 0$).

In the next section, we shall study when this condition was satisfied for varying values of n using the empirical results obtained on the TREC CDS shared task detailed above.

4 Results

4.1 Task extraction results

The task extraction process resulted in the occurrences of diagnoses, tests and treatments shown in Table 1.

Table 1 Occurrences of diagnoses, tests and treatments as a results of the task extraction process.

Task type	Number of occurrences	Median occurrences per article (SD)
Diagnoses	3,151,334	3 (5.1)
Tests	1,396,290	1 (2.7)
Treatments	5,994,993	7 (7.6)

Mentions of treatments were more common that tests or diagnoses. Mentions of tests were least common.

The number of occurrences per articles is shown in the histogram of Figure 5. Most articles had a small number of task mentions. Treatments exhibited a different trend of either containing no mentions or contain a median of 7 mentions.

A sample medical article from the user interface showing the annotated diagnoses and treatments is shown in Figure 6.

4.2 Visualisation of retrieval results

A screenshot of the user interface, presenting the results of a search for 'malaria', is shown in Figure 7. Three barplots provide an overview of the significant diagnoses (red), tests (orange) and treatments (green). The individual search results are displayed below the plots and include the article title, journal title, publication date and snippet (portion of the article where the search terms were found). Below the article title a summary of the diagnoses, tests and treatment contained *within that* article are displayed.

4.3 Retrieval effectiveness evaluation results

The retrieval effectiveness results for different task filtering are shown in Table 2. The 'No filter' method represented the baseline method of a clinician's initial search. The other task filters simulated the clinicians filtering the search results in a task-oriented manner. The results show that task-oriented filtering led to a statistically significant improvement in precision @ 10 and mean reciprocal rank. Filtering on tests exhibited the greatest improvement, followed by filtering on diagnosis and, finally, filtering on treatments.

Next, we consider how different topic types and different task-oriented filters affected retrieval. Figure 8 provides precision @ 10 results (y-axis) for different task-oriented filters (x-axis) and different topic types (colored legend). Each boxplot is comprised of the precision @ 10 for the group of topics within the particular topic type and filter.

Table 2 Retrieval results for task-oriented search. All results showed statistical significance over 'No filter' baseline (paired t-test, $p < 0.01$).

Task-oriented filtering	Prec.@10 (%Δ)	Mean recip. rank (%Δ)
No filter	0.2867	0.4349
Diagnoses	0.3250 (+13%)	0.5271 (+21%)
Tests	0.3283 (+15%)	0.5324 (+22%)
Treatments	0.3167 (+10%)	0.5113 (+16%)
By topic type	0.3183 (+11%)	0.5320 (+22%)

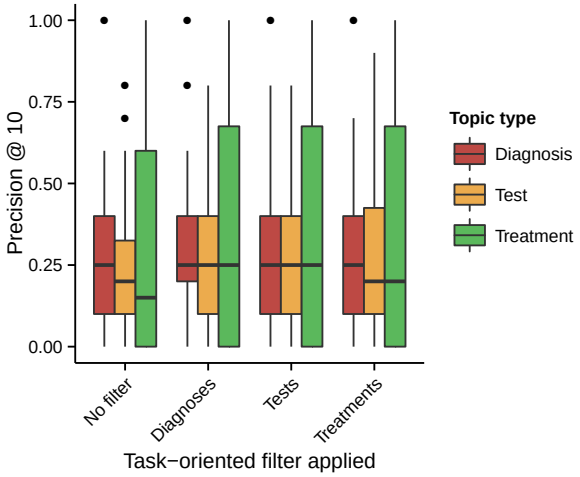

Fig. 8 How different task-oriented filters (x-axis) and different topic types (colored legend) affected precision @ 10 (y-axis). Task-oriented filtering mainly improved topics that required searching for tests and treatments. Dealing with treatments related information proved more challenging than diagnoses and tests.

First, we consider the 'No filter' case (first x-axis category) and the effect of topic type on precision @ 10. Treatment topics exhibited poor performance compared to diagnosis topic (test topics were in between). This shows that searching for treatments is generally harder than diagnoses and tests.

In contract, when task-oriented filtering is applied we observed that the poor performance on tests and treatments was mostly alleviated. Stated alternatively, the gains from using task-oriented filtering all came from test and treatment topic types. For treatment-oriented filtering (last x-axis category), the performance on test and treatment topic types was lower. Once again, this shows that dealing with treatment related information is generally more challenging. (We consider why this may be the case in the discussion.)

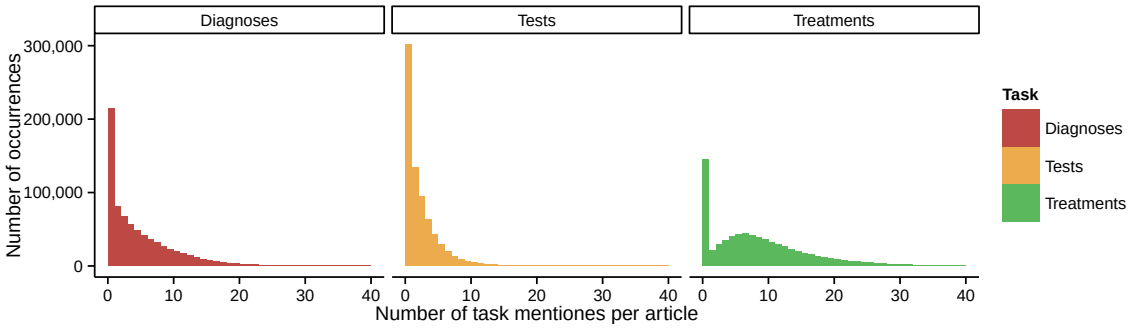

Fig. 5 The number of task mentions per article.

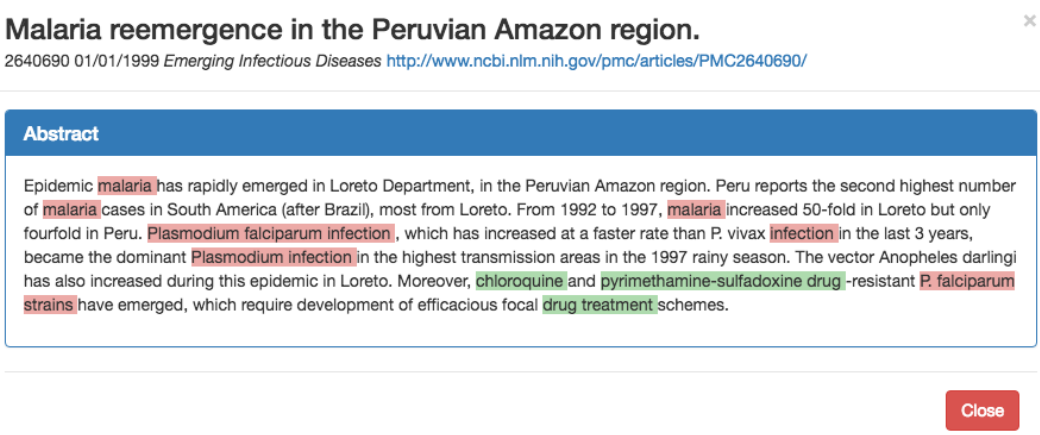

Fig. 6 An sample medical article from the user interface showing annotated diagnoses (red) and treatments (green).

4.4 Cost model evaluation results

The cost model measures the number of articles that a clinician would have to view before reaching k relevant articles and allows us to compare the cost of interaction when using an interface with filtering vs. an interface with no filtering. Specifically, we recall from Section 3.8.2 that the clinician is better off using the filtering interface when inequality 6 is satisfied.

In Figure 9 we plot the right hand side (RHS) of inequality 6 (i.e., $N_{NF}(n) - N_F(n)$, for varying values of $n = k$) according to the empirical results of our experiments. The interface with filtering provided a better (cheaper) interaction when the left hand side (LHS) of inequality 6 was smaller than RHS. However, recall that LHS has zero as a lower bound, because costs were assumed positive, i.e., the LHS cannot be negative. This means that the filtering interface was a better choice when the ratio between the cost involved with filtering and that of assessing an article was at any point above the zero line but below the RHS line. The values of the LHS ratio that satisfy such a condition have been represented by the colored areas in Figure 9. For example, when seeking tests (i.e., topic types is Tests) and wanting to find $k = 13$ relevant articles, the clinician was better off using the diagnoses task-oriented filter if the cost of filtering was up to 15 times more expensive than the cost of assessing an article; if the relative cost of filtering was instead more expensive than that, then the clinician was better off not using the filter.

According to the results shown in Figure 9, in the majority of the cases and across all task types, there was a filter that could be applied to the interface such that the filtering provided cheaper interaction than the no filtering condition, provided that the cost of filtering was below a certain value compared to the cost of assessing an article (typically in the region of filtering being up to 10-20 times more expensive than assessing). We also note that the tests filter provided the best savings when the task was to seek for diagnoses or tests; while the treatments filter provided the best savings when seeking treatments themselves. Using the diagnoses filter to seek for diagnoses never provided savings over using no filter; it did provide savings for specific values of k when applied to seeking tests and treatments.

We further use the developed cost model to understand the benefit of using an interface with filters, by

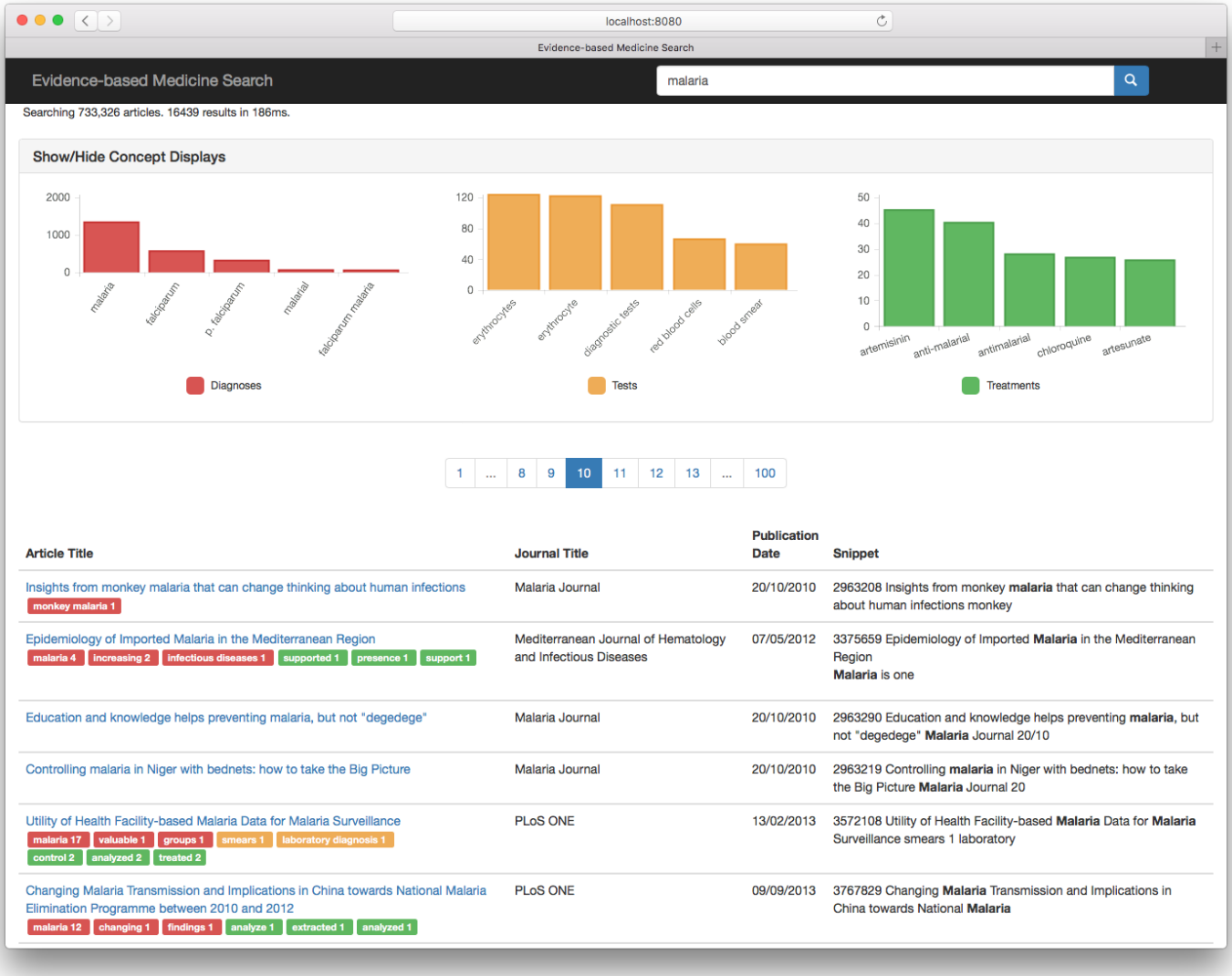

Fig. 7 Screenshot showing the results of a search for 'malaria'. Three barplots provide an overview of the significant diagnoses (red), tests (orange) and treatments (green). Individual search results are shown below the histograms.

considering specific instantiations of the filtering and assessing costs derived from the literature. These will identify savings in terms of workload clinicians mat achieve using the developed interface.

To estimate the time required to use the filtering option, we used the GOMS Keystroke Level Model [5] and the timings reported in previous work [26,1]. The GOMS model associates to each low level interaction with an interface a time estimate in seconds. When using the filter option, the user was likely to move the hands from the keyboard to the mouse (H = 0.4 seconds) because the previous action may have been typing the query, point the mouse to the filter option (P = 1.1 seconds), click on the filter (C = 0.2 seconds), mentally prepare for the actions (M = 1.35 seconds); the model also includes an overhead for the system response time (R = 0.8 seconds). In total, selecting a filter required a cost of 3.85 seconds; to this cost estimate we need to add the time required to visually and mentally analyse the filter options (A). We treat this as a variable in the analysis below because we do not have access to reasonable estimates of this cost for our (or similar) interface. Thus, we express the cost of filtering as $C_f = 3.85 + A$.

To estimate the time required to assess an article, we rely on data reported by previous work. Azzopardi et al. [1] found that to assess documents in a common web-style search interaction, users took on average 16 seconds (E1). Turpin et al. [31] instead reported times of 88 seconds (E2) to evaluate a web document for relevance (as a document relevance assessor though, rather

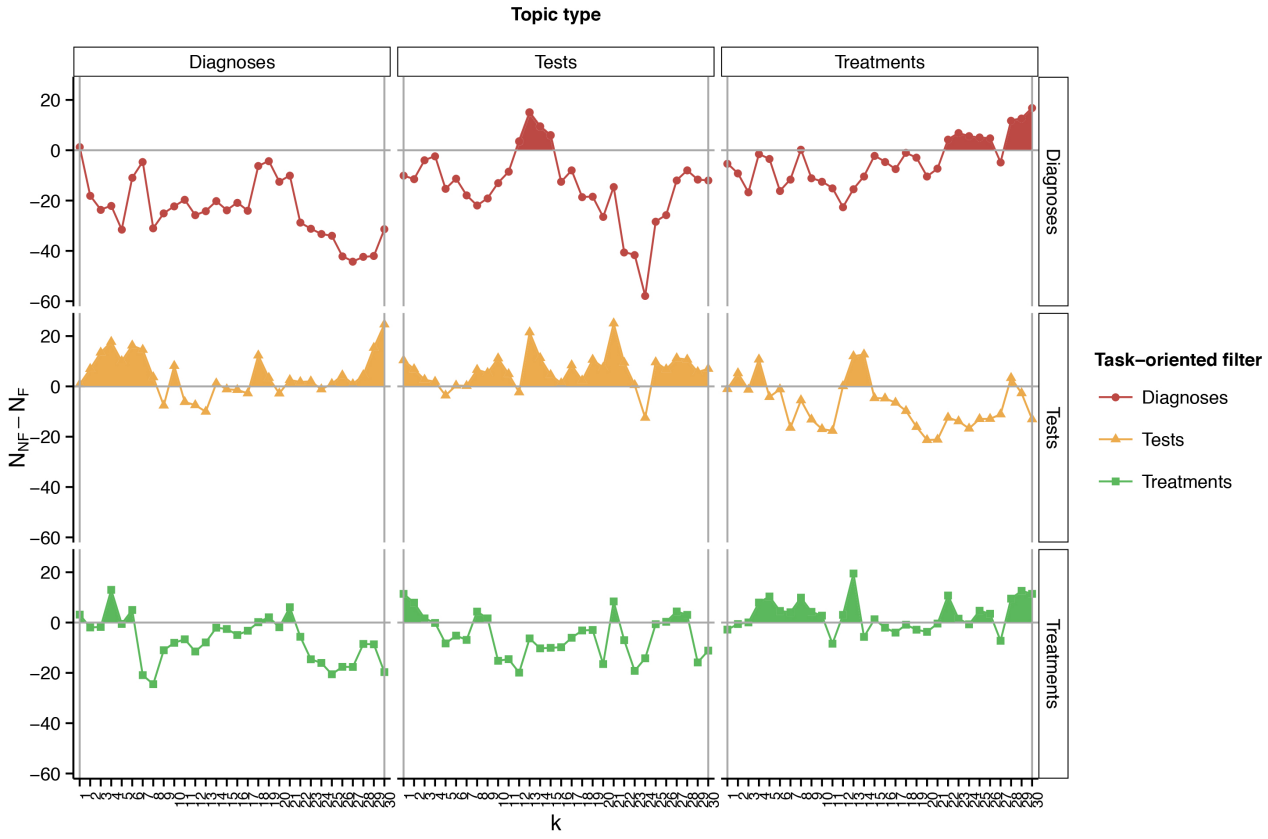

Fig. 9 When is filtering worthwhile? Here we report the values obtained for the RHS of inequality 6 using the data collected in our experiments. The coloured areas identify conditions when values of the LHS of inequality 6 indicate task-oriented filtering was more convenient than the 'No-Filter' interface.

Table 3 Cost estimates, in seconds, for the time required to assess an article and relative maximum value of the time to evaluate filtering options (A) for filtering to be worthwhile, with respect to the difference in number of documents to be assessed to achieve a fixed level of recall (n) when using filtering vs no filtering.

Est.	C_d value	max. A
E1	16	A = 16 * ($N_{NF}(n) - N_F(n)$) - 3.85
E2	88	A = 88 * ($N_{NF}(n) - N_F(n)$) - 3.85
E1	185	A = 185 * ($N_{NF}(n) - N_F(n)$) - 3.85

than a user). Finally, in the medical domain, Koopman and Zuccon [14] reported that clinicians took on average 185 seconds (E3) to assess a document for relevance, where in that study a document was one or more electronic health records. We use these three different estimations to instantiate the cost of assessing a document, C_d.

Table 3 reports the maximum values of time required by the clinician to evaluate filtering options (A) for the filtering interface to be worthwhile (cheaper than the counterpart interface with no filters). The values of A are computed with respect to the three different estimates of the cost (E1–3), and the difference in numbers of articles to be assessed with vs. without filtering.

When used along with the data recorded in Figure 9, the estimates of Table 3 suggest, e.g., that, when seeking for 13 relevant articles related to tests ($N_{NF}(13) - N_F(13) = 15$), filtering with the diagnoses filter is worthwhile if the time A required to chose which filter to apply among those provided by the interface is at most 236.15 seconds for E1 ($\approx$ 4 minutes), 1,316.15 seconds for E2 ($\approx$ 22 minutes), and 2,771.15 seconds for E3 ($\approx$ 46 minutes). Here it is fair to assume that the time required to decided upon which filter is in the order of tens of seconds, or up to a minute, and thus, in this case, filtering would provide substantial time savings. Note that if the tests filter rather than the diagnoses filter was used in the same conditions (and $N_{NF}(13) - N_F(13) = 21$), then the saving would be even more considerable.

A similar reasoning could be used to derive what is the minimum number of articles that the user will have to save assessing for the filtering interface to be

worthwhile, i.e., $N_{NF}(n) - N_F(n)$. Following the assumption that the time required to decide upon which filter to select is up to one minute ($A = 60$ seconds), then $N_{NF}(n) - N_F(n)$ is 4 articles for E1 and only 1 article for E2 and E3.

5 Discussion

The task extraction process (Section 3.3) provided a means to identify and annotate mentions of diagnoses, tests and treatments found in free-text medical articles. These annotations provided a from of semantic enrichment to the medical articles. In our study, this was used to improve information retrieval effectiveness; however, the same annotations could be used in many other applications where it is desirable to demarcate diagnoses, tests and treatments.

In general web-search, organising information around different categories is often referred to as faceted retrieval — the different task-oriented annotations represent the facets. In web-search, one of drawbacks of faceted retrieval is that the various categories need to be manually created and maintained. (Although there have been attempts to alleviate this via semi-supervised creation methods [29].) The advantage of the task extraction process outlined in this study is that the process is completely automated: the categories were taken from the concepts contained within the UMLS thesaurus and the extraction process is unsupervised.

The task-oriented annotations were successfully integrated within a retrieval method via significant task filtering. This proved an effective retrieval method: statistically significant improvements in precision @ 10 and mean reciprocal rank were observed when filtering was compared to an initial search with no filter. This showed that taking into account different clinical tasks at retrieval time led to improvements in retrieval effectiveness.

The task-oriented approach can be seen as a type of facet-based retrieval, which aims to reduce metal workload on the searcher [33]. Our economic analysis using the cost model evaluation showed that filtering was able to reduce the overhead on clinicians, even when a very conservative estimate of the cost of filtering at 60 seconds was used. The time it would take a clinician to decide on a filter would likely be far lower than 60 seconds, thus the overall savings would be greater.

The analysis of how different tasks affected retrieval showed that some tasks were harder than others. When no filtering was applied, poorer performance was observed on tests and particularly on treatments. The task-oriented approach mitigated this by improving test and treatment topics. The fact that each task type displayed different retrieval results may reveal that each task had differing requirements from a clinician searcher point-of-view. As such, a retrieval system that is more adaptive to topic type would be advantageous and warrants future investigation.

In general, dealing with treatment information (topic type and task-oriented filtering) proved the most challenging from the retrieval effectiveness standpoint. One reason for this is that treatments may suffer more from the vocabulary mismatch problem [17,15] — the difference between how a treatment is expressed in different settings (e.g., between the clinician's query and the medical article). A major source of vocabulary mismatch is in the way medications (which are considered treatments) are expressed; for example, a drug can be expressed as its brand name, its generic name, or its active ingredients. The impact of vocabulary mismatch, we posit, was the source of the challenge with treatments.

In the current system, the clinician was presented with a list of search results (showing article title and journal name) and, if they selected a result, they could view the entire article. However, many articles are lengthy and clinicians may only be interested in specific portions (likely, the portion matching their task). As such, a 'passage' retrieval system that displays only the relevant section(s) would be preferred. The task-oriented annotation may actually aid in this regard. The relevant passage to display to the clinician could be derived based on the location of the clinicians query keywords, their stipulated task and the location of matching task annotations in the article. The investigation of task-oriented passage retrieval is left to future work.

In this study, we simulated the clinician filtering search results with a single diagnosis, test or treatment. Multiple filters were not evaluated, although the user interface did support this and in a real-world setting multiple filters would be used. It would have been possible to simulate the clinician applying multiple filters. However, the ultimate evaluation of the effectiveness of the system would be to conduct a direct user study with clinicians (for example, an A/B test [12] with and without task-oriented filtering). This is left for future work.

6 Conclusion and future work

Clinicians pose queries around the three clinical tasks of searching for diagnoses, searching for tests and searching for treatments. As such, we investigated incorporating these three tasks as part of a task-oriented approach

to searching for evidence-based medicine resources in digital libraries.

The task-oriented approach has a number of components: extracting task-specific information from free-text medical articles; indexing task-specific information in a information retrieval system; task-specific retrieval that identifies significant diagnoses, tests and treatments from a set of search results; and task-oriented visualisation and interaction via a user interface.

An empirical evaluation showed that taking into account different clinical tasks led to improvements in retrieval effectiveness. Our analysis also showed that some tasks were harder than others; specifically, dealing with treatments proved the most challenging (likely due to the vocabulary mismatch problem).

A cost-benefit analysis of interacting with the system showed that task-oriented filtering represented a cost saving to the clinicians when compared to no task-oriented filtering.

Future work includes: investigating additional clinical categories beyond diagnoses, tests and treatments; evaluating the effect of multiple task-oriented filters; and an extensive user study.

References

1. Azzopardi, L., Kelly, D., Brennan, K.: How query cost affects search behavior. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 23–32. ACM (2013)
2. Azzopardi, L., Zuccon, G.: Building and using models of information seeking, search and retrieval: Full day tutorial. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, pp. 1107–1110. ACM, New York, NY, USA (2015)
3. Azzopardi, L., Zuccon, G.: An analysis of the cost and benefit of search interactions. In: International Conference on the Theory of Information Retrieval (ICTIR). Newark, USA. (2016)
4. Azzopardi, L., Zuccon, G.: Two scrolls or one click: A cost model for browsing search results. In: European Conference on Information Retrieval, pp. 696–702. Springer (2016)
5. Card, S.K., Moran, T.P., Newell, A.: The keystroke-level model for user performance time with interactive systems. Communications of the ACM **23**(7), 396–410 (1980)
6. Demner-Fushman, D., Lin, J.: Knowledge extraction for clinical question answering: Preliminary results. In: Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains, pp. 9–13 (2005)
7. Demner-Fushman, D., Lin, J.: Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics **33**(1), 63–103 (2007)
8. Druss, B.G., Marcus, S.C.: Growth and decentralization of the medical literature: implications for evidence-based medicine. Journal of the Medical Library Association **93**(4), 499–501 (2005)
9. Ely, J., Osheroff, J., Gorman, P., Ebell, M., Chambliss, M., Pifer, E., Stavri, P.: A taxonomy of generic clinical questions: classification study. British Medical Journal **321**(7258), 429–432 (2000)
10. Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Yee, K.P.: Finding the flow in web site search. Communications of the ACM **45**(9), 42–49 (2002)
11. Hersh, W.: Information retrieval: a health and biomedical perspective, 3rd edn. Springer Verlag, New York (2009)
12. Hofmann, K., Li, L., Radlinski, F., et al.: Online evaluation for information retrieval. Foundations and Trends in Information Retrieval **10**(1), 1–117 (2016)
13. Koopman, B., Zuccon, G.: Why assessing relevance in medical IR is demanding. In: Proceedings of the SIGIR Workshop on Medical Information Retrieval (MedIR). Gold Coast, Australia (2014)
14. Koopman, B., Zuccon, G.: Why assessing relevance in medical ir is demanding. In: Medical Information Retrieval Workshop at SIGIR 2014, p. 16 (2014)
15. Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., Lawley, M.: Information retrieval as semantic inference: A graph inference model applied to medical search. Information Retrieval **19**(1), 6–37 (2015)
16. Kwasnik, B.H.: The role of classification in knowledge representation and discovery. Library trends **48**(1) (2000)
17. Lancaster, F.W.: Vocabulary Control for Information Retrieval, 2nd edn. Arlington, Virginia, Arlington, Virginia (1986)
18. Larsen, P.O., von Ins, M.: The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics **84**(3), 575–603 (2010)
19. Lau, A.Y., Coiera, E., et al.: How do clinicians search for and access biomedical literature to answer clinical questions? Studies in health technology and informatics **129**(1), 152 (2007)
20. Limsopatham, N., Macdonald, C., Ounis, I.: A Task-Specific Query and Document Representation for Medical Records Search. In: Proceedings of the 35th European Conference on Information Retrieval (ECIR). Moscow, Russia (2013)
21. Liu, Z., Chu, W.W.: Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Information Retrieval **10**(2), 173–202 (2007)
22. National Library of Medicine: Detailed indexing statistics: 1965-2015 (2016). URL https://www.nlm.nih.gov/bsd/index_stats_comp.html
23. Richardson, W.S., Wilson, M.C., Nishikawa, J., Hayward, R.S., et al.: The well-built clinical question: a key to evidence-based decisions. ACP Journal Club **123**(3), A12–3 (1995)
24. Roberts, K., Simpson, M.S., Voorhees, E., Hersh, W.R.: Overview of the TREC 2015 clinical decision support track. In: Text REtrieval Conference (TREC) (2015)
25. Simpson, M.S., Voorhees, E.M., Hersh, W.: Overview of the TREC clinical decision support track. In: Text REtrieval Conference (TREC) (2014)
26. Smucker, M.D.: Towards timed predictions of human performance for interactive information retrieval evaluation. In: Proceedings of The Third International Workshop on Human-Computer Interaction and Information Retrieval (HCIR 2009) (2009)
27. Soergel, D.: The rise of ontologies or the reinvention of classification. Journal of the Association for Information Science and Technology **50**(12), 1119 (1999)

28. Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: SIGIR Medical Information Retrieval (MedIR) Workshop (2016)
29. Stoica, E., Hearst, M.A.: Nearly-automated metadata hierarchy creation. In: Proceedings of HLT-NAACL 2004: Short Papers, pp. 117–120. Association for Computational Linguistics (2004)
30. Trieschnigg, D., Hiemstra, D., de Jong, F., Kraaij, W.: A cross-lingual framework for monolingual biomedical information retrieval. In: Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 169–178. ACM (2010)
31. Turpin, A., Scholer, F., Jarvelin, K., Wu, M., Culpepper, J.S.: Including summaries in system evaluation. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 508–515. ACM (2009)
32. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18**(5), 552–556 (2011)
33. White, R.W.: Interactions with search systems. Cambridge University Press (2016)