



Focused Query Expansion with Entity Cores for Patient-Centric Health Search

Erisa Terolli¹✉, Patrick Ernst², and Gerhard Weikum¹

¹ Max Planck Institute for Informatics, 66123 Saarbrücken, Germany

{eterolli,pernst,weikum}@mpi-inf.mpg.de

² Amazon, Berlin, Germany

Abstract. The Web provides a plethora of contents about diseases, symptoms and treatments. Most notably, users turn to health forums to seek advice from doctors and from peers with similar cases. However, the benefit of forums mostly lies in community QA and browsing. Expressive querying for patient-centric needs is poorly supported by search engines. This paper overcomes this issue by enriching user queries with judiciously chosen entities and classes from a large knowledge graph. Candidate entities are extracted from the full text of user posts. To counter topical drift that would arise from picking *all* entities, we devise ECO, a novel method that computes a *focused entity core* for query expansion. Experiments with contents from health forums and clinical trials demonstrate substantial gains that ECO achieves over state-of-the-art baselines.

1 Introduction

Motivation. The Internet provides a wealth of online content about health topics, including linked open data about drugs and diseases (e.g., drugbank.ca and disease-ontology.org), scientific articles about biomedical research in PubMed¹, online portals with encyclopedic entries to inform doctors and patients (e.g., mayoclinic.org and patient.info/health), all the way to online health communities (e.g., patient.info/forums and healthboards.com). All these contents are indexed by search engines, but the query result quality is fairly poor (compared to queries about music, movies, games etc.); it is often a tedious process to find relevant answers [1].

Advances on health search and QA [24] have mostly focused on specific kinds of information needs and content sources: short consumer queries that can tap health portals on topics such as “Alzheimer’s treatments” or “Aricept side effects” (e.g., [18, 46]), expert queries on scientific articles (e.g., [36, 41]) such as “pancreatic cancer risk with DPP4 inhibitors”, and specialized retrieval over electronic patient records or clinical notes (e.g., [7, 25, 44]). In contrast, *patient-centric* needs focusing on queries about individual health situations posed by patients themselves or general physicians on their behalf, have received little-to-no attention.

¹ <https://pubmed.ncbi.nlm.nih.gov/>.

Aricept has made my mum worse!!

#1

Hello All

My mum was diagnosed with early stages Alzheimers at the end of June but she has been suffering loss of memory for probably 3 years but as usual we put it down to 'old age'. My dad is 85 and in wonderful health, mum is 76 She had her first appointment with the local memory clinic and they prescribed 5mg Aricept. Day 1 she was fine, no side effects; day 2 she started feeling sick and then was very ill for the best part of 24 hours. She also hallucinated did not recognise her home, my dad, saw my gran (deceased) etc. We stopped the medication immediately, called the memory clinic (who didn't call back) so we have been trying to cope, for the first time, with my mum in a horrible state of confusion and anger, where she thinks we are poisoning her, trying to steal her money, not let her go home... the paranoia is endless. Can anyone please tell me how this has happened? How is she SO confused?

Fig. 1. Example of user post from forum.alzheimers.org.uk.

Example. Consider someone with Alzheimer's Disease who has taken specific drugs for years (e.g., Aricept or Risperidone) but starts to suffer from various forms of confusion. Looking up portal pages about Alzheimer's or Aricept, it is tiresome to find advice for the individual user's case, and searching PubMed is not useful either. The best source rather would be *online health forums*, where patients share experiences and doctors offer advice. The user has several options:

- i) Post a question in the forum. Figure 1 shows an example with a post title and a description in the post body. Then, the user would wait for good replies by doctors or other patients in the QA community.
- ii) Browse the forum, navigating over posts and topical links. This is tedious and time-consuming, but may eventually lead to helpful results such as the one shown in Fig. 2.
- iii) Fill the forum's search box to query posts of other users. This faces the problem that the user's individual needs are not easily cast into a crisp set of keywords. Using the post title alone is too unspecific. Using the full post body leads to a long, verbose and diffuse query.

A general physician (GP) who searches on behalf of the patient may consider also tapping *clinical trials*: empirical studies of patient cohorts (e.g., clinicaltrials.gov). However, the GP would also struggle with the limitations of the search engine. The goal here is to find information that is *individually relevant* for the patient, taking into account the specific description in the user's post or the doctor's initial assessment.

Aricept

My mother went through a similar phase of accusing my father of affairs and she had paranoid delusions. These did get worse concurrently with being on Aricept, but it is simply impossible for me to say whether this was down to the Aricept or the progress of the disease. As FifiMo post indicates, Aricpet can make patients aggressive. As the delusions and paranoia persisted when she came off Aricept, I'm inclined to say it was probably mainly the progression of the disease in my mother's case. Risperidone has toned down the symptoms of paranoia, but it hasn't got rid of it altogether.

Fig. 2. Example of reply from forum.alzheimers.org.uk.

Problem. This paper addresses the case of *patient-centric* information needs over online contents of patient experiences. The primary source for this purpose is online health communities. Users are patients who are unhappy with their current treatment, because they have non-standard symptoms and the diagnosis is unclear, or because they suffer from adverse side effects of their therapies. The goal is to find, for an individual case, similar patients and specifically related advice by doctors. For the example post of Fig. 1, we aim to automatically retrieve the post of Fig. 2 as a highly related and useful result.

In addition to health forums, we also consider searching clinical trials for patient-centric needs, expecting that studies with similar cohorts can be helpful.

Design Space and Approach. We aim to aid users and doctors by automatically generating *user-specific* and *coherent* queries from the description that a user puts in a forum post. To this end, a number of design choices could be pursued.

An Information Retrieval (IR)-style approach would employ *query expansion* [6], by combining the user question with the terms in the full text of the post body, with term weights derived from forum statistics. However, this will lead to broad and noisy queries. A machine-learning approach could learn to classify relevant posts, with training data based on “thank you” indicators in the forum’s threads. However, the *training data* will be *scarce and noisy*, and the approach would not work for highly individualized needs. A Semantic-Web approach could identify *named entities* in the user posts and link them to entries in a *knowledge graph* and other Linked-Data resources. The question could then be translated into a crisp entity-aware query (e.g., [38]). However, this provides no guarantee that the query keeps its focus, as user posts often contain cues for remotely related entities.

The design choice put forward in this paper is a combination of the IR and Semantic-Web paradigms. We build on query expansion, and use extracted named entities and a knowledge graph to generate expanded queries.

Contribution. To counter the dilution from adding too many entities, we devise a novel method to identify an *entity core* for each query, which is a compact sub-graph of the knowledge graph. We utilize KnowLife² [12,13] which integrates various Semantic-Web resources like UMLS, DrugBank etc. Entities qualify for query expansion if and only if 1) they are highly *relevant* for the user post and 2) they are *coherent* with each other so that jointly they have a clear focus.

In the Alzheimer’s example provided earlier, the list would include memory loss, aricept, feeling sick, hallucination, death, aggression, confusion, poisoning, and many more. Some of these are merely peripheral and misleading. A coherent core should focus on the key entities and classes: aricept, hallucination, dementia, and a few more. Our method, called ECO (for *Entity Cores*), computes this entity core (EC) using advanced graph algorithms [17,20] and harnesses the EC for judicious query expansion.

² <http://knowlife.mpi-inf.mpg.de/>.

ECO is designed for patient-centric search over health contents, but can be carried over to other domains.

The salient contributions of this work are:

- A novel method for query expansion, by computing entity cores that identify the most relevant and coherent terms for focused expansion.
- Experimental studies with model patients for 20 different diseases, studying two cases: search over health forums and search over clinical trials. The results show the superiority of the ECO method over baselines of entity-aware query expansion.
- Data and code are accessible at: <http://eco.mpi-inf.mpg.de/>.

2 Related Work

Health Search. Early work on health search (e.g., [31,32]) focused on query rewriting based on the MeSH ontology and interactive user support for better recall and result diversification. Recent works (see the tutorial [24] and references therein) have mostly shifted the attention to clinical texts and leveraging domain knowledge for expert search. A major exception is [46] on consumer health search over general Web pages, which aims to bridge end-user and clinical vocabularies via knowledge bases like UMLS and dictionaries like CHV (Consumer Health Vocabulary).

Search over health forums has been addressed in few projects (e.g., [12,34]), with basic methods for retrieval and ranking. The demo paper [14] presented a system architecture for personalized search, but merely sketched a high-level methodology without technical detail. [43] analyzed health-related queries in large search engine logs. [21] investigated the quality of community-QA responses from health forums, finding that they are more useful than results from Web search. Studies on health forums investigated dimensions like misinformation or emotions (e.g., [22,33]). The work of [10] tackled the task of detecting narrative patient posts in health forums, by means of a supervised classifier. Topical classification for health content in Reddit discussion threads was investigated by [5], proposing word-embedding-based clustering methods. None of these works addresses the search problem.

Benchmark competitions like CLEF Consumer Health Search Task [42] and TREC Precision Medicine Track [36] focus on broad queries over general web pages and specialized search over scientific publications, respectively. Neither considers online health communities. CLEF addresses general queries such as “infectious disease prevention”; one of the tasks is personalization, but this is with regard to the user’s level of expertise and comprehension of search results, not regarding the user’s individual health situation.

Query Expansion. Query expansion is a classical IR topic (see, e.g., [6]). As sources for expansion terms, most works considered either initial search results assuming pseudo relevance feedback or background corpora for computing semantic relatedness measures between terms (see, e.g., [15,16]). A recent

trend is to incorporate latent embeddings learned from large text collections into the relatedness scores for query expansion (e.g., [27, 28]). The works of [9, 40, 45] studied query reformulation for medical search over clinical trials, health records and scientific articles. None of these considers the user’s individual health situation. Recently, health-specific language embeddings, like Bio-BERT, and neural learning have been utilized to advance question answering over scientific PubMed articles [2, 19, 29].

Semantic Web and Knowledge Graphs. Semantic-Web research on health data has mostly addressed the horizontal integration of structured data. The iASiS project [26] pursues the goal of semantic data integration towards personalized precision medicine. The Horus.AI project [11] builds services for patient monitoring on specific conditions (e.g., diabetes). The Thalia project [41] provides a semantic search engine for PubMed articles, to support biomedical experts. SemEHR [44] harnesses semantic background knowledge to enhance search over clinical notes. In these and further projects of similar kind, online health communities are out of scope, and layperson queries is not an issue either.

Knowledge graphs (KGs) have been leveraged as a source of relevant entities and types for query expansion or query translation. A major focus here is on bridging the gap between user vocabulary, such as “high blood sugar”, and biomedical terminology, such as “hyperglycemia” [24, 35].

Entity-aware Query Expansion. Query expansion by including (appropriately weighted) terms from entities and types in general-purpose KGs, such as Freebase or Wikipedia-derived, has been investigated by [3, 8, 30, 35, 38] and others. Especially for entity-seeking queries [4] this has proven to be a powerful asset, in broad domains like searching for companies, products or entertainment works.

For health search, these approaches have been pursued to a lesser extent. Notable works, where the KG is constituted by domain-specific knowledge from MeSH, UMLS, CHS or health-centric parts of Wikipedia, include [12, 23, 46]. For consumer health search with short keyword queries, [46] conducted a systematic experimental study with a large variety of KG configurations. One of their findings is that it is crucial to configure all details properly for good performance, and identifying such good configurations is a difficult task by itself.

In our previously unexplored setting with forum contents and patient-centric queries, we expect these difficulties to be even more pronounced. In particular, unrestricted KG-based expansion with a large number of query-related entities and types as additional search terms, does not work at all. Even when term weights are carefully tuned, the expanded queries tend to be diluted and suffer from topical drifts. This is a key motivation for our approach of *focused* expansion using entity cores. In our experiments, we consider the prior works on KG-based expansion as baselines.

3 The ECO Method for Coherent Query Enrichment

3.1 Overview

ECO aims to answer patient-centric information needs consisting of

- an informational search query: a short phrase or few keywords, e.g., mentioning a disease or side effect;
- a patient-specific case description: free text about the user’s individual health situation and anamnesis.

The post title “*Aricept has made my mum worse*” together with the case description in the post body, which are shown in Fig. 1, are a good example how a online forum post can cover such information needs comprehensively. Given such information, our approach is able to generate coherent queries, which are tailored to the information needs of the user. The generated queries are executed over health-forum contents or, alternatively, clinical trials. Figure 3 gives an overview of the ECO architecture. In the following, we discuss its key components.

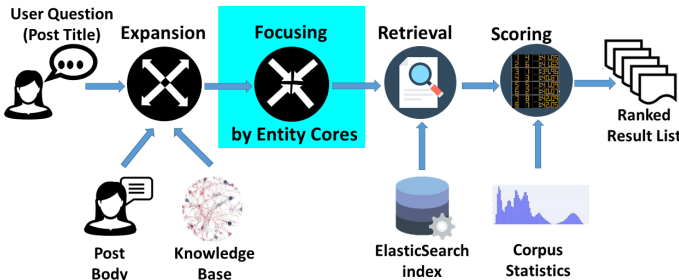


Fig. 3. ECO overview.

Health Corpus & Index. We retrieve search results from a corpus consisting of 1 million crawled and indexed forum threads and 100k clinical trials. A forum thread is a discussion starting with an initial post, having a short informational post title and an elaborate post body which contains a description of the user’s individual situation (as shown in Fig. 1). We leverage such posts as starting point for generating queries, since they extensively cover patient-centric information needs as aforementioned. This post is followed by a sequence of replies by other users. Each thread belongs to a topical sub-forum. For instance, for the health forums we considered in Sect. 4.2, such sub-forums contain threads for a specific (family of) disease(s) or drug(s). For example, on healthboards.com most of the Thyroid-centered posts can be found in the *Endocrine > Thyroid Disorders* and *Endocrine > Hormone Problems* sub-forums. Even though this categorization is descriptive, relevant posts are often spread out across sub-forums, which can be inter-related.

Clinical trials are taken from clinicaltrials.gov and are semi-structured study records, which obey a fixed schema. This schema represents a broad-spectrum of patient-specific information in the form of semantic structures, such as medical conditions, genders and age-ranges considered within the trial and free-text fields, such as study description and summary.

All documents are semantically enriched by the components described in Sect. 3.2. In order to execute search queries, we store and index all free-texts and various semantic assets, such as extracted entities and categories, using ElasticSearch³.

Knowledge Graph. We use the KnowLife resource as our knowledge graph (KG), comprising:

- entities from UMLS (Unified Medical Language System⁴, together with basic relations from its source vocabularies (e.g. mereological properties over anatomical concepts or dosage forms of drugs),
- subject-property-object triples for relational statements compiled by KnowLife [13] from a variety of sources, capturing symptoms, treatments of diseases, side effects of drugs, etc.,
- types from the DeepLife project [12], containing general categories (e.g., *endocrine system disorders* subsuming *thyroid disease* among others) as well as categories derived from facts (e.g. *risk factors for thyroid disease*).

In total, the KG contains 3.2 million entities, 323,862 types, 2,170,660 property triples and is stored as RDF triples in a Neo4J graph database.

Query Processing. Given a patient-centric information in the form of initial forum posts, ECO starts with the few keywords stated in the post title and enriches it in two stages, by leveraging the knowledge graph as follows:

- **Expansion:** Information extraction is performed on the provided case descriptions, i.e. the body of health forum post, to identify entities in the KG that are specific for the medical situation of the patient. These entities are added to the query.
- **Focusing:** The expanded query is often too broad, with the risk of drifting away from the user’s intent and needs. Therefore, we refocus the query by computing a coherent core of most relevant entities. This way, the expanded query is reduced into a more concise form, to ensure that query answers are focused on the user’s individual needs.

The focused query, comprising the gathered information (i.e. keywords, entities and semantic categories), is executed on ElasticSearch using a custom scoring function to compute the final search result ranking as explained in Sect. 3.4.

³ <https://www.elastic.co>.

⁴ <https://www.nlm.nih.gov/research/umls/>.

3.2 Query Expansion

A case description usually contains crucial information which tailors a general medical condition to a individual situation. For instance, while the post title in Fig. 1 is about a patient's bad reaction to drug *Aricept*, the corresponding post body substantiates this medical condition with situation-specific symptoms, such as confusion, hallucination, etc. We incorporate such information by inferring the most important medical entities as follows:

Named Entity Recognition (NER). To extract entities we pre-process texts with StanfordCoreNLP⁵ for tokenization and part-of-speech tagging, and then run the OpenNLP Chunker⁶ to generate an initial set of noun phrase candidates. This set is extended by applying a small number of rules, like splitting or merging prepositional phrases, conjunctions, and proper/common nouns.

Named Entity Disambiguation (NED). To prune the set of candidates and link them to the KG, we use the algorithm proposed by [39]. This NED method is based on Locality Sensitive Hashing (LSH) with min-wise independent permutations for matching candidate phrases against entity names and their semantic types. We leverage type information from the KG to disambiguate between multiple entity candidates that match the same noun phrase. Whenever multiple entities have high matching scores, we pick the one with the more specific type. As the KG provides a ranked list of entities for each exactly-matching name (using information from UMLS), we further disambiguate by picking the highest ranked entity. If two entities share the same rank, the entity with the highest number of occurrences in different UMLS vocabularies is preferred.

Using the type hierarchy of the KG we prune out abstract entities of uninformative types such as *physical objects* or *concepts*, constraining the entity set to symptoms, diseases, medical findings, and pharmacological substances.

Category Expansion. For each entity in the expanded query, we retrieve its semantic categories from the KG. The categories do not only encode type information (e.g., the pharmacological class of a drug), but also relational facts harvested by KnowLife from large text corpora such as PubMed, Wikipedia articles, MayoClinic pages and more (e.g., the diseases for which a certain drug is prescribed). For instance, for *Alzheimer* we retrieve the categories *Mental or Behavioral Dysfunction* (a type category) and also *causes of memory impairment* (a fact category) among many others.

3.3 Query Focusing

Often, patient case descriptions are all but precise. They contain relevant information as well as peripheral or general information that digresses from the actual health issue. Therefore, the key focus is often buried under a substantial amount of secondary or irrelevant points. The expansion step alone cannot resolve this

⁵ <https://stanfordnlp.github.io/CoreNLP/>.

⁶ <https://opennlp.apache.org/>.

concern. This calls for a second step to re-focus the expanded query. In the second stage, we exploit the KG by considering the relationships between entities, this way enforcing:

- the coherence between entities in the query to counter topical drift, and
- the conciseness of the query itself by removing entities that are not in the core of the query intent.

This step does not only filter out irrelevant entities, but also produces a more comprehensive query. It discovers relevant semantic background knowledge for the patient’s medical condition by exploring neighboring entities related to entities mentioned in the case description. We model this task as a graph-algorithmic problem. First, the KG excerpt under consideration defines a *Query Graph* as follows:

Definition 1. A *Query Graph*, denoted by $QG = (V, E)$, is a directed graph with labeled vertices V and labeled edges E . V consists of the entities that appear in a patient’s question’s title and the full text of the corresponding post. E consists of the relational statements that exist between entities of V in the underlying knowledge graph.

Our goal is to extract the *most informative* and *focused* sub-graph from the QG . This resembles the task of graph summarization, where summaries take the form of dense subgraphs, aiming to represent the gist of the query. On one hand, such a graph should be as comprehensive as possible, but on the other hand we also need to factor in the varying degrees of informativeness of the included entities. To incorporate these two requirements, the ECO method maps the task into a *Prize Collecting Steiner Tree (PCST)* problem [20]. The PCST problem is a generalization of Steiner Trees, which considers both edge and node weights and relaxes the requirement that all terminal nodes are included in the resulting subgraph. Our method for computing Entity Cores is based on a PCST algorithm.

Definition 2. For a given query q posted in sub-forum S , let $QG = (V, E)$ be the Query Graph constructed from q , enhanced with node rewards $r(u)$ and edge costs $c(u, v)$ where u and v are nodes in QG as follows:

- $r(u) = \text{term-frequency}(u, S) / \# \text{sub-forums containing } u$
- $c(u, v) = 1 - \text{PMI}^2(e_u, e_v)$ where PMI^2 is the squared pointwise mutual information between two entities [37].

The **Entity Core (EC)** for this query is a connected subgraph $T' = (V', E')$ of QG that maximizes $f(T') = \sum_{v \in V'} r(v) - \sum_{u, v \in V'} c(u, v)$ and satisfies the condition that T' contains the node for sub-forum S .

An EC could be a general subgraph, but the nature of the objective function guarantees that only trees can be optimal. This is because including additional paths between already included nodes only increases the cost without improving the reward.

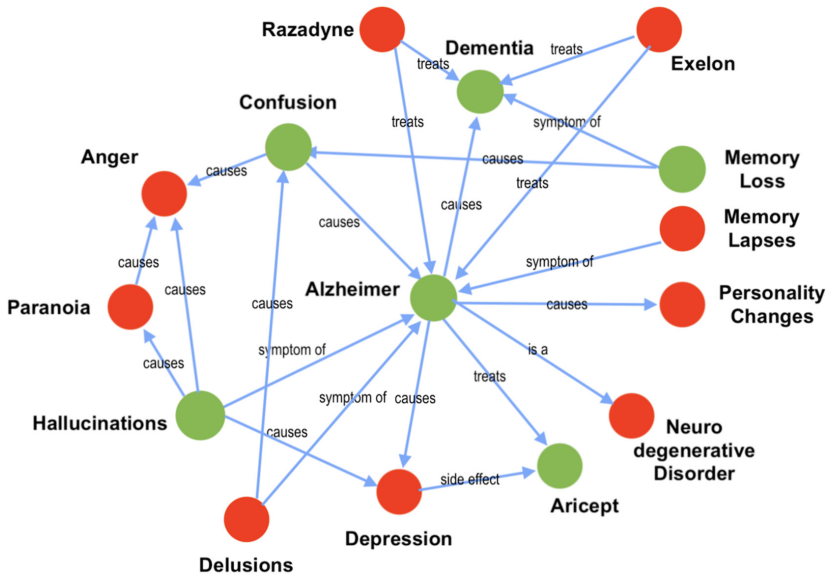


Fig. 4. Query Graph with Entity-Core Nodes Depicted in Green Color (Color figure online)

The output of ECO’s focusing step is the Entity Core (EC). The final query consists of three parts: all terms from the original user question, entities in the EC, and their semantic categories.

Approximation Algorithm. Not surprisingly, computing EC’s is NP-hard, but greedy methods are good approximations in many settings. We adopt the framework of [17], which consists of an iterative clustering method that groups nodes in the graph by merging existing clusters. More specifically, the algorithm can be divided in two stages: 1) growth and 2) pruning. During the growth stage, a set of active clusters and their respective spanning trees is maintained. The algorithm proceeds by iteratively merging or deactivating clusters until a desired number of active clusters is reached. In the pruning stage, unnecessary nodes are removed from the spanning trees of the last active clusters. This algorithm runs in nearly-linear time ($O(n \log^k n)$ with constant k) and has a factor-2 approximation guarantee.

As an illustration, Fig. 4 depicts a Query Graph for the example post in Fig. 1, with the computed Entity-Core nodes colored in green. As we can see, the EC reduces the overly broad Query Graph into a coherent and concise set of query terms.

3.4 Scoring and Ranking of Query Answers

Our scoring function for ranking query answers is based on a weighted linear combination of TF-IDF scores (tf = term frequency, idf = inverse document

frequency). Weights are derived from corpus statistics, and hyper-parameters are determined using a withheld validation set. A query Q is a triple $Q = (T, E, C)$, where T is the set of keywords, E is the set of extracted entities, and C is the set of semantic categories for E . Correspondingly, all documents D are divided into three fields (D_t, D_e, D_c) , i.e free-text, entities, and categories, which are indexed separately. With these representations we compute a ranking score s for a document and query as:

$$score(D, Q) = \lambda_T \sum_{t \in T} \frac{\mathbf{idf}(t)^2 * \mathbf{tf}(t, D_t)}{\sqrt{D_T}} + \lambda_C \sum_{c \in C} \frac{\mathbf{idf}(c)^2 * \mathbf{tf}(c, D_c)}{\sqrt{D_C}} + \lambda_E \sum_{e \in E} \frac{\mathbf{idf}(e)^2 * \mathbf{tf}(e, D_e)}{\sqrt{D_E}}$$

where $\sqrt{D_{\{T,E,C\}}}$ is a field-length normalization factor. Appropriate values for the hyper-parameters $\lambda_{\{T,E,C\}}$ are obtained by grid search (see Sect. 4.2).

Table 1. Health forum corpus

Source	Subforums	Users	Threads	Posts
healthboards.com	236	316,658	751,304	1,213,383
ehealthforum.com	285	338,079	297,356	4,251,533
patient.co.uk	800	18,326	44,618	151,583

4 Experimental Studies

To study the performance of the ECO method and compare it with state-of-the-art baselines, we conduct experiments with two different kinds of online contents: health forums and clinical trials. For the evaluation of the retrieved results, we use two kinds of assessments:

- relevance judgements by crowdsourcing workers
- authoritative judgements by two medical doctors, for a sub-set of the results.

The assessments by medical professionals primarily serve to validate the soundness of the crowdsourcing results. In addition, they reflect different perspectives for health forums where lay users and doctors may have different views on whether a reply is useful or not (not a point for clinical trials, though). Hence, we report on both settings separately.

4.1 Setup

Competitors. We compare a suite of strategies for generating queries:

- *Baseline - Title*: using all terms present in the post title, that is, in the user-question itself.
- *Baseline - Title+Post*: using all terms present in the title and full text of the post.
- *Baseline - Entity Expansion*: expanding the title-based query with biomedical entities from the full user post.
- *Baseline - Entity+Type Expansion*: expanding the title-based query with biomedical entities and their semantic types.
- *Baseline - Steiner Tree Expansion*: refining the title-based query with terms for all entities present in the Steiner Tree, computed over the Query Graph.
- *Entity Core Expansion (ECO)*: enriching the title-based query with all entities present in the Entity Core as in Definition 2.

Note that the Steiner-Tree-based expansion is not really a prior-works baseline, as it already makes use of our KG-based query graph construction. However, it is simpler than ECO, hence considered as another point of comparison.

We evaluated the rankings of query results by Top-k Precision (PRE@k), Top-k Mean Average Precision (MAP@k), and Normalized Discounted Cumulative Gain (NDCG@k).

Hyper-Parameters. We perform grid search to set the hyper-parameters $\lambda_{T,E,C}$ of the answering scoring of Sect. 3.4, using a small validation set of 10 withheld queries. The resulting values are $\lambda_T = 1.0$, $\lambda_E = 0.6$ and $\lambda_C = 0.1$.

Crowdsourcing Assessments. For gathering human judgements, we conducted crowdsourcing tasks over the [appen.com](https://www.appen.com) platform to assess the retrieval quality of the different query formulation strategies. To this end, crowd workers were asked to judge if a retrieved forum thread is relevant for a particular query along with the full user post (i.e., reflecting the individual health situation of the user whose post the query was derived from). The retrieved answer threads were presented to the worker as a combination of the root post that initiates the thread and the post with the highest word overlap measured by Jaccard similarity with the query post.

For quality assurance, we designed a set of test cases intermingled with the actual assessment tasks, and we cross-checked the workers' answers with our gold-standard results. The gold-standard set consisted of a set of 10 questions, that were evaluated by at least two experts that had perfect agreement between them for the relevance labels of the test cases. Poorly performing workers (2237 out of 14437) who failed to answer correctly the test question assigned to them were eliminated. We obtained 3 judgments for each query and paid 2.5 cent for each assessment. On average, trusted annotators needed 2.5 min to finish a task of 4 items. Overall, the inter-annotator agreement measured by Krippendorff's Alpha is 0.46.

Professional Assessments. To validate the relevance data obtained by crowdsourcing and to have authoritative judgements for evaluating our approach, we asked two professionals, both medical doctors, to label the results. A result is considered to be relevant, if both doctors label it as relevant. Due to the limited availability of the annotators, we reduce the evaluation set by randomly selecting 15 queries out of our 100 test queries.

4.2 Health Forums

Data. Our experimental data is obtained by crawling multiple health forums with a total of 1,048,428 discussion threads, from three main sources as given in Table 1. The forums are organized into sub-forums on more than 100 diseases, syndromes and drugs, from which we selected the following 20 topics with high coverage: Depression, Eating Disorders, Skin Cancer, Alzheimer Disease, Acid Reflux, Arthritis, Asthma, Back Pain, Carpal Tunnel Syndrome, Crohn Disease, Diabetes, Fibromyalgia, High Blood Pressure and Hypertension, Insect Bites, Low Blood Pressure and Hypotension, Meningitis, Multiple Sclerosis, Pancreas Disorders, Sinusitis, Vision and Eye Disorders. For each of these disorders, we identified 5 typical user posts which serve as queries in our experiments. As such, our test workload consists of 100 individual queries. To make our experiments transparent and reproducible for third parties, we release the datasets on both the forum-search and clinical-trials experiments at: <http://eco.mpi-inf.mpg.de/>.

Expansion Strategies. Table 2 compares the Title and Title+Post baselines with straightforward expansion strategies using Entities and Entities+Types. The first observation is that a simple expansion using Title+Post baseline degrades the results compared to Title only. This illustrates the difficulty of generating queries that capture specific user’s situation yet stay focused and concise. Entity expansion outperforms the baseline in all evaluation metrics, where all results have statistical significance (with a p-value ≤ 0.01 for a paired t-test). Expansion with types and categories (in addition to entities), is significantly better than the baseline (with p-value ≤ 0.01) but does not improve over expanding merely with entities. Overall, we conclude that incorporating entities from the full text of user posts is crucial and significantly improves search result quality. The additional incorporation of entity types does not give notable benefits.

Focusing Strategies. For comparing ECO against the expansion-only strategies, we focus on the best-performing baselines using solely entities. Table 3 compares the results of two focusing strategies against the best expansion-only strategy *Entity Expansion*. The table clearly shows that focusing with Steiner Trees cannot improve the results, and actually loses against expansion-only by all metrics (with p-value <0.01). In contrast, the ECO method yields additional benefits in retrieval performance with significant gains. This underlines the need for judiciously re-focusing the expanded query, where Entity Cores turn out to

Table 2. Crowdsourcing evaluation of expansion baselines for health forums.

Approach	PRE		MAP		NDCG	
	@5	@10	@5	@10	@5	@10
Title	0.59	0.59	0.72	0.67	0.80	0.80
Title + Post	0.49	0.46	0.72	0.67	0.8	0.79
Entity expansion	0.68	0.67	0.8	0.76	0.86	0.87
Entity + Type expansion	0.66	0.64	0.79	0.75	0.86	0.87

Table 3. Crowdsourcing evaluation of focusing strategies for health forums.

Approach	PRE		MAP		NDCG	
	@5	@10	@5	@10	@5	@10
Entity expansion	0.68	0.67	0.8	0.76	0.86	0.87
Ex + ST	0.59	0.58	0.73	0.69	0.8	0.81
ECO	0.75	0.74	0.81	0.79	0.87	0.88

be much better than Steiner Trees. The superiority of ECO is confirmed also in the evaluation by medical professionals, as shown in Table 4. Here, too, ECO significantly outperforms both baselines.

Retrieval Time. Generating and executing focused expanded queries with ECO takes 1 to 5 s ($\mu = 1.86$, $\sigma = 0.97$). The analysis of posts for detecting entities takes 5 to 60 s, with high variance, as it is approximately proportional to the post length. Both entity markup and query processing could be sped up by more engineering.

4.3 Clinical Trials

To demonstrate the versatility of our approach, we also test its applicability for clinical trials, where doctors would search on behalf of a patient. Our experimental data consists of 97,390 clinical trials from clinicaltrials.gov. We evaluate ECO on the 15 randomly selected queries.

The crowdsourcing results in Table 5 demonstrate that ECO is able to achieve large performance gains across all metrics compared to the previously best baseline. The inter-annotator agreement between crowd workers is 0.49, far from perfect but remarkably high.

Since clinical trials are difficult to interpret for lay users, we also evaluate the results with judgements by medical professionals. Even though the doctors' assessments of relevance and utility tend to be more conservative than the crowdsourcing judgements, the ECO method significantly outperforms the baselines for all metrics (with $p\text{-value} < 0.01$) as shown in Table 6. This shows that the judiciously re-focused use of KG entities and categories does successfully bridge the

Table 4. Medical doctor evaluation for health forums.

Approach	PRE		MAP		NDCG	
	@1	@5	@1	@5	@1	@5
Title	0.33	0.28	0.33	0.49	0.33	0.56
Entity expansion	0.33	0.31	0.33	0.54	0.33	0.63
ECO	0.40	0.39	0.40	0.62	0.40	0.69

Table 5. Crowdsourcing evaluation for clinical trials

Approach	PRE		MAP		NDCG	
	@1	@5	@1	@5	@1	@5
Title	0.67	0.63	0.67	0.79	0.67	0.86
Entity expansion	0.87	0.64	0.87	0.82	0.87	0.90
ECO	0.93	0.87	0.93	0.91	0.93	0.95

terminologies of users (in post titles as queries) and medical experts (in result documents). Altogether, this confirms that ECO is able to achieve large gains over the baselines also under the meticulous examination by professionals.

Table 6. Evaluation by medical professionals for clinical trials

Approach	PRE		MAP		NDCG	
	@1	@5	@1	@5	@1	@5
Title	0.40	0.40	0.40	0.50	0.40	0.53
Entity expansion	0.40	0.41	0.40	0.55	0.40	0.60
ECO	0.60	0.69	0.60	0.74	0.60	0.81

5 Conclusion

This work addressed the under-explored topic of supporting patient-centric information needs by search over health contents. Our experiments, with evaluation by both crowdsourcing users and medical professionals, demonstrated the viability of our ECO method. In comparison to state-of-the-art baselines with query expansion by entities and classes from a KG, the experimental results clearly showed that the re-focusing step, based on entity cores, is crucial for the superior performance of ECO.

We focused on two kinds of health contents: forums of online communities and reports on clinical trials, as these are the best sources on patient experiences. Nevertheless, we plan to explore the suitability of our approach for other kinds of health documents, such as PubMed articles or health news.

In this work, we have focused on searching health forums and clinical trials, as health is by itself an important domain with high impact. In general, our methodology can be adapted to other domains as well, such as finance, food or travel. For example, we could address a travel discussion forum where people ask about and exchange experiences about visa issues, sightseeing beaten paths, local food and culture, etc. These are subject to individual preferences, so that ranking answers by personal relevance is important. We would use travel-centric KGs, and use ECO for focused query expansion. This is left for future work.

References

1. Abrahamson, J.A., Fisher, K.E., Turner, A.G., Durrance, J.C., Turner, T.C.: Lay information mediary behavior uncovered: exploring how nonprofessionals seek health information for themselves and others online. *J. Med. Library Assoc. JMLA* **96**(4), 310 (2008)
2. Alsentzer, E., et al.: Publicly available clinical bert embeddings. arXiv preprint [arXiv:1904.03323](https://arxiv.org/abs/1904.03323) (2019)
3. Balaneshinkordan, S., Kotov, A.: An empirical comparison of term association and knowledge graphs for query expansion. In: Ferro, N., et al. (eds.) *ECIR 2016*. LNCS, vol. 9626, pp. 761–767. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1_65
4. Balog, K.: *Entity-Oriented Search*. Springer Nature, Cham (2018). <https://doi.org/10.1007/978-3-319-93935-3>
5. Barros, J.M., Buitelaar, P., Duggan, J., Rebholz-Schuhmann, D.: Unsupervised classification of health content on reddit. In: *Proceedings of the 9th International Conference on Digital Public Health*, pp. 85–89 (2019)
6. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. *ACM Comput. Surv. (CSUR)* **44**(1), 1–50 (2012)
7. Chamberlin, S.R., et al.: A query taxonomy describes performance of patient-level retrieval from electronic health record data. medRxiv, p. 19012294 (2019)
8. Dalton, J., Dietz, L., Allan, J.: Entity query feature expansion using knowledge base links. In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 365–374 (2014)
9. De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., Bruza, P.: Medical semantic similarity with a neural language model. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 1819–1822 (2014)
10. Dirkson, A., Verberne, S., Kraaij, W.: Narrative detection in online patient communities. In: *Texts@ECIR*, pp. 21–28 (2019)
11. Dragoni, M.: Semantic ai for healthcare: The horus. ai platform. In: *Second International Workshop on Semantic Web Meets Health Data Management (SWH 2019) co-located with the 18th International Semantic Web Conference (ISWC 2019)*. vol. 2515, pp. 1–4. CEUR-WS. org (2019)
12. Ernst, P., et al.: DeepLife: an entity-aware search, analytics and exploration platform for health and life sciences. In: *ACL*, pp. 19–24 (2016)
13. Ernst, P., Siu, A., Weikum, G.: Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinform.* **16**(1), 157 (2015)

14. Ernst, P., Terolli, E., Weikum, G.: LongLife: a platform for personalized search for health and life sciences. In: 18th Semantic Web Conference, pp. 237–240. ceur-ws.org (2019)
15. Fang, H., Zhai, C.: Semantic term matching in axiomatic approaches to information retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 115–122 (2006)
16. Hazimeh, H., Zhai, C.: Axiomatic analysis of smoothing methods in language models for pseudo-relevance feedback. In: ICTIR, pp. 141–150. ACM (2015)
17. Hegde, C., Indyk, P., Schmidt, L.: A nearly-linear time framework for graph-structured sparsity. In: ICML (2015)
18. Jimmy, Zuccon, G., Palotti, J.R.M., Goeuriot, L., Kelly, L.: Overview of the CLEF 2018 consumer health search task. In: Working Notes of CLEF (2018)
19. Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X.: PubMedQA: a dataset for biomedical research question answering. arXiv preprint [arXiv:1909.06146](https://arxiv.org/abs/1909.06146) (2019)
20. Johnson, D.S., Minkoff, M., Phillips, S.: The prize collecting steiner tree problem: theory and practice. In: SODA, pp. 760–769 (2000)
21. Kanthawala, S., Vermeesch, A., Given, B., Huh, J.: Answers to health questions: internet search results versus online health community responses. *J. Med. Internet Res.* **18**(4), e95 (2016)
22. Khanpour, H., Caragea, C.: Fine-grained information identification in health related posts. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1001–1004 (2018)
23. Kondylakis, H., et al.: Semantically-enabled personal medical information recommender. In: ISWC (2015)
24. Koopman, B., Zuccon, G.: WSDM 2019 tutorial on health search (HS2019): a full-day from consumers to clinicians. In: WSDM, pp. 838–839 (2019)
25. Koopman, B., Zuccon, G., Bruza, P.: What makes an effective clinical query and querier? *JASIST* **68**(11), 2557–2571 (2017)
26. Krithara, A., et al.: iASiS: towards heterogeneous big data analysis for personalized medicine. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), pp. 106–111. IEEE (2019)
27. Kuzi, S., Carmel, D., Libov, A., Raviv, A.: Query expansion for email search. In: SIGIR, pp. 849–852. ACM (2017)
28. Kuzi, S., Shtok, A., Kurland, O.: Query expansion using word embeddings. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, pp. 1929–1932 (2016)
29. Lee, J., et al.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4), 1234–1240 (2020)
30. Liu, X., Chen, F., Fang, H., Wang, M.: Exploiting entity relationship for query expansion in enterprise search. *Inf. Retrieval* **17**(3), 265–294 (2014)
31. Luo, G., Tang, C.: On iterative intelligent medical search. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3–10 (2008)
32. Luo, G., Tang, C., Yang, H., Wei, X.: MedSearch: a specialized search engine for medical information retrieval. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 143–152 (2008)
33. Mukherjee, S., Weikum, G., Danescu-Niculescu-Mizil, C.: People on drugs: credibility of user statements in health communities. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 65–74 (2014)

34. Pang, P.C.I., Verspoor, K., Pearce, J., Chang, S.: Better health explorer: designing for health information seekers. In: OzCHI, pp. 588–597. ACM (2015)
35. Patel, C., et al.: Matching patient records to clinical trials using ontologies. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 816–829. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_59
36. Roberts, K., et al.: Overview of the trec 2017 precision medicine track. In: TREC (2017)
37. Role, F., Nadif, M.: Handling the impact of low frequency events on co-occurrence based measures of word similarity. In: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR-2011). Scitepress, pp. 218–223 (2011)
38. Rospoche, M., Corcoglioniti, F., Dragoni, M.: Boosting document retrieval with knowledge extraction and linked data. *Semantic Web* **10**(4), 753–778 (2019)
39. Siu, A., Nguyen, D.B., Weikum, G.: Fast entity recognition in biomedical text. In: Proceedings of Workshop on Data Mining for Healthcare (DMH) at Conference on Knowledge Discovery and Data Mining (KDD). ACM Press, New York (2013)
40. Soldaini, L., Yates, A., Goharian, N.: Learning to reformulate long queries for clinical decision support. *JAIST* **68**(11), 2602–2619 (2017)
41. Soto, A.J., Przybyla, P., Ananiadou, S.: Thalia: semantic search engine for biomedical abstracts. *Bioinformatics* **35**(10), 1799–1801 (2019)
42. Suominen, H., et al.: Overview of the CLEF eHealth evaluation lab 2018. In: Bellot, P., et al. (eds.) CLEF 2018. LNCS, pp. 286–301. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_26
43. White, R.W., Horvitz, E.: From health search to healthcare: explorations of intention and utilization via query logs and user surveys. *JAMIA* **21**(1), 49–55 (2013)
44. Wu, H., et al.: SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J. Am. Med. Inform. Assoc.* **25**(5), 530–537 (2018)
45. Zhu, D., Wu, S., Carterette, B., Liu, H.: Using large clinical corpora for query expansion in text-based cohort identification. *J. Biomed. Inform.* **49**, 275–281 (2014)
46. Zuccon, G., Koopman, B., et al.: Payoffs and pitfalls in using knowledge-bases for consumer health search. *Inf. Retrieval J.* **22**(3–4), 350–394 (2019)