Inf Retrieval J (2016) 19:189–224
DOI 10.1007/s10791-015-9269-8

MEDICAL INFORMATION RETRIEVAL

How users search and what they search for in the medical domain

Understanding laypeople and experts through query logs

João Palotti[1] · Allan Hanbury[1] · Henning Müller[2] · Charles E. Kahn Jr.[3]

Received: 31 December 2014 / Accepted: 19 September 2015 / Published online: 24 October 2015

Abstract The internet is an important source of medical knowledge for everyone, from laypeople to medical professionals. We investigate how these two extremes, in terms of user groups, have distinct needs and exhibit significantly different search behaviour. We make use of query logs in order to study various aspects of these two kinds of users. The logs from America Online, Health on the Net, Turning Research Into Practice and American Roentgen Ray Society (ARRS) GoldMiner were divided into three sets: (1) laypeople, (2) medical professionals (such as physicians or nurses) searching for health content and (3) users not seeking health advice. Several analyses are made focusing on discovering how users search and what they are most interested in. One possible outcome of our analysis is a classifier to infer user expertise, which was built. We show the results and analyse the feature set used to infer expertise. We conclude that medical experts are more persistent, interacting more with the search engine. Also, our study reveals that, conversely to what is stated in much of the literature, the main focus of users, both laypeople and professionals, is on disease rather than symptoms. The results of this article, especially through the classifier built, could be used to detect specific user groups and then adapt search results to the user group.

Keywords Query log analysis · Health search · User behavior

✉ João Palotti
palotti@ifs.tuwien.ac.at

Allan Hanbury
hanbury@ifs.tuwien.ac.at

Henning Müller
henning.mueller@hevs.ch

Charles E. Kahn Jr.
charles.kahn@uphs.upenn.edu

[1] Vienna University of Technology, Vienna, Austria

[2] University of Applied Sciences and Arts Western Switzerland (HES-SO), Delémont, Switzerland

[3] University of Pennsylvania, Philadelphia, PA, USA

1 Introduction

Among all topics available on the internet, medicine is one of the most important in terms of impact on the user and one of the most frequently searched. A recent report states that one in three American adult Internet users have sought out health advice online to diagnose a medical condition (Fox and Duggan 2013). This tendency is the same in Europe, where a recent report from the European Commission estimates that 60 % of the population have used the Internet to search for health-related information in 2014 (Eurobarometer 2014), with numbers even higher in several member states. Both reports show that the most common tasks performed are either searching for general information on health-related topics, such as diet, pregnancy and exercise, or searching for information on specific injuries or diseases. They also found that mostly the search starts in a search engine and young users are more likely to search for this kind of information.

Physicians are also very active Internet users (Kritz et al. 2013). PubMed which indexes the biomedical literature reports more than one hundred million users (Dogan et al. 2009), where two-thirds are experts (Lacroix and Mehnert 2002). Nevertheless, studies on how experts search on the Internet for medical content are relatively rare (Younger 2010).

We divide the users of medical search engines into *laypeople* and *experts*, where laypeople are considered to be searchers that do not have a deep knowledge about the medical topic being searched, and experts do have a deep knowledge about the medical topic being searched. Our assumption is that laypeople wish to see more introductory material returned as search results, whereas experts wish to see detailed scientific material returned as search results. At first glance, this could easily be interpreted as a division into patients and medically-trained professionals. Nevertheless, it often occurs that a patient or patient's relatives become experts on a disease or condition affecting themselves or a family member, sometimes becoming more knowledgeable in a narrow domain than medically-trained professionals. There is also the case of a medical professional searching in a medical topic outside of his/her main expertise (e.g., a cardiac surgeon looking for information on a skin disease), where the information need may be initially satisfied by less scientific documents, although likely not very basic documents due to the medical background. For these reasons, we specifically avoid defining medical professional and patient classes.

Distinguishing laypeople and experts can significantly improve their interactions with the search engine (White et al. 2009; Palotti et al. 2014a). Currently, users may get different results for their queries if they are in different locations, but not if they have different levels of expertise. We make the assumption that it is possible to distinguish the level of expertise of the searcher based on the vocabulary used and the search style. While it would be realistic to represent a continuum of expertise levels, we define two classes (laypeople and experts) in this study, allowing us to investigate the most relevant differences between the classes.

Recently, many studies showed successful cases of exploring the user's expertise, in particular for general search engines (White et al. 2009; Schwarz and Morris 2011; Collins-Thompson et al. 2011). Schwarz and Morris (2011) show that the popularity of a webpage among experts is a crucial feature to help laypeople identify credible websites. Collins-Thompson et al. (2011) discuss that re-ranking general search engine results to match the user's skills of readability can provide significant gains, however estimating user profiles is a non-trivial task and needs to be further explored.

This study investigates how users search for medical content, building profiles for experts and laypeople. Understanding the needs of these two distinct groups is important for designing search engines, whether it is used for boosting easy-to-read documents or for suggesting queries to match the search expertise. Additionally, whenever it is possible, we also compare search for medical information with regular search for other topics.

This work is conducted through the analysis of user interactions logged by search engines. Log analysis is unobtrusive and captures the user behaviour in a natural setting (Jansen et al. 2008). We used Metamap,[1] which is the state-of-the-art tool to recognise and map biomedical text to its corresponding medical concepts, to provide a richer set of information for each query. Little is known in the literature on how to identify medical concepts in short Web queries, therefore we also evaluated Metamap for this task.

In particular, this work addresses the following questions:

1. How suitable is MetaMap for analysing short queries?
2. Which characteristics allow laypeople and experts to be distinguished based on
   (a) How they search in medical content?
   (b) What they search for in medical content?
3. To what extent do these characteristics match or disagree with those identified in other published studies?
4. What are the most useful features to automatically infer user expertise through the query logs?

In our analysis, we use health related queries from the America Online (AOL) query log, as well as the Health on the Net (HON) search engine log to represent the logs generated to a significant extent by laypeople. Medical professionals also use general search engines to seek health content, however their queries are drowned in the laypeople queries. White et al. (2009), for example, hypothesise that search leading to PubMed was made by experts. Using this hypothesis, only 0.004 % of the whole AOL log was issued by medical professionals (also referred to as experts).

Besides the fact that PubMed is more frequently used in a research environment rather than in a clinical environment (Kritz et al. 2013), it is also frequently visited by laypeople (Lacroix and Mehnert 2002). Therefore, we use the logs from the evidence-based search engine TRIP Database and the radiology image search engine American Roentgen Ray Society (ARRS) GoldMiner to represent queries entered by physicians usually when facing a practical problem.

Several analyses are presented: from general statistics of the logs to complex inference on what is the search focus in each individual search session. We contrast our results with others from the literature and provide our interpretation for each phenomenon found.

The remainder of this paper is organised as follows. Section 2 presents a literature review and positions our work with respect to other articles in the literature. In Sect. 3, we describe the datasets used and the preprocessing steps applied. In Sect. 4 we present and evaluate MetaMap, the tool used to enrich the information contained in the query logs. We start our analysis in Sect. 5, where we examine the general user behaviour and the most popular queries, terms and topics searched. In Sect. 6, we introduce the concept of search session into our analysis. In Sect. 7, we present a Random Forest classifier to infer user

[1] http://metamap.nlm.nih.gov/.

expertise and analyse the feature's importance. Section 8 presents our findings and limitations. Finally, conclusions and future work are presented in Sect. 9.

2 Related work

As soon as modern search engines appeared, the first studies on query logs started (Jansen et al. 1998; Silverstein et al. 1999). Jansen et al. (1998) and Silverstein et al. (1999) analysed the logs from Excite and Altavista respectively, popular search engines at that time. Both articles point out some important results such as the fact that the vast majority of users issue only one single query and rarely access any result page beyond the first one.

The most recent general search engine to disclose query logs to researchers was America Online (AOL) in 2006 (Pass et al. 2006). The AOL data were afterwards used in various studies, such as Brenes and Gayo-Avello (2009), which provides methods to group users and their intents, and Duarte Torres et al. (2010), who analyse queries targeting children's content. In this work, we compare the analysis made in the literature for general search engines (Jansen et al. 1998; Silverstein et al. 1999) with medical domain search engines, and we adopt a method similar to Duarte Torres et al. (2010) to divide the AOL logs into queries related or not to health. It is important to mention that the AOL log had known privacy problems in the past, resulting in some users being identified even though the logs were supposedly anonymised. Despite this problem, we opt to use this dataset for several reasons. One reason is that it can be freely downloaded, as well as the code used for all the experiments of this paper, making the experiments reproducible.[2] Another reason is that studies of how medical annotation tools such as MetaMap perform in the wild are not well known. Finally, in the absence of a more recent large search engine query log we consider that the AOL logs are still the best choice for researchers in academia. A complete reference of the previous 20 years of research on log analysis and its applications is well described by Silvestri (2010).

There are a number of studies analysing query logs in the medical domain. We highlight here some important work for this research, including work based on general search engines (Spink et al. 2004; White et al. 2009; Cartright et al. 2011; White and Horvitz 2012), as well as specialised ones (Herskovic et al. 2007; Dogan et al. 2009; Tsikrika and Müller 2012; Meats et al. 2007; Zhang 2014). We also highlight some important work on user expertise and behaviour. Figure 1 depicts each one of these areas, including relevant work on general search for non-health-related content. For a matter of organisation, we divide the rest of this section into three parts, one for each topic. As shown in Fig. 1, some papers may be relevant to more than one topic.

2.1 General search engines

We describe here studies on health-related query logs in general search engines, starting with Spink et al. (2004), who studied medical queries issued in 2001 in Excite and AlltheWeb.com. They showed that medical web search was decreasing since 1999, suggesting that users were gradually shifting from general-purpose search engines to specialised sites for health-related queries. Also, they found that health-related queries were equivalent in length, complexity and lack of reformulation to general web searching.

[2] https://github.com/joaopalotti/logAnalysisJournal.

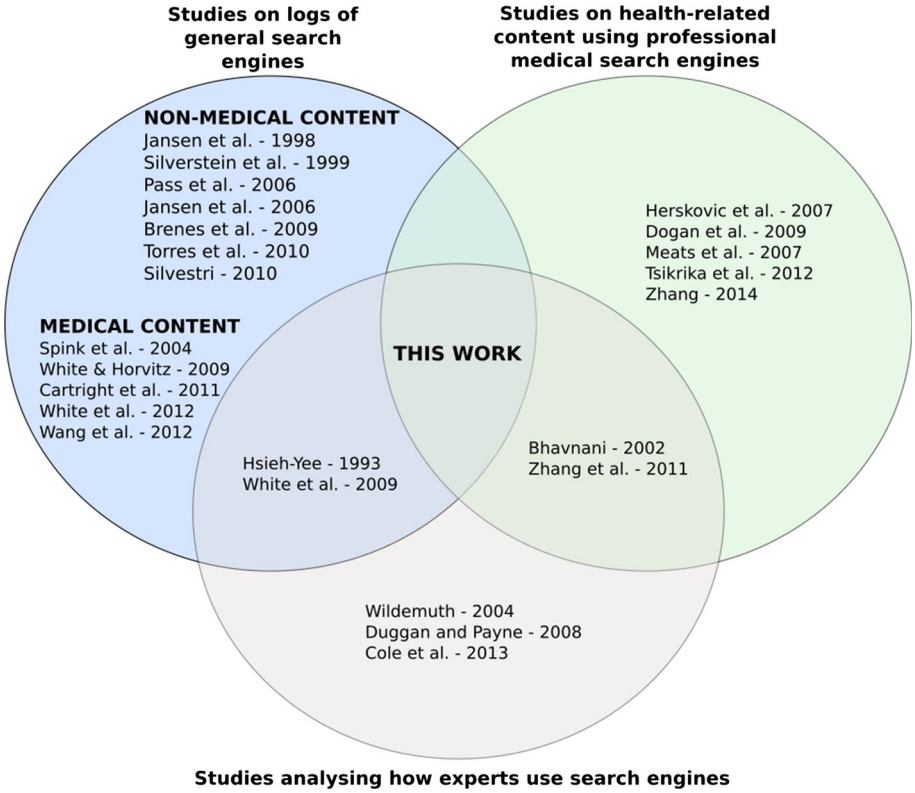

Fig. 1 Our work studies both general and specialised search engines and investigates how users with different expertise levels search for health content

More recently, White and Horvitz (2009), White and Horvitz (2012) studied how users start looking for a simple symptom and end up searching for a serious disease, a phenomenon they named cyberchondria. They used the logs of the Windows Live Toolbar to obtain their data and list of keywords to annotate symptoms and diseases in queries, while we used the US National Library of Medicine MetaMap to do the same. Similar to our work, they define user sessions as a series of queries followed by a period of user inactivity of more than 30 min and they made use of the Open Directory Project (ODP) hierarchy to identify medical sessions.

Another important work is Cartright et al. (2011). The authors presented a log-based study of user behaviour when searching for health information online. The authors classified user queries into three classes: symptoms, causes and remedy. They analysed the change of search focus along a session, and showed that it is possible to build a classifier to predict what is the next focus of a user in a session. We decided to use the same classes in order to make our study comparable, however we used the semantic annotator of MetaMap instead of hand coded rules.

Not studying the query logs, but the ranking lists of major general search engines, Wang et al. (2012) compared the results of Google, Yahoo!, Bing, and Ask.com for one single query *breast cancer*. Among their conclusions is the fact that results provided rich information and highly overlapped between the search engines. The overlap between any

two search engines was about half or more. Another work that compares a large number of search engines is Jansen and Spink (2006), in which nine search engines with logs varying from 1997 to 2002 were used. Nevertheless, the latter did not focus on medical queries.

2.2 Expertise and search engines

One of the first studies to report how expertise influences the process of search dates from the 1990s. In this work, Hsieh-Yee (1993) reported that experienced library science students could use more thesauri, synonymous terms, combinations of search terms and spend less time monitoring their searches than novices. Later, Bhavnani (2002) studied search expertise in the medical and shopping domains. He reported that experts in a topic can easily solve the task given even without using a search engine, because they already knew which website was better adapted to fill their needs. Bhavnani also reported that experts started their search by using websites such as MedlinePlus,[3] instead of a major search engine, while laypeople started with Google.

White et al. (2009) showed a log-based analysis of expertise in four different domains (medicine, finance, law, and computer science), developing an expertise classifier based on their analysis. Apart from showing that it is possible to predict user expertise based on their behaviour, they showed that experts have a higher success rate only in their domain of expertise, with success in a session being defined as a clicked URL as the final event in a session. Therefore, an expert in finance would have a comparable or worse success rate in medicine than a non-expert. An important difference between our work and White's work is the approach used to separate experts from non-experts. They assume that search leading to PubMed was made by medical experts and search leading to ACM Digital library (ACM-DL)[4] was made by computer science experts. In the medical domain this is a weak premise for two reasons: (1) it is estimated that one-third of PubMed users are laypeople (Lacroix and Mehnert 2002), (2) PubMed is more important for medical researchers than practitioners (Kritz et al. 2013). Tracing a parallel between medicine and computer science, a general practitioner would be like a software developer that does not necessarily need to consult the ACM-DL (the correspondent for PubMed) to perform his/her work. One could manually expand the list of expert sites to include, for example, StackOverflow[5] or an API website for experts in CS and treatment guidelines or drug information sites for medicine but it would be a laborious task and unstable over time. Hence, to cope with this challenge, we use the logs of different search engines made for distinct audiences.

An important user study was conducted by Wildemuth (2004). He evaluated how the search tactics of microbiology students changed over an academic year, while the students' topic knowledge was increasing. The students were asked questions about the topic at three different times: before starting the course, when finishing the course, and 6 months after the course. As their expertise increased, the users were able to perform a better term selection for search, being more effective. The most common pattern used across all three occasions was the narrowing of the retrieved result set through the addition of search concepts, while at the beginning users were less effective in the selection of concepts to include in the search and more errors were made in the reformulation of a query. Later, Duggan and Payne (2008) explored the domains of music and football to evaluate how the

[3] MedlinePlus is a web-based consumer health information system developed by the American National Library of Medicine (NLM): http://www.medlineplus.gov/.

[4] http://dl.acm.org/.

[5] http://stackoverflow.com/.

user knowledge of a topic can influence the probability of a user answering factual questions, finding that experts detect unfruitful search paths faster than non-experts.

Recently, there have been a few user studies in user expertise prediction. For example, Zhang et al. (2011) and Cole et al. (2013) are based on TREC Genomics data. The former employed a regression model to match user self-rated expertise and high level user behaviour features such as the mean time analysing a document and the number of documents viewed. They found that the user's domain knowledge could be indicated by the number of documents saved, the user's average query length, and the average rank position of opened documents. Their model, however, needs to be further investigated because the data was limited, collected in a controlled experiment, and from only one domain. Similarly, but using only eye movement patterns as features, the latter conducted a user study instead of log analysis and employed a linear model and random forests to infer the user expertise level. Their main contribution is demonstrating that models to infer a user's level of domain knowledge without processing the content of queries or documents is possible, however they only performed one single experiment and in one single domain.

2.3 Medical-specialised search engines

For specialised medical search engines, Herskovic et al. (2007) analysed an arbitrary day in PubMed, the largest biomedical database in the world. They concluded that PubMed may have a different usage profile than general web search engines. Their work showed that PubMed queries had a median of three terms, one more than what is reported for Excite and Altavista. Subsequently, Dogan et al. (2009) studied an entire month of PubMed log data. Their main finding comparing PubMed and general search engines was that PubMed users are less likely to select results when the result sets increase in size, users are more likely to reformulate queries and are more persistent in seeking information. Whenever possible, our analysis is compared with the statements made for PubMed.

Meats et al. (2007) conducted an analysis on the 2004 and 2005 logs of the TRIP Database, together with a usability study with nine users. Their work concluded that most users used a single term and only 12 % of the search sessions utilised a Boolean operator, underutilising the search engine features. Tsikrika and Müller (2012) examined query logs from ARRS GoldMiner, a professional search engine for radiology images. They studied the process of query modification during a user session, aiming to guide the creation of realistic search tasks for the ImageCLEFmed benchmark. Meats used 620,000 queries and Tsikrika only 25,000, while we use nearly three and nine times more queries from TRIP and GoldMiner, respectively, allowing us to perform a deeper analysis.

Zhang (2014) analysed how 19 students solved 12 tasks using MedlinePlus. The tasks were created based on questions from the health section of Yahoo! Answers. Although the log analysis made is very limited due to the artificial scenario created and the small number of users, Zhang could investigate browsing strategies used by users (amount of time searching and/or browsing MedlinePlus) and the users' experience with MedlinePlus (usability, usefulness of the content, interface design) through questionnaires and recording the users performing the tasks. Our study is limited to only the query logs, however a large analysis is made for different websites and the user behaviour is captured in a very natural setting.

2.4 This work

As illustrated in Fig. 1, this work closes a gap. It studies both general and specialised search engines, as well as taking into consideration different user expertise levels. Throughout the rest of this work, we compare our methodology and results with the studies cited in this section.

3 Datasets and pre-processing steps

In this section, we describe the datasets used in this study and the preprocessing steps applied to them.

3.1 Sorting the data by expertise level

We make the assumption that experts and laypeople are more likely to use different search engines to satisfy their information needs. Therefore we assume that almost all queries entered into a particular search engine are entered by only one of the two classes of users under consideration. This assumption is justified as we are using search logs from search engines clearly aimed at users of specific expertise. This assumption is also more inclusive than another assumption that has been used to separate medical experts from laypeople: that only searches leading to PubMed were made by medical experts White et al. (2009). As discussed in Palotti et al. (2014a), this assumption would only tend to detect medical researchers, as medical practitioners make less use of PubMed (Kritz et al. 2013). We do not take into account that many users are in between laypeople and experts as levels can vary.

On one extreme, we have AOL laypeople users. There might be a few medical experts using AOL, but their queries are drowned in the laypeople queries. Also focused on patients, HON is a search engine for laypeople searching for reliable health information. The main target audience is laypeople concerned about the reliability of the information they access. On the other extreme, mainly targeting physicians looking for medical evidence, the TRIP database can also be accessed by patients but these few patients might be already considered specialists on their diseases. Finally, the GoldMiner search engine is made by radiologists and for radiologists, patients have practically no use for this kind of information, but a variety of physicians might access the system. We position each dataset on an expertise axis in Fig. 2, to help understanding how each dataset relates to each other.

Fig. 2 The datasets used here are plotted on an expertise scale. The expertise level increases as a dataset is placed more to the *right-hand side* of the scale

3.2 Data

Four query logs from search engines taking free text queries were divided into five datasets in our analysis: two focused on laypeople queries, two made up of queries from medical professionals and one consisted of queries not related to health or medical information.

The query logs that are assumed to consist almost completely of queries submitted by laypeople were obtained from medical-related search in America Online's search service (Pass et al. 2006)[6] and from the Health on the Net Foundation website (HON[7]).

The AOL logs were obtained from March to May 2006. We divided them into two non-overlapping sets: **AOL-Medical** and **AOL-NotMedical**. For this purpose, the click-through information available in the AOL data was used. A common approach to decide what the topic of a URL is, is checking if it is listed in the Open Directory Project (ODP)[8] (Cartright et al. 2011; Collins-Thompson et al. 2011; Duarte Torres et al. 2010; White et al. 2009; White and Horvitz 2012). For the clicked URLs that are not present in ODP, some researchers use supervised learning to automatically classify them (Collins-Thompson et al. 2011; White et al. 2009; White and Horvitz 2012). However, it is very important to note that this approach cannot be used here, as 47 % of the AOL log entries lack the clicked URL information.

Alternative approaches can be designed. One is to keep only queries in which the clicked URL is found in ODP, excluding all the rest. Although valid, this approach results in removing 73 % of all queries, as only 27 % of the queries had a clicked URL found in ODP. This has a strong impact in the behaviour analysis, such as a strong reduction in the number of queries per session. Another possibility is doing as in Cartright's work (Cartright et al. 2011), in which a list of symptoms was used to filter sessions on health information. However this approach creates a strong bias when analysing what users are searching for, as it certainly results in a dataset in which everyone searches for symptoms.

Our solution is based on user sessions—this approach is not as restricted as when analysing single queries and does not suffer from the bias of filtering by keywords. First we divide the query log into user sessions, continuous queries from the same user followed by an inactivity period exceeding 30 min. After this, we attribute one of the following labels for each clicked URL, if any: (1) *Medical*, (2) *Not Medical*, or (3) *Not Found*. This depends on whether the URL is (1) found in any Medical category listed in Table 1; (2) found in any other category: News, Arts, Games, Health/Animals, Health/Beauty, etc; or (3) not found in either of these. Last, we assign to the whole session the Medical label only if the proportion of URLs on Medical information is greater than a threshold $\mathbf{t}$. Medical search sessions classified this way are attributed to the set **AOL-Medical**, while the rest goes to the **AOL-NotMedical** set. Figure 3 illustrates the session assignment procedure. For the experiments performed in this work, we use $\mathbf{t} = \mathbf{0.5}$. This value is a fair trade-off between two extremes: considering an entire session as being on medical information because one single URL on medical information was clicked (see second part of Fig. 3), and considering an entire session as being on medical information only if all the known clicked links are on medical information (see the first part of Fig. 3).

For the first part of Fig. 3, it is important to note that the first query could be considered to belong to another session, as the user intent might be different from the rest of the session. The second and third queries, drug names that are clearly for medical content,

[6] Obtained from http://www.gregsadetsky.com/aol-data/.

[7] http://www.hon.ch/HONsearch/Patients/index.html.

[8] http://www.dmoz.org/.

Table 1 ODP categories used to filter the AOL-Medical

ODP category	URL examples
\Top\Health\Medicine	http://www.nlm.nih.gov
	http://www.webmd.gov
\Top\Health\Alternative	http://www.acupuncturetoday.com
	http://www.homeopathyhome.com
\Top\Health\Dentistry	http://www.dental-health.com
	http://www.animated-teeth.com
\Top\Health\Conditions_and_Diseases	http://www.cancer.gov
	http://www.cancer.org
\Top\Health\Organisations\Medicine	http://www.ama-assn.org
	http://www.aafp.org
\Top\Health\Resources	http://health.nih.gov
	http://www.eyeglassretailerreviews.com

These categories are the most relevant ones related to medicine in ODP hierarchy (see http://www.dmoz.org/Health/Medicine/)

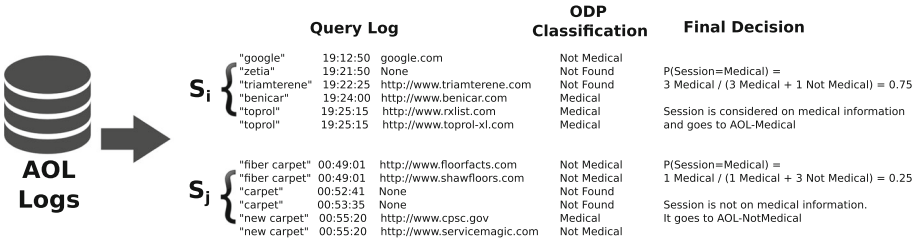

Fig. 3 Two real user sessions extracted from AOL logs, S_i is classified as a search for medical content, while S_j is not

were not used to calculate whether the session was on medical information or not, as their clicked URLs were not found in ODP. After the label estimation is done, all the queries of a session are assigned to the same class, therefore all six queries in S_i are assigned to AOL-Medical.

While only 27 % of the queries have their URLs found in ODP, using the session approach described above allows us to have 50 % of all sessions with at least one URL found in ODP. Altogether, 68 % of all AOL queries were evaluated, as they belong to sessions that had at least one clicked URL in ODP. A more accurate way to define sessions is a field of research by itself (He and & Göker 2000; Jones and Klinkner 2008; Gayo-Avello 2009) and it is not the goal of this work.

The **HON** dataset is composed of anonymous logs ranging from December 2011 to August 2013. This non-governmental organisation is responsible for the HONcode, a certification of quality given to websites fulfilling a pre-defined list of criteria (Boyer et al. 2011). HON provides a search engine to facilitate the access to the certified sites. Although the majority of the queries are issued in English, the use of French or Spanish is frequent. Aiming to reduce noise, every query in the HON dataset was re-issued in a commercial

search engine and the snippets of the top 10 results were used as input for an automatic language detection tool (Lui and Baldwin 2012), which presented a precision of 94 % in filtering English queries.

As expert datasets, we use the logs from the Turning Research Into Practice (**TRIP**) database[9] and ARRS **GoldMiner**.[10] The former is a search engine indexing more than 80,000 documents and covering 150 manually selected health resources such as MEDLINE and the Cochrane Library. Its intent is to allow easy access to online evidence-based material for physicians (Meats et al. 2007). The logs contain queries of 279,280 anonymous users from January 2011 to August 2012. GoldMiner consists of logs from an *image* search engine that provides access to more than 300,000 radiology images based on text queries of text associated with the images. Although the usage of an image search engine is slightly different from document search, previous work in the literature (Tsikrika and Müller 2012; Hollink et al. 2011) showed that the user search behaviour is similar. We had access to more than 200,000 queries with last logged query being issued in January 2012. Due to a confidentiality agreement, we cannot reveal the start date of this collection. The GoldMiner search engine is interesting because its users are so specialised and it therefore represents the particular case of catering to experts in a narrower domain inside medicine. As GoldMiner is so specialised, the number of laypeople using it is likely small. It is therefore a good example of the extreme specialisation end of the expert continuum, allowing the effects of this specialisation on the vocabulary and search behaviour of the users to be found.

3.3 Pre-processing log files

The first challenge dealing with different sources of logs is normalising them. Unfortunately, there is clickthrough URL information available only for the AOL and HON datasets, limiting a detailed click analysis. Therefore, we focus on a query content analysis, using only the intersection of all possible fields: (1) timestamp, (2) anonymous user identification, and (3) keywords. Neither stop word removal nor stemming were used.

Sessions were defined as follows. They begin with a query and continue with the subsequent queries from the same user until a period of inactivity of over 30 min is found. This approach for sessions, as well as the 30-min threshold, is widely used in the literature (Cartright et al. 2011; White and Horvitz 2012; Jones and Klinkner 2008). We excluded extremely prolific users (over 100 queries in a single session), since they could represent "bots" rather than individuals.

4 Enriching the query logs with MetaMap

The US National Library of Medicine MetaMap was intensively used in this work to enrich the information contained in the query logs, adding annotations regarding the concepts searched in the queries. MetaMap is widely used to map biomedical text to the Unified Medical Language System (UMLS) Metathesaurus, a compendium of many controlled vocabularies in the biomedical sciences (Aronson 2001). This mapping can serve for different tasks, such as query expansion (Aronson and Rindflesch 1997; Goeuriot et al. 2014), concept identification and indexing (Aronson et al. 2000; Névéol et al. 2009),

[9] http://www.tripdatabase.com/.

[10] http://goldminer.arrs.org.

question answering (Demner-Fushman et al. 2007), knowledge discovering (Weeber et al. 2000), and more related to this work, enrich query logs to understand user goals (Herskovic et al. 2007; Dogan et al. 2009). To explain how mapping queries to UMLS can give us some insights about the user intent, we first have to explain what UMLS is and how MetaMap maps text to UMLS. We explain how the mapping works in the next section and we evaluate the mapping in Sect. 4.2.

4.1 MetaMap

A Metathesaurus can be defined as a very large, multi-purpose, and multi-lingual vocabulary resource that contains information about biomedical and health related concepts, their various names, and the relationships among them (NLM 2009). In its 2013 version, the UMLS Metathesaurus has more than one hundred different controlled vocabulary sources and a large amount of internal links, such as alternative names and views of the same concept.

The top part of Table 2, showing concept C004238, is the original version of the classical UMLS example from NLM (2009). It illustrates how different atoms can have the same meaning. Atoms are the basic building blocks from which the Metathesaurus is constructed, containing the concept names or strings from each of the source vocabularies. The atoms shown are part of two vocabularies PSY (Psychological Index Terms), and MSH (Medical Subject Headings, MeSH), mapping different strings and terms to the same concept, C0004238, which states that atrial fibrillation is a pathological function. The other row of this table shows another concept, C1963067, mapped from the vocabulary NCI (National Cancer Institute), which states that atrial fibrillation can be an adverse event

Table 2 A concept is potentially linked to various AUI (atom), SUI (string), and LUI (term). We used MetaMap to map a user query, e.g., "Atrial Fibrillation" to the different existing concepts (C0004238, C1963067)

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs)
C0004238 (Pathologic Function) Atrial fibrillation (preferred) Atrial fibrillations Auricular fibrillation Auricular fibrillations	**L0004238** Atrial fibrillation (preferred) Atrial fibrillations	**S0016668** Atrial fibrillation (preferred)	**A0027665** Atrial fibrillation (from MSH)
			A0027667 Atrial fibrillation (from PSY)
		S0016669 (plural variant) Atrial fibrillations	**A0027668** Atrial fibrillations (from MSH)
	L0004327 (synonym) Auricular fibrillation Auricular fibrillations	**S0016899** Auricular fibrillation (preferred)	**A0027930** Auricular fibrillation (from PSY)
		S0016900 (plural variant) Auricular fibrillations	**A0027932** Auricular fibrillations (from MSH)
C1963067 (Finding) Atrial fibrillation (Atrial fibrillation adverse event)	…		Auricular fibrillations (from NCI)

Note that each concept is associated to one single semantic meaning

associated with the use of a medical treatment or procedure, although we do not know which medical treatment or procedure.

The job of MetaMap is to map a biomedical text to its corresponding concept(s). MetaMap generates a candidate set for a piece of text, based on its internal parser and variant generation algorithm, which takes into account acronyms, synonyms, inflections and spelling variants of the text. Then, based on metrics such as centrality, variation, coverage and cohesiveness, MetaMap ranks each candidate (Aronson 2001). Occasionally, more than one candidate may have the same score. We collect all the top candidate(s) and its (their) associated semantic type(s), shown in bold below the CUIs in Table 2. In the running example, a text containing only '*atrial fibrillation*' is mapped to both C0004238 and C1963067 with the same top score, and the types 'Pathologic Function' and 'Finding' are assigned to the query. To the best of our knowledge, MetaMap is the state of the art for mapping biomedical text to UMLS concepts.

An interesting way to capture the user intent is mapping the queries to a well known domain corpus. In this work we use the Medical Subject Headings, MeSH, as it is a rich and well structured hierarchy that has already been studied to examine user query logs (Herskovic et al. 2007), allowing us to compare the behaviour of the users studied here with PubMed users. The whole MeSH hierarchy contains more than 25,000 subject headings in the 2013 version, the one used in this work, containing 16 top categories such as '*Anatomy*' and '*Diseases*'. Figure 5 shows an example of the MeSH hierarchy with the first level of the disease branch expanded.

We use the approach of Herskovic et al. (2007) in this paper, mapping each query onto one or more MeSH terms with MetaMap. As shown in Fig. 4, one query can be mapped to multiple MeSH identifiers. For example, the query '*atrial fibrillation*' is mapped to both MeSH ids *C14.280.067.198* and *C23.550.073.198*, both in the topmost *Disease* category (represented by the starting letter 'C' as show in Fig. 5). After the mapping to MeSH is done, we can easily have an overview regarding the subjects the users are more interested in. In this case we would conclude that this user is interested in diseases, as her/his only query maps only to category 'C', more specific in cardiovascular diseases, C14, and pathological conditions, C23.

After preprocessing, each query is converted into the following format: *<timestamp, userID, query, semanticTypes, meshIDs>*, where the *timestamp*, *userID* and *query* are originally query log fields, while *meshIDs* and *semanticTypes* are the set of semantic types and MeSH identifiers generated by MetaMap. These two fields are examined in details in Sects. 5.2.2 and 6.2. Figure 4 illustrates how the queries were enriched with the information provided by MetaMap and the final format.

Finally, it is important to mention that the queries were mapped to concepts in the UMLS 2013 AA USAbase Strict Data and no special behaviour parameter was used. We manually examined the behaviour for two important parameters: allowing acronyms/abbreviations (-a) and using the word sense disambiguation module (-y), and decided not to activate them. Our experiments show that activating the former parameter decreases the precision significantly for the sake of a small increase in recall, as MetaMap is already capable of matching some of the most frequently used abbreviations (HIV, HPV, AIDS, COPD). For the latter, we have an inverse scenario, where we had a small gain in precision but a larger loss in recall, as MetaMap always picks only one possibility when more than one concept is possible. It means that MetaMap would be forced to choose between concepts C0004238 and C1963067 of Table 2, even when both are equally likely. The last important reason for not using any other parameter is that we want to compare our results with Herskovic et al. (2007), in which no special option was used either. For the

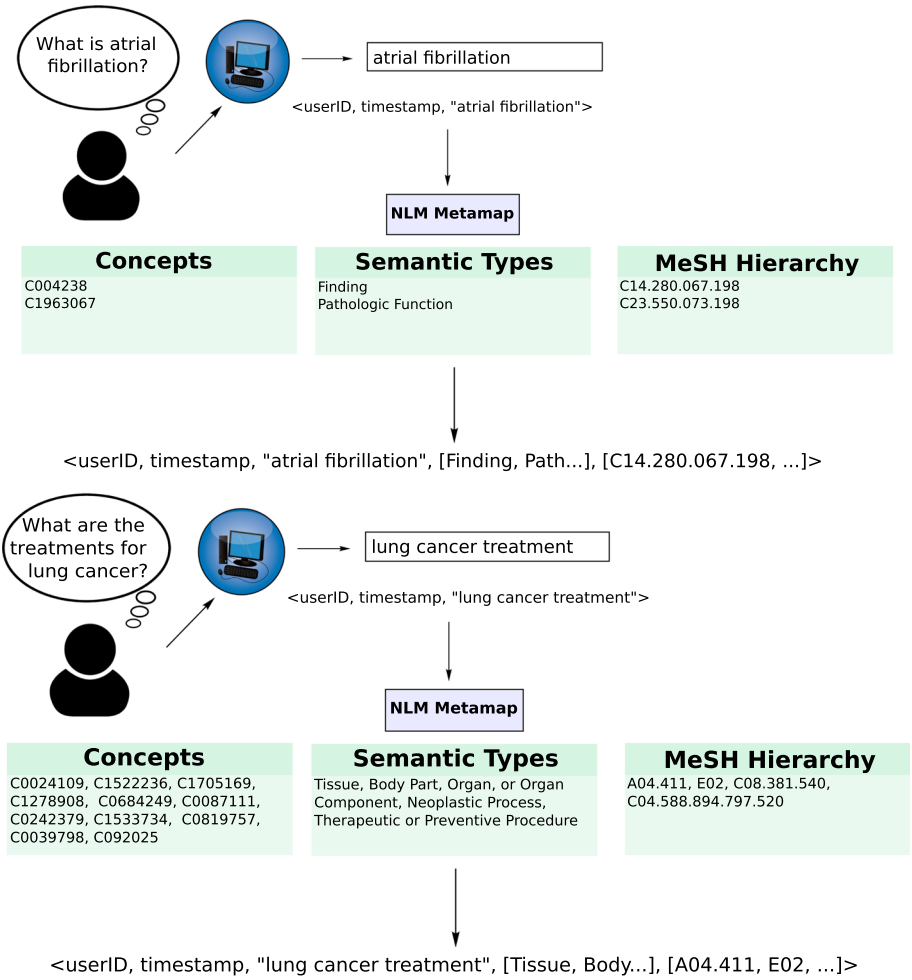

Fig. 4 Two different user queries are enriched with information extracted with MetaMap. In the top part, the same example used in Table 2 is processed by MetaMap. In the *bottom part*, the query "lung cancer treatment" is more ambiguous and results in different mappings, such as *Lung* (*Entire lung*)/*Cancer Treatment* (*Cancer Therapeutic Procedure*) and *Lung Cancer* (*Malignant neoplasm of lung*)/*Treatment* (*Therapeutic procedure*)

experiments shown in Sect. 5.2.2 we used the parameter (-R) to restrict MetaMap to use only MeSH as vocabulary source.

4.2 Evaluation of the mapping

As recently reported by MetaMap's authors Aronson and Lang (2010), a direct evaluation of MetaMap against a manually constructed gold standard mapping to UMLS concepts has almost never been performed. Usually, indirect evaluations are made, where the effectiveness of a task is measured with and without MetaMap. For example, query expansion using the related concepts of a concept identified by MetaMap versus not using it. Here we

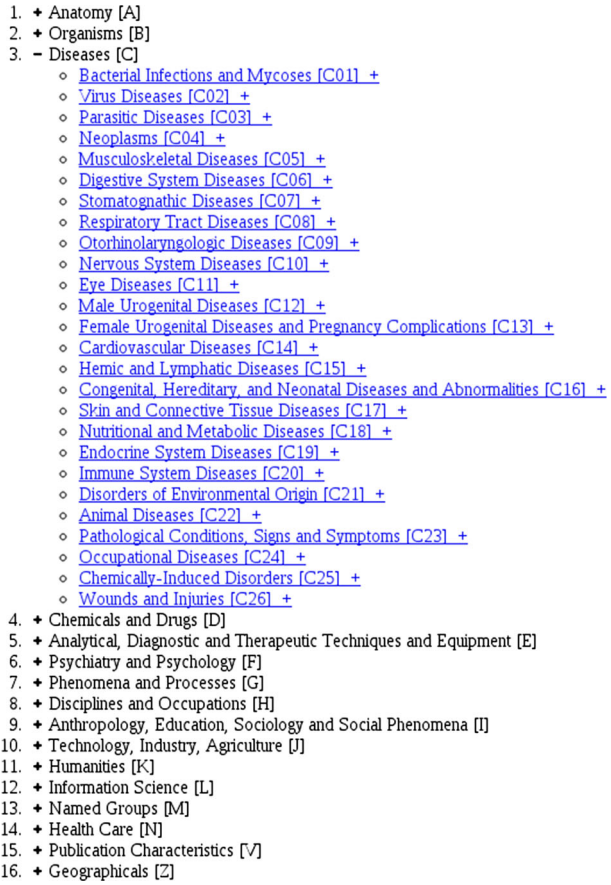

Fig. 5 MeSH hierarchy with the disease branch expanded

are interested in the few articles that evaluate the effectiveness of MetaMap, especially the ones focused on mapping user queries.

In 2003, Pratt and Yetisgen-Yildiz (2003) compared MetaMap mappings to UMLS with mappings made by six physicians and nurses. For the 151 concepts in their ground truth, MetaMap could match 81 concepts exactly, 60 partially and could not match only 10 concepts, of which 6 were not available in UMLS. In a scenario considering partial matches (e.g., mapping to '*angiomatosis*' instead of '*leptomeningeal angiomatosis*'), MetaMap had an F1 of 76 %. In another experiment in the same year, Denny et al. (2003) built a bigger gold standard dataset of 4281 concepts to evaluate MetaMap, reaching a precision of 78 %, recall of 85 % and F1 of 81 %.

More recently, Névéol et al. (2009) reported results on using MetaMap to detect disease concepts on both literature and query corpus. The results showed that MetaMap had a better effectiveness for long sentences (F1 of 76 %) than for short queries (F1 of 70 %), but they also pointed out that the average inter-annotator agreement of the three assessors for the query corpus was 73 %, showing that MetaMap results are not far from humans performing the same task. Using 1000 queries from partly the same datasets that are used

here: AOL, HON and TRIP, Palotti et al. (2014b) also showed an F1 of 70 % for query mappings.

Névéol et al. (2011) created an annotated set of 10,000 queries that were mapped to 16 categories, in a similar way to what is done in Sect. 6.2, where the semantic types produced by MetaMap are used to define our own categories. We used Névéol's dataset to calibrate our mappings for our 'Cause' and 'Remedy' categories (see Sect. 6.2), as well as to take decisions regarding MetaMap's parameters. We used the 'Disorder' category of Névéol as an equivalent of our 'Cause' category, and we combined '*Chemical and Drugs*' (antibiotic, drug or any chemical substance), '*Gene, Proteins and Molecular Sequences*' (name of a molecular sequence) and '*Medical Procedures*' (activity involving diagnosis, or treatment) as the closed possible class of our 'Remedy' class. We could reach an F1 of 78 % for the 'Cause' category ($P = 75\,\%, R = 81\,\%$) and 72 % for 'Remedy' ($P = 70\,\%, R = 73\,\%$). These figures are in line with what is known for MetaMap when mapping medical abstracts to concepts, encouraging us to use it for mapping short queries to concepts as well.

4.3 Using the mappings

In the following sections, we show how we exploit the mappings made by MetaMap to enrich query logs. In Sect. 5, we use the mappings to analyse individual queries, following a very similar approach carried out by Herskovic et al. (2007), being able to compare our results for individual queries. Later, in Sect. 6, the focus is on the session level. An interesting work which we took as a basis for comparison is Cartright et al. (2011), which defines three classes: symptoms, diseases and remedies. Note that we could group the MeSH hierarchy into these three classes, but we prefer to use the semantic types provided by MetaMap, as it is more intuitive and it was already done in the literature (for example Jadhav et al. 2014; Névéol et al. 2011; Palotti et al. 2014a, b).

5 Individual query analysis

One goal of this section is to study how users search, based first on simple statistics to model their behaviour. Also, we start exploring the content of their queries, but considering all the queries without dividing into user sessions.

5.1 How users search

We start by showing a few simple but important statistics about the logs. The aim of this section is to understand the user behaviour through general statistics, as well as to show how each log is composed. In Table 3 we depict several metrics that are used to characterise user interactions, and compare their values to those in related studies. Duarte Torres et al. (2010) use AOL logs to study queries performed by kids. White et al. (2009) use a keyword-based method to filter domain specific queries and divide them into those issued by laypeople and those issued by experts. Their work also considers other types of queries, such as queries on computer science or financial information. We show only the data for the medical domain. Herskovic et al. (2007) and Dogan et al. (2009) analyse different periods of PubMed logs. For all datasets, "N/A" is used when the information is not available.

Table 3 General statistics describing the query logs

Dataset	This work					Literature					
	Laypeople		Experts		Non-medical	AOL-Kids	AOL-NKids	Laypeople	Experts	PubMed	
	AOL-M	HON	TRIP	GM	AOL-NM	Duarte Torres et al. (2010)		White et al. (2009)		Herskovic et al. (2007)	Dogan et al. (2009)
Logs initial date	Mar 2006	Dec 2011	Jan 2011	N/A	Mar 2006	Mar 2006		May 2007		Jan 2006	Mar 2008
Logs final date	May 2006	Aug 2013	Aug 2012	Jan 2012	May 2006	May 2006		Jul 2007		Jan 2006	Mar 2008
No. of users	47,532	47,280	279,280	45,090	655,292	N/A	N/A	37,243	7971	624,514	NA
No. of queries	215,691	140,109	1,788,968	219,407	34,427,132	485,561	N/A	673,882	362,283	2,689,166	58,026,098
No. of unique queries	69,407	85,824	486,431	90,766	9,695,882	10,252	N/A	N/A	N/A	N/A	N/A
No. of sessions	79,711	77,977	344,038	100,848	10,555,562	21,009	N/A	68,036	26,000	740,215	23,017,461
Avg. terms per query	2.61 (±1.71)	2.72 (±2.05)	3.40 (±2.33)	2.28 (±2.54)	2.46 (±1.87)	3.23	2.5	2.92	3.30	3[a]	3.54
Avg. char per query	16.22 (±9.11)	18.11 (±11.48)	21.22 (±9.69)	16.64 (±10.20)	15.98 (±9.67)	N/A	N/A	20.76	24.05	N/A	N/A
Avg. queries per session	2.71 (±2.50)	1.80 (±2.48)	5.20 (±5.95)	2.18 (±2.57)	3.26 (±4.65)	8.76	2.8	9.90	13.93	N/A	4.05
Avg. time per session (s)	258 (±531)	208 (±592)	471 (±758)	163 (±520)	384 (±809)	1238	N/A	1549.74	1776.45	N/A	N/A

N/A data was not available

[a] median used as the mean was not informed

The query logs from the related work shown in Table 3 belong to the same time period as the AOL logs. Query logs from HON, TRIP and GM are considerably newer than the others. Nevertheless, Table 3 shows that AOL-Medical and HON are very similar in many aspects, such as the average number of terms per query and the average time per session. The biggest difference between these two query logs was found for the average number of queries per session, however the difference is small if compared to any other datasets.

The average terms/characters per query can be an indicator of the complexity and difficulty of the users to express their information needs. We note that AOL-Medical and HON queries are shorter than TRIP queries, and that TRIP logs are similar to PubMed logs in terms of query length. White's work also found that expert queries are more complex than layperson queries.

The average number of queries per session and time per session, although considerably smaller than what was found by White's work, follow the same pattern, with TRIP data having longer sessions than HON and AOL-Medical. We could not find an explanation for so long queries in White et al. dataset. We show only the sessions made by experts and laypeople in the medical domain from White's work, but in their original paper they report that sessions are considerably smaller when the same set of people query in other domains: having a mean session length of less than five queries, and the mean time per session is never longer than 800 s.

We aggregate the log into two groups in Table 4: *laypeople* and *experts*, making the comparison of our datasets with the literature possible. As done by White et al. (2009), we use Cohen's $\underline{d}$ to determine the effect size of each variable between each pair of groups. We randomly sampled 45,000 users from TRIP and merged them with the 45,090 users from GoldMiner, making all datasets have a comparable number of users. Cohen's $\underline{d}$ is a useful metric for meta-analysis (Cohen 1988) that uses the means and standard deviations of each measurement to calculate how significant a difference is. Although there are controversies about what is a "small", "medium" or "large" effect size, a recommended procedure is to define a Cohen's $\underline{d}$ effect size of 0.2 or 0.3 as a "small" effect, around 0.5 as "medium" effect and greater than 0.8 as a "large" effect (Cohen 1988). White et al. built a classifier to detect user expertise based on a superset of the features shown in

Table 4 General statistics—stratified by expertise

Dataset	Laypeople	Experts	Cohen's $\underline{d}$	
Total number of users	94,812	90,090	E – L	E – L from White et al. (2009)
Total number of queries	355,800	504,745		
Total number of unique queries	149,648	181,051		
Total number of sessions	157,688	155,965		
Mean terms per query	2.65 (± 1.85)	2.91 (±2.09)	0.13	0.20
Mean chars per query	16.97 (± 10.16)	19.18 (±10.16)	0.22	0.30
Mean queries per session	2.26 (± 2.53)	3.24 (±4.29)	0.28	0.38
Mean time per session (s)	233 (±562)	271 (±629)	0.06	0.11

L laypeople, *E* experts

Table 4. They argued that these are valuable features based on Cohen's d value, as well as feature importance calculated by their regression classifier. Although considered to have a "small" effect, this was big enough to help separate experts from laypeople. We reached very similar Cohen's d values to White's paper, hypothesising that the behaviour could be used to predict expertise in other logs as well. In particular, we found the same ranking that White et al. found among the four features presented in Table 4.

5.2 What users search for

In order to understand what users are looking for, we investigate popular terms and queries issued. Also, we use MetaMap to map queries to the MeSH hierarchy, finding the high level topics associated with the user queries.

5.2.1 Terms and queries

We depict the most popular queries, terms (here excluding the stop words), and abbreviations used in all logs, as well as their frequency among the queries in Table 5. As expected, AOL-NotMedical contains navigational queries and diverse terms related to entertainment. Similarly, some of the most popular queries in AOL-Medical are navigational, with the website '*webmd.com*' appearing twice in the top 10 queries, and the Mayo Clinic also a common query. Both of these navigational queries also appear in the HON search log. The analysis of AOL-Medical terms shows common medicine-related concepts, with people searching for information about different cancer types in more than 3 % of the cases.

Most of the top queries in the TRIP log are related to disease. In TRIP logs, we found '*area:*' in 3 % of the queries, '*title:*' in 2.2 %, '*to:*' in 1.5 % and '*from:*' in 1.8 %, in total these keywords were used in 6.7 % of the queries, however, we do not show these terms in Table 5, as they do not reveal what the users search, but how they search. These patterns were not found in the other datasets. The use of more advanced terms is also found in PubMed logs (Herskovic et al. 2007), we hypothesise that some users might just copy and paste their queries from PubMed into the TRIP search engine, resulting in queries such as '*palliative care (area:oncology)*', indicating that the user wants material about palliative care specifically for the area of oncology. 'Title' is used in PubMed for performing a search only in the title of the indexed articles, while 'from:' and 'to:' specify periods of time in which a document was published.

The topmost query in the HON log and its top 3 terms are '*trustworthy health sites*'. It shows that many of the queries are from users that do not know which are the medical websites that they can trust, and also demonstrates a misunderstanding by the end users of the nature of the content indexed by the HON search engine (only HONcode-certified websites are indexed).

For the GoldMiner queries and terms, we clearly see the increase in the terminological specificity of the most popular keywords used.

5.2.2 Mapping to MeSH

MeSH is a hierarchical vocabulary used by US National Library of Medicine for indexing journal articles in the life sciences field. A query log analysis using MeSH was also carried out by Herskovic et al. (2007) for the PubMed logs in order to understand what are the

Table 5 Top queries and terms and their relative frequency (%) among all queries

Rank	Laypeople				Experts					
	AOL-Medical		HON		TRIP		GoldMiner		AOL-NotMedical	
	String	Freq.	String	Freq.	String	Freq.	String	Freq.	String	Freq.
Queries										
1	webmd	0.98	trustworthy health sites	4.24	skin	0.29	mega cisterna magna	0.44	google	0.95
2	web md	0.41	cancer	0.51	diabetes	0.22	baastrup disease	0.40	ebay	0.40
3	shingles	0.27	webmd	0.47	asthma	0.17	toxic	0.23	yahoo	0.37
4	mayo clinic	0.26	sleep apnea syndromes	0.27	hypertension	0.14	limbus vertebra	0.22	yahoo.com	0.28
5	lupus	0.25	lymphoma	0.22	stroke	0.13	cystitis cystica	0.20	mapquest	0.25
6	herpes	0.20	breast cancer	0.21	osteoporosis	0.11	thornwaldt cyst	0.14	google.com	0.23
7	diabetes	0.19	hypertension	0.18	low back pain	0.10	buford complex	0.13	myspace.com	0.22
8	fibromyalgia	0.18	mayoclinic.com	0.16	copd	0.10	splenic hemangioma	0.13	myspace	0.21
9	pregnancy	0.16	obesity	0.16	breast cancer	0.09	throckmorton sign	0.12	www.yahoo.com	0.12
10	hernia	0.16	drweil.com	0.14	pneumonia	0.09	double duct sign	0.12	www.google.com	0.12
Terms										
1	cancer	3.40	health	6.39	treatment	3.03	cyst	3.17	free	1.24
2	hospital	3.00	sites	4.37	cancer	2.56	mri	1.89	google	1.04
3	pain	2.25	trustworthy	4.28	pain	2.13	disease	1.80	county	0.65
4	symptoms	2.14	cancer	2.74	care	2.10	ct	1.75	yahoo	0.62
5	disease	2.03	disease	1.53	children	1.98	fracture	1.68	pictures	0.60
6	blood	1.87	diabetes	1.17	therapy	1.81	tumor	1.65	lyrics	0.52
7	medical	1.62	treatment	0.96	diabetes	1.80	syndrome	1.47	school	0.51
8	webmd	1.21	syndrome	0.87	disease	1.78	liver	1.26	myspace	0.49
9	surgery	1.14	heart	0.83	pregnancy	1.70	pulmonary	1.22	ebay	0.46
10	syndrome	1.13	pain	0.80	acute	1.41	bone	1.16	sex	0.44
11	breast	1.11	care	0.77	syndrome	1.39	renal	1.13	florida	0.45

Table 5 continued

Rank	Laypeople				Experts					
	AOL-Medical		HON		TRIP		GoldMiner		AOL-NotMedical	
	String	Freq.	String	Freq.	String	Freq.	String	Freq.	String	Freq.
12	center	1.09	effects	0.75	management	1.14	sign	1.12	sale	0.41
13	health	1.04	medical	0.67	stroke	1.07	lung	1.11	city	0.40
14	heart	0.90	blood	0.65	surgery	1.06	brain	1.08	home	0.39
15	diabetes	0.86	pregnancy	0.61	prevention	1.05	cell	1.00	state	0.39

most popular topics searched by the users. We use the same weighting schema used in Herskovic's work: if n categories are detected in one query, we give the weight of $1 / n$ to these categories.

General statistics calculated for the mapping of user queries to MeSH terms are shown in Table 6. Here, we are testing MetaMap for the annotation of non-medical queries as well, which to the best of our knowledge was never studied.

An interesting result is the fact that around 50 % of AOL-NotMedical queries were successfully mapped to a MeSH concept. To investigate this, we collected a large random sample of mapped queries and analysed them. We found that MetaMap is able to find many concepts not directly linked to medicine, such as geographic locations, animals and plants, food and objects. For example, 'www' (*L01.224.230.110.500*), used in 10 % of all AOL queries, is recognised and annotated as *Manufactured Object*. Also, locations are usually very commonly found and help to explain the high mean MeSH depth found for this dataset, second row in column AOL-NM (California is mapped to both *Z01.107.567.875.760.200* and *Z01.107.567.875.580.200*). It is important to have this in mind when building systems like in Yan et al. (2011), in which the MeSH depth is used to model document scope and cohesion. When looking at false positive mappings, especially the ones mapping to diseases and symptoms, we detected that MetaMap's errors fall into two main categories: (1) English common words: tattoo (tattoo disorder), pokemon (ZBTB7A gene), and (2) abbreviations: park (Parkinson disease), dvd (Dissociated Vertical Deviation). For both types of errors, MetaMap or a system using it, would have to use the context (words around the mapping) to detect that Pokemon is used as a cartoon or a game, and not as a gene name. Specifically for the second case, it would be desirable if MetaMap could allow the use of a pre-defined list of acronyms to increase its precision. In the current implementation, MetaMap has a parameter for user defined acronyms (-UDA), but it is just used to expand more acronyms instead of overwriting its pre-defined ones. Also for AOL-NL, the third and fourth rows indicate the suitability of using mappings to MeSH for distinguishing between medical and non-medical queries. Queries from the medical logs have a larger number of MeSH terms and disease terms than AOL-NM. If the errors analysed above could be amended using the query context or session, for example, then a mapping to MeSH could be helpful to detect queries or sessions on medical information.

Going further, we present in Fig. 6 the most popular categories for the first level of the MeSH hierarchy. We also show the results obtained by Herskovic et al. (2007) for PubMed, in order to compare our findings. We show only the categories that have more than 5 % of the queries containing MeSH terms mapped to it.

Table 6 General MeSH statistics

Metric	Laypeople		Experts		AOL-NM
	AOL-M	HON	TRIP	GM	
Percentage of queries containing MeSH terms	77.87	77.81	85.96	79.02	50.51
Mean MeSH depth	3.99	3.83	3.86	4.01	4.37
Mean MeSH terms per query	2.14	2.19	2.78	2.07	1.12
Mean disease terms per query	0.81	0.60	0.99	1.17	0.05

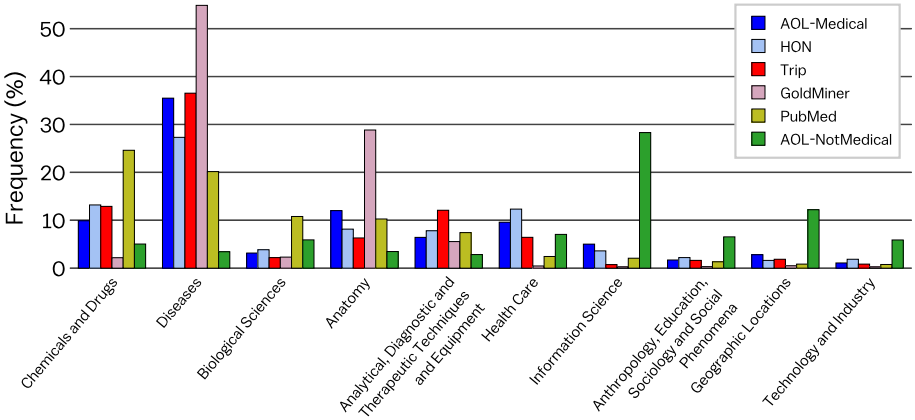

Fig. 6 Popular categories according to MeSH mappings (Color figure online)

When Herskovic and colleagues did this experiment, they found that PubMed users were more interested in the category *Chemical and Drugs*. In general, the distributions over the categories for the AOL-Medical, HON and TRIP search logs are similar. However, differently from PubMed, we found that the users are generally most interested in *Diseases*, and then *Chemicals and Drugs*. The results for GoldMiner show another trend for the second most popular category, focused on anatomy rather than on drugs, likely because radiologists often have to append to their query the part of the body that they are interested in. In its actual version, GoldMiner has a filter for age, sex and modality (e.g., CT, X-ray), but it has no filter for body parts. This analysis suggests that it could be interesting to add a filter for body regions as well.

Last, the four classes to the right of Fig. 6 partly explain the high percentage of AOL-NotMedical terms mapped to MeSH terms. Also, the high percentage of these least medical categories, together with low percentage of relevant medical categories, the four classes to the left of Fig. 6, can be used as a discriminative feature to distinguish between medical and non-medical logs.

6 Analysing sessions

From now on, we consider user sessions instead of separate queries. Once more, we study first the user behaviour, then the content of each session.

6.1 Session characteristics

A series of queries, part of an information seeking activity, is defined as a session. We consider that, after issuing the first query, a user may act in four different ways: (1) repeat exactly the same query, (2) repeat the query adding one or more terms to increase precision, (3) reduce the number of terms to increase recall, or (4) reformulate the query changing some or all the terms used. We ignore the first case because we cannot be sure if a user is really repeating the same query or just changing the result page, as some search engines record the same query as a result of a page change.

Table 7 depicts the changes made by users during the sessions. If during one single session a user adds a term to the previous query and then changes a few words, we count one action in the row Exp.Ref (for expansion and reformulation – the order is not important). At the end, we divide the number of actions of each row by the total actions in the query log. Hence, Table 7 shows that the most frequent user action is the reformulation alone but it is more likely to happen in search engines targeting laypeople, e.g., 84 % of the sessions in the AOL-Health logs and 63 % of HON had only reformulations. The last row of Table 7 shows that expert users might be more persistent than laypeople, as more than 10 % of the sessions in the professional search engines are composed of every type of action, while in laypeople logs this number is less than a third of this. In the literature, White et al. (2009) also hypothesise that expert users are more persistent than laypeople.

To better understand the last row of Table 7, we plot the two graphs in Fig. 7. The first graph is the cumulative distribution of session length, showing that TRIP has clearly longer sessions, with 20 % of the sessions being longer than five queries. The second graph shows the last row of Table 7 distributed over different session length. In this graph we can see how TRIP and GoldMiner users tend to perform more actions even for short sessions, as 20 % of sessions of length 4 have already done all 3 actions. We also studied the user behaviour when the query repetition is allowed and we found a very similar situation.

6.2 What are the sessions about?

In this section, we attribute meaning to the users' queries in order to better understand their behaviour in a medical search context. We decide to use the same classes defined in Cartright et al. (2011): symptom, cause and remedy, so that a direct comparison can be performed. A difference of their method and ours is that we classify the queries into the semantic types using MetaMap, as done in Jadhav et al. (2014), Névéol et al. (2011), Palotti et al. (2014a, b), instead of handmade rules.

In Fig. 8, we show all concepts that have a frequency of at least 5 % in any query log. Additionally, we show the type *'Sign and Symptom'* because it is an important concept in our further analysis. We show only these 10 semantic types for a matter of readability, as currently MetaMap recognises 133 semantic types and it is not possible to visualise them all.[11] The single most common type in all the medical logs is '*Disease and Syndrome*'. As

Table 7 Aggregated percentages for query modifications along the sessions

Action	Laypeople		Experts		AOL-NM
	AOL-M	HON	TRIP	GM	
Expansion	6.66	13.83	14.85	5.96	3.71
Reduction	1.23	2.23	4.35	9.61	0.84
Reformulation	84.74	63.56	43.96	49.56	80.27
Exp. and Red.	0.37	1.29	5.09	3.54	0.57
Exp. and Ref.	5.43	13.90	15.27	8.28	9.66
Red. and Ref.	1.01	2.21	5.63	12.01	2.09
Exp. Red. Ref.	0.56	2.98	10.85	11.04	2.86

[11] A complete list of all semantic types can be found online: http://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml.

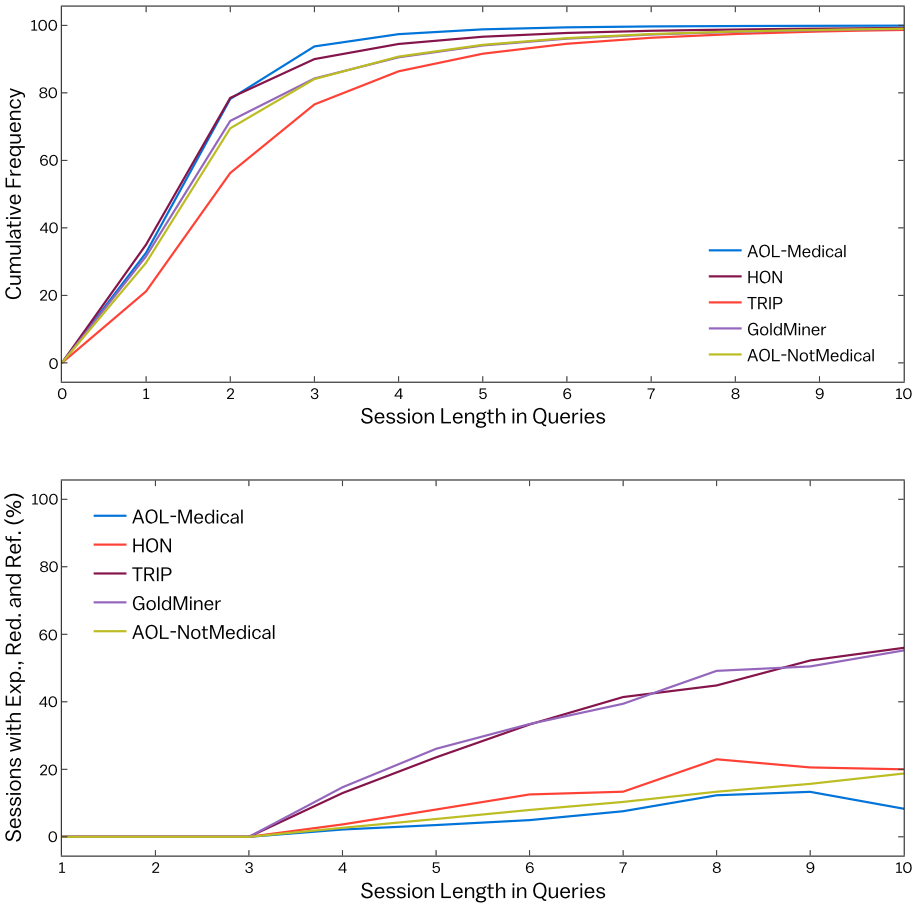

Fig. 7 The *top* graph shows the cumulative distribution of sessions length in terms of number of queries. The users of HON and GoldMiner tend to have shorter sessions, while the users of TRIP have longer sessions. The graph *below* shows the percentage of users that in a single session perform all three actions (expand, reduce and reformulate the previous query) for sessions of different sizes. As we can expect, this percentage increases as the query length increases but it is much higher for the expert users (Color figure online)

we expect, the top types in AOL-NotMedical are not really related to the medical domain, and the second most common semantic type for GoldMiner is related to parts of the body, as one might expect for radiology queries.

After a meticulous analysis of the semantic meaning assigned for the queries and the experiments described in Sect. 4.2, we defined the following classification based on the three classes created in Cartright et al. (2011) (some examples of queries classified for each type are given for a better understanding):

- *Symptom:* Sign or Symptom (*cough*; *sore*; *headache*; *red eyes*), Findings (*stress*; *testicular cyst*)
- *Cause:* Anatomical Abnormality (*hiatial hernia*), Cell or Molecular Dysfunction (*macrocytos*), Congenital Abnormality (*scoliosis*), Disease or Syndrome (*diabetes*;

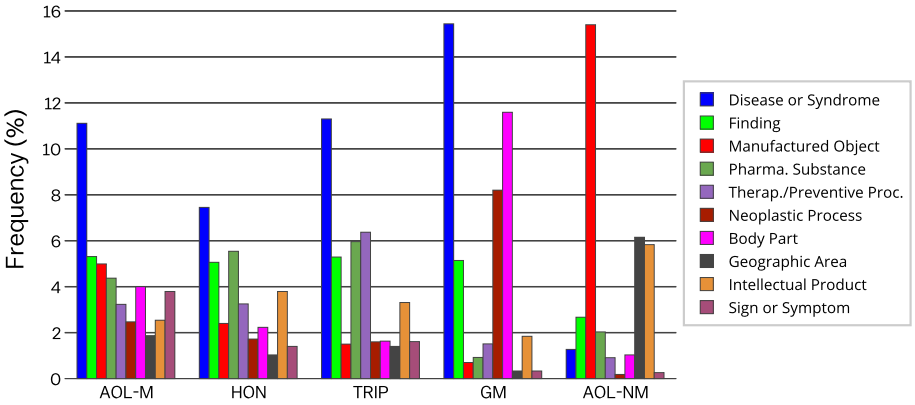

Fig. 8 The top most frequently used semantic types (frequency in percentage). Many of the most used types are aggregated to study the user focus described in Table 8 (Color figure online)

heart failure), Experimental Model of Disease (*cancer model*), Injury or Poisoning (*achilles tendon rupture*), Mental or Behavioural Dysfunction (*bipolar disorder*), Neoplastic Process (*lung cancer*; *tumor*), Pathologic Function (*atypical hyperplasia*)

- *Remedy:* all 28 types belonging to the high-level group Chemicals and Drugs, which includes Clinical Drug (cough syrup), Antibiotic (penicillin), Pharmacologic Substance (tylenol; mietamizol), Amino Acid, Peptide, or Protein (vectibix; degarilex), Immunologic Factor (vaccine; acc antibody), Vitamin (quercetin, vitamin B12), Therapeutic or Preventive Procedure (treatment; physiotherapy), etc.

We analyse the most popular semantic types found in the queries and show them in Table 8, together with a direct comparison to Cartright et al. (2011). The largest difference between all four medical logs analysed in this paper and the Cartright et al. results is in the symptom category. For the latter, 63.8 % of the sessions are focused on symptoms, while between 5.5 and 9.1 % are focused on symptoms in our analysis. The main reason for Cartright's result is linked to the way in which they created their dataset: *keeping only sessions that had at least one query containing a term in a wordlist extracted from a* ***list of symptoms*** *from the Merck medical dictionary*. Their preprocessing step therefore explains

Table 8 User focus when searching for medical content in a single session

Intent	Laypeople		Experts		AOL-NotMedical	Cartright et al. (2011)
	AOL-Medical	HON	TRIP	GoldMiner		
None	34.0	40.4	16.8	21.2	82.9	3.9
Symptom	9.1	6.3	5.5	6.4	3.9	63.8
Cause	24.3	20.9	26.0	58.2	3.3	5.3
Remedy	14.7	16.2	17.4	3.3	7.5	1.1
Symptom and cause	6.8	6.1	7.2	6.4	0.5	22.6
Symptom and remedy	2.1	2.6	4.5	0.5	0.9	2.0
Cause and remedy	7.1	5.0	15.9	3.0	0.8	0.4
All three	1.9	2.5	6.7	1.0	0.2	0.8

the fact that most of the sessions were concentrated only on searching for symptoms. Conversely, our analysis reveals that the most common user focus is on causes rather than on symptoms. Also, the second most common focus is on a way to cure a disease. It is important to note that Cartright et al. logs date from 2009, it means they are 3 years younger than AOL, but roughly 3 years older than HON, also suggesting that large divergence found is due to the preprocessing steps and not to an evolution on how the users search.

Once more, GoldMiner presents a different behaviour, we hypothesise that the low number of sessions on remedies is explained by the fact that radiologists are not interested in remedies when searching for images as they are rather in the diagnosis phase. It is interesting to note that searching for causes and remedies in the same session is a very frequent task for medical professionals in the TRIP logs, with 16 % of the sessions searching for both remedies and causes.

In Table 9, we show the behaviour modifications along a session. One oscillation is characterised by a transition from one focus type to another and then back to the original type. Originally, this study was made to support the hypothetico-deductive searching process in which a user cyclically searches for a symptom, then a cause and then returns to symptom (Cartright et al. 2011). The symptom-cause pattern was also found in our experiments, but with a more balanced distribution in relation to the other patterns. Again, the large number of behaviours involving symptoms found in Cartright et al. (2011) is likely an artefact of how the dataset was constructed. We see that the cause-remedy pattern plays a very important role, especially in the TRIP log, in which this is the most common pattern. Finally, the least frequent pattern found in all four datasets is the symptom-remedy one. The study of the behaviour modification was used in Cartright et al. (2011) to build a classifier to predict what is the next user action, allowing a search system to support medical searchers by pre-fetching results of possible interest or suggesting useful search strategies.

7 User classification

We have seen in Sects. 5 and 6 that experts and laypeople use different search strategies. In this section, we take advantage of these differences to build an automatic classifier that can assist search systems, exploring the user domain knowledge.

The expertise inference can be directly applied by a search engine to tailor the results shown, e.g., boosting easy to read documents for laypeople (Walsh and Volsko 2008;

Table 9 Cycle sequence along a single session

Pattern	Interaction	Laypeople		Experts		Cartright et al. (2011)
		AOL-Medical	HON	TRIP	GoldMiner	
	Sessions with oscillations (%)	23.07	13.48	64.61	8.60	16.2
Symptom-cause	Symptom → Cause → Symptom	19.2	15.6	13.2	22.7	51.4
	Cause → Symptom → Cause	19.9	18.8	14.5	35.3	38.4
Symptom-remedy	Symptom → Remedy → Symptom	8.2	11.8	10.8	4.1	5.1
	Remedy → Symptom → Remedy	8.1	14.2	11.6	3.8	2.7
Cause-remedy	Cause → Remedy → Cause	18.2	18.4	24.8	20.3	1.5
	Remedy → Cause → Remedy	26.4	21.2	25.1	13.8	0.9

Palotti et al. 2015; Collins-Thompson et al. 2011), or search aids, such as query suggestions to match the searcher expertise. Also, the search strategies employed by experts could be used to support non-experts in learning more about domain resources and vocabulary (White et al. 2009).

In order to take advantage of the user domain expertise, it is necessary to be able to identify whether a user is an expert or not. We employed a Random Forest classifier.[12] to solve this binary classification problem, since it is a well-known machine learning technique and has shown to be suitable for this task before (Cole et al. 2013; Palotti et al. 2014a).

The classifier relies upon a set of features to take its decision between the two modelled classes: expert or layperson. We list 14 features proposed in this work in Table 10, and group them into two sets: (1) user behaviour features, and (2) medicine-related features. The first set is made from the analysis of Sects. 5.1 and 6.1, while the second one covers Sects. 5.2 and 6.2.

To form our dataset, we merged the users from AOL-Medical and HON logs into the laypeople class, and the users from TRIP and GoldMiner logs into the expert class. As noted in Sect. 5.1, the number of users from TRIP logs is considerably larger than the other logs, therefore we repeat the sampling made to generate Table 4, and 94,812 users are created for the layperson class and 90,090 for the expert one.

We performed a 10-fold cross-validation experiment and present the results in Table 11. We employed as the baseline a simple classifier that always outputs the positive class, which could reach an F_1 of 67.8 %. The next two rows of Table 11 show the classification

Table 10 Features used in the expertise classification task

Group	Feature	Explanation
User behaviour features	AvgCharPerQuery	Average number of characters and terms used by the user in each query
	AvgTermsPerQuery	
	AvgQueryPerSession	The average number of queries and time per session
	AvgTimePerSession	
	AvgExpansions	Compares the ith query to $(i-1)$th query and counts the expansions, reductions and reformulations made
	AvgReductions	
	AvgReformulation	
Medical related features	AvgSymptomsPerQuery	The average number of symptoms/causes/remedies/none of them per query
	AvgCausesPerQuery	
	AvgRemediesPerQuery	
	AvgOtherTypePerQuery	
	PercQueriesWithMeSH	Percentage of queries that could be mapped to any MeSH concept
	AvgMeSHPerQuery	Average number of MeSH concepts identified in all queries
	AvgMeSHDepth	The average depth of all identified concepts

Two groups were created using the features discussed in the previous sections of this work: Experts and Laypeople

[12] The Random Forest classifier is based on the python machine learning module scikit-learn (http://scikit-learn.org/) Hyper-parameters were optimised using a grid-search approach.

Table 11 Classification results: compared to the baseline, the random forest classifier using all the features can reach an improvement of 26 % when detecting experts

Classifier	Pos. class	Acc.	Prec.	Rec.	F1
Baseline	Layp.	51.3	51.3	100.0	67.8
Positive class	Exp.	48.7	48.7	100.0	65.5
Random forest	Layp.	75.7	76.3	76.4	76.3
User behaviour features	Exp.		75.1	75.0	75.0
Random forest	Layp.	67.1	67.4	69.5	68.5
Medical features	Exp.		66.8	64.6	65.7
Random forest	Layp.	83.5	84.1	83.6	83.9
All features	Exp.		82.8	83.4	83.1

performance when using only a single group of features. Clearly for our experiments, the user behaviour features were more important than the medical ones: while the medical features marginally improved the F_1 score for each class, the user behaviour features could reach an improvement of 14 % over the baseline for detecting experts. The last row of Table 11 shows the performance of the classifier using all features. We highlight an improvement of more than 20 % over the baseline for both classes.

The Random Forest classifier also allows us to compute the Gini importance score for each feature. This value (from 0.0 to 1.0) is higher when the feature is more important, indicating how often a particular feature was selected for a split in a random forest, and how large its overall discriminative value was for the classification problem under study. We show in Fig. 9 all the features according to the Gini importance score. Befitting the

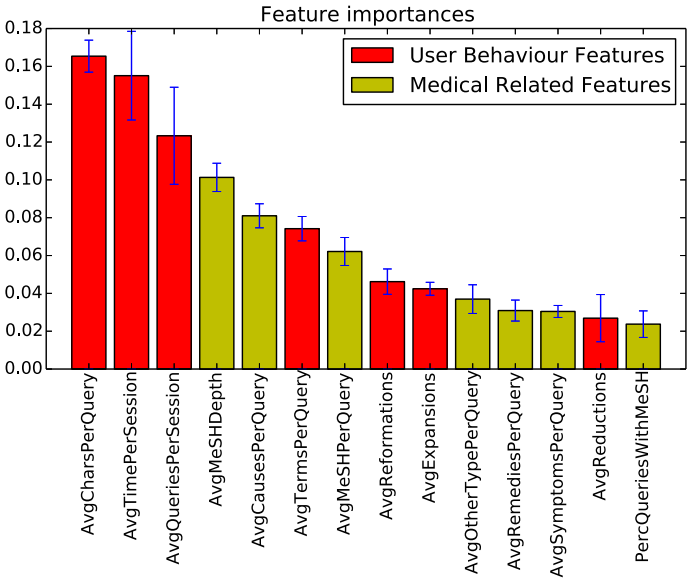

Fig. 9 Feature importance according to the Gini importance score generated by the Random Forest classifier. The *error bars* represent the standard deviation from the mean value for each feature (Color figure online)

results of Table 11, the most important features were predominantly user behaviour features.

8 Discussion

We presented the analysis of four different query logs divided into five datasets. We discuss each of the initial research questions in the next subsections, with the coverage of the third research question (relation to previously published results), covered in each subsection.

8.1 MetaMap and short queries

This study relies on the accuracy of MetaMap to enable the intent of the searchers to be identified. As MetaMap was designed for annotation of documents and not queries, we did an evaluation of its performance for short queries. Using an existing dataset of 10,000 manually annotated queries (Névéol et al. 2011), we evaluated MetaMap on two of the categories used in this paper: cause and remedy. The category symptom was not evaluated as it is not included in the dataset used. It is found that MetaMap can annotate the cause category with an F1 of 78 % and the Remedy category with an F1 of 72 %. While these values are not directly comparable to other results published, they correspond to the level of accuracy measured for related tasks: MetaMap was shown to map disease concepts in queries with an F1 of 70 % (Névéol et al. 2009), and a mapping into five classes in Palotti et al. (2014b) on 1000 queries was done with an F1 of 70 %. Most importantly, inter-annotator agreement for the manual annotation of the query corpus in Névéol et al. (2009) was 73 %. This demonstrates that the results obtained by annotating the queries by MetaMap are at the same level as those obtained by manual annotation, implying that the MetaMap annotations are sufficiently accurate for this study.

8.2 How is search conducted for medical content?

This section covers the behaviour of the users when searching for medical information. Analyses were done both at the level of individual queries and of sessions. It was found that the mean terms per query and mean chars per query were higher for experts than for laypeople with a small effect (measured by Cohen's $\underline{d}$ value). This supports the small effect also detected for these characteristics by White et al. (2009). Moving toward sessions, we found also longer sessions in both terms of mean queries per session and mean time spent per session, with a small effect, as detected in White's work. Although White et al. studied search logs from a general purpose commercial search engine, for which assumptions had to be made about the behaviour of users in order to detect experts and laypeople, we were able to find very similar effect size, including the same importance ranking for the four characteristics measured. These small effects were sufficient to successfully train a classifier to predict expert and layperson classes in White et al. (2009) and also in this paper.

When analysing the user behaviour in terms of sessions, we conclude that experts are more persistent than laypeople, as more than 10 % of the sessions in the professional search engines were composed of all possible query modification actions (expansion, reduction, reformulation). This was also found in White's work, where they noted that

sessions conducted by domain experts were generally longer than non-expert sessions and that domain experts consistently visited more pages in a session. Alternatively, longer sessions could mean that experts are struggling to find relevant information. It supports the current efforts of the information retrieval community to help experts finding scientific material to improve their clinical decisions (Roberts et al. 2014). It would be interesting to study if the increase of expertise of laypeople can change their user behaviour over time as suggested by Wildemuth (2004), but this will likely require years of search engine logs.

8.3 What are the users searching for?

The investigation of what the users search for led us to conclusions that are significantly different from results published in the literature. In both of our analyses, the one based on the MeSH hierarchy and the one based on semantic types, we observed that users are more concerned with diseases rather than symptoms, converse to what (Cartright et al. 2011) found. This difference is large: Cartright et al. found that searches for symptoms occur in 63.8 % of the sessions, while our results showed that symptoms were only in 5.5–9.1 % of sessions, depending on the search engine. In our analysis, the cause category appeared most often, in 20.9–58.2 % of the sessions, depending on the search engine. This large difference is likely due to the fact that Cartright et al. had to make assumptions about the characteristics of a medical query in order to extract medical queries from the logs of a general purpose search engine, whereas we used search logs from domain-specific medical search engines in three of the four cases. This allowed us to make the very strong assumption that users will always enter medical queries into these search engines. Understanding what users are searching for is an essential step towards providing more relevant search results.

We also identified patterns supporting the hypothetico-deductive searching processes, especially for the cause-remedy component, in which both laypeople and experts cycle through searching for causes and remedies in sessions so as to discover potential treatments for a disease. Finally, we found that TRIP users, mainly users falling into our expert class, use the hypothetico-deductive method very often, in more than 60 % of their sessions, versus <25 % for AOL and HON. This supports the hypothesis that experts have much more complex information needs, which are not well addressed by the current search systems (Roberts et al. 2014).

An interesting kind of search in the medical domain is the one for self-diagnosis purposes (Fox 2011), which often arises before consulting a medical professional (or to help the decision to consult). Previous research has shown that exposing people with no or scarce medical knowledge to complex medical language may lead to erroneous self-diagnosis and self-treatment and that access to medical information on the Web can lead to the escalation of concerns about common symptoms (e.g., cyberchondria) (White and Horvitz 2009). Also current commercial search engines are far from being effective in answering such queries (Zuccon et al. 2015), presenting on average only three highly relevant results in the top 10 results. In the same manner that experts can assist non-experts in detecting credible content on the Web (Schwarz and Morris 2011), a search system capable of inferring user expertise can learn about the decisions taken by experts to better support non-experts. In the case of self-diagnosis, the symptom-cause cycle in the expert search logs can be explored to provide query suggestions for non-experts.

After consulting a medical professional, non-experts often query about a disease or about a treatment that was recommended to them (Fox 2011). When literally copying-and-pasting the complex terms into a search box, they are presented with documents that are

potentially as complex as their queries (Goeuriot et al. 2014). Inferring user medical knowledge can help matching non-experts with the suitable documents for them even for complex queries, significantly diminishing harmful situations and misunderstandings.

8.4 What are the most useful features to infer user expertise?

We grouped the features collected in this work into two distinct sets: user behaviour features and medicine-related features. Judging by the experimental results, the user behaviour features are indispensable; and, while the medicine-related features alone were not very effective, they showed to provide large gains in all metrics, when combined with the user behaviour ones.

When analysing the features through the Gini importance coefficient, the average MeSH depth was considered the best medical feature by the classifier, a feature that was also highly ranked in Palotti et al. (2014a). For the user behaviour features, the main four metrics analysed here were also important in White et al. (2009), while features based on query modification in a session seemed not to be well used by the classifier.

9 Conclusion

In this paper, we conducted a detailed study of medical information search behaviour through query logs. We studied how users search for medical documents, as well as what they search for. Results were compared to those in published studies analysing search logs in the medical domain. Almost all recent studies about the behaviour of searchers looking for medical information have been based on the search logs of a large commercial general purpose search engine. This paper performs the important task of reproducing these studies as far as possible on search logs from other search engines to find out to what extent these results can be supported or not. An important difference with this study compared to published studies is the use, in three of the four cases, of domain-specific medical search engines targeted at either experts or laypeople, meaning that we have very strong priors about who is using the search engines and what they are searching for. This avoids assumptions that have to be made in order to extract medical queries or extract expert or laypeople queries from the search log of a general purpose search engine.

Our results support those published in the literature for the following outcomes: (1) It is possible to distinguish between medical experts and laypeople based on search behaviour characteristics; (2) experts issue more queries and modify their queries more often, meaning that they can be either more persistent than laypeople or that their information need is more complex and more difficult to reach.

A large difference with respect to what is published in the literature was found for what the users are searching for. Our analysis showed that diseases were the focus of the largest number of sessions (20.9–58.2 %), as opposed to symptoms (63.8 % in Cartright et al. 2011). We suggest that this difference is mainly due to the criteria used to extract medical queries from the search logs of a general purpose search engine, which skewed the results toward symptoms. This result suggests that the occurrence of Cyberchondria (White and Horvitz 2009) is less prevalent, especially on domain-specific medical search engines. A further result from this study that is potentially useful for search systems is the study of features for distinguishing experts and laypeople, showing that although the behavioural

features were the most discriminative ones, the combination of behavioural features with medicine-related features reached the best results.

One of the limitations of this study is the lack of clickthrough information, which would have allowed us to perform a more detailed analysis of search behaviour. A further limitation is that MetaMap can only annotate English text. Laypeople in particular prefer to query in their own language, as is clear from the high number of non-English queries that were removed from the HON search logs for this study. There is certainly a vast amount of work to be done for supporting such a query analysis for languages other than English, in particular due to the lack of such detailed language resources for many languages. MeSH on the other hand exists for many languages and mapping tools do exist. Still, detecting language of very short queries is not easy to do, so a multilingual scenario has many additional challenges.

The results of our analysis can be used to better understand the users through building detailed user profiles based on user behaviour in order to provide users with documents and query suggestions suited to their level of expertise. We can also identify new features for improving a search engine, such as the suggestions arising from this analysis to add a filter or facets for body regions to the GoldMiner search engine.

By using logfiles of several domain-specific medical search engines, this paper explores complementary information to most analyses of medical log files that either use general search engine logs or PubMed logfiles. This allows us to obtain information on user groups in a different way compared to general search engines where assumptions have to be made that can influence the analysis of the group behaviour.

Acknowledgments This research was partly funded by the European Union Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 257528 (KHRESMOI), partly funded by Horizon 2020 program (H2020-ICT-2014-1) under Grant Agreement No. 644753 (KCONNECT), and partly funded by the Austrian Science Fund (FWF) Project No. I1094-N23 (MUCKE).

References

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program (pp. 17–21).

Aronson, A. R., & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. In *Proceedings of the AMIA annual symposium* (pp. 485–489).

Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G., Nelson, S. J., Rindflesch, T. C., & Wilbur, W. J. (2000). The NLM Indexing Initiative (pp. 17–21), Lister Hill National Center for Biomedical Communications (LHNCBC), National Library of Medicine, Bethesda, MD 20894, USA.

Aronson, A. R., & Lang, F. (2010). An overview of metamap: Historical perspective and recent advances. *JAMIA*, *17*(3), 229–236.

Bhavnani, S. K. (2002). Domain-specific search strategies for the effective retrieval of healthcare and shopping information. In *CHI '02 extended abstracts on human factors in computing systems* (pp. 610–611), CHI EA '02. ACM.

Boyer, C., Baujard, V., & Geissbuhler, A. (2011). Evolution of Health Web certification through the HONcode experience. *Studies in Health Technology and Informatics*, *169*, 53–57.

Brenes, D. J., & Gayo-Avello, D. (2009). Stratified analysis of AOL query log. *Information Sciences*, *179*(12), 1844–1858.

Cartright, M.-A., White, R. W., & Horvitz, E. (2011). Intentions and attention in exploratory health search. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 65–74), SIGIR '11, New York, NY, USA, ACM.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). London: Routledge.

Cole, M. J., Gwizdka, J., Liu, C., Belkin, N. J., & Zhang, X. (2013). Inferring user knowledge level from eye movement patterns. *Information Processing and Management*, *49*(5), 1075–1091.

Collins-Thompson, K., Bennett, P. N., White, R. W., de la Chica, S., & Sontag, D. (2011). Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 403–412), CIKM '11, New York, NY, USA, ACM.

Demner-Fushman, D., Humphrey, S. M., Ide, N. C., Loane, R. F., Mork, J. G., Ruch, P., Ruiz, M. E., Smith, L. H., Wilbur, W. J., & Aronson, A. R. (2007). Combining resources to find answers to biomedical questions. In *Proceedings of the sixteenth text retrieval conference, TREC 2007, Gaithersburg, Maryland, USA, November 5–9, 2007*.

Denny, J. C., Smithers, J. D., Miller, R. A., & Spickard, A. (2003). "Understanding" medical school curriculum content using KnowledgeMap. *Journal of the American Medical Informatics Association*, *10*(4), 351–362.

Duarte Torres, S., Hiemstra, D., & Serdyukov, P. (2010). Query log analysis in the context of information retrieval for children. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval* (pp. 847–848), New York, ACM.

Duggan, G. B., & Payne, S. J. (2008). Knowledge in the head and on the web: Using topic expertise to aid search. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 39–48), CHI '08.

Eurobarometer. (2014). European citizens' digital health literacy. Technical report, European Commision.

Fox, S. (2011). Health topics. Technical report, The Pew Internet & American Life Project.

Fox, S., & Duggan, M. (2013). Health online 2013. Technical report, The Pew Internet & American Life Project.

Gayo-Avello, D. (2009). A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, *179*(12), 1822–1843.

Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G. J. F., & Müller, H. (2014). ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred Health Information Retrieval. In *Working notes for CLEF 2014 conference, Sheffield, UK, September 15–18, 2014* (pp. 43–61).

He, D., & Göker, A. (2000). Detecting session boundaries from web user logs. In *Proceedings of the BCS-IRSG 22nd annual colloquium on information retrieval research* (pp. 57–66).

Herskovic, J., Tanaka, L., Hersh, W., & Bernstam, E. (2007). A day in the life of PubMed: Analysis of a typical day's query log. *Journal of the American Medical Informatics Association*, *14*(2), 212–220.

Hollink, V., Tsikrika, T., & de Vries, A. P. (2011). Semantic search log analysis: A method and a study on professional image search. *Journal of the American Society for Information Science and Technology*, *62*(4), 691–713.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the Association for Information Science and Technology*, *44*, 161–174.

Islamaj Dogan, R., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed® user search behavior through log analysis. *Database*, *2009*, bap018.

Jadhav, A. S., Sheth, A. P., & Pathak, J. (2014). Online information searching for cardiovascular diseases: An analysis of mayo clinic search query logs. *Studies in Health Technology and Informatics*, *205*, 702–706.

Jansen, B. J., & Spink, A. (2006). How are we searching the world wide web?: A comparison of nine search engine transaction logs. *Information Processing and Management*, *42*(1), 248–263.

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, *32*(1), 5–17.

Jansen, B., Spink, A., & Taksai, I. (2008). *Handbook of research on web log analysis. Information science reference*. Hershey, PA: IGI Global Publishing.

Jones, R., & Klinkner, K. L. (2008). Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on information and knowledge management* (pp. 699–708), CIKM '08, New York, NY, USA, ACM.

Kritz, M., Gschwandtner, M., Stefanov, V., Hanbury, A., & Samwald, M. (2013). Utilization and perceived problems of online medical resources and search tools among different groups of european physicians. *Journal of Medical Internet Research*, *15*(6), e122.

Lacroix, E.-M., & Mehnert, R. (2002). The US National Library of Medicine in the 21st century: Expanding collections, nontraditional formats, new audiences. *Health Information and Libraries Journal*, *19*(3), 126–132.

Lui, M., & Baldwin, T. (2012). Langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations* (pp. 25–30), ACL '12, Stroudsburg, PA, USA, Association for Computational Linguistics.

Meats, E., Brassey, J., Heneghan, C., & Glasziou, P. (2007). Using the Turning Research Into Practice (TRIP) database: How do clinicians really search? *Journal of the Medical Library Association*, *95*(2), 156–163.

Névéol, A., Kim, W., Wilbur, W. J., & Lu, Z. (2009). Exploring two biomedical text genres for disease recognition. In *Proceedings of the workshop on current trends in biomedical natural language processing* (pp. 144–152), BioNLP '09, Stroudsburg, PA, USA, Association for Computational Linguistics.

Névéol, A., Dogan, R. I., & Lu, Z. (2011). Semi-automatic semantic annotation of pubmed queries: A study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics*, *44*(2), 310–318.

NLM. (2009). *UMLS reference manual*. Bethesda (MD): National Library of Medicine (US).

Palotti, J., Hanbury, A., & Muller, H. (2014a). Exploiting health related features to infer user expertise in the medical domain. In *Proceedings of WSCD workshop on web search and data mining*. Wiley.

Palotti, J., Stefanov, V., & Hanbury, A. (2014b). User intent behind medical queries: An evaluation of entity mapping approaches with metamap and freebase. In *Proceedings of the 5th information interaction in context symposium* (pp. 283–286), IIiX '14, ACM.

Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G. J. F., Lupu, M., & Pecina, P. (2015). ShARe/CLEF eHealth Evaluation Lab 2015, Task 2: User-centred Health Information Retrieval. In *Working notes for CLEF 2015 conference, Toulouse, France, September 8–11, 2015*.

Pass, G., Chowdhury, A., & Torgeson, C. (2006). A picture of search. In *Proceedings of the 1st international conference on scalable information systems*, InfoScale '06, New York, NY, USA, ACM.

Pratt, W., & Yetisgen-Yildiz, M. (2003). A study of biomedical concept identification: Metamap vs. people. In *AMIA annual symposium proceedings* (Vol. 2003, pp. 529–533). American Medical Informatics Association.

Roberts, K., Simpson, M., Demner-Fushman, D., & Voorhees, E., Hersh, W. (2014). State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS Track.

Schwarz, J., & Morris, M. (2011). Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1245–1254), CHI '11, New York, NY, USA, ACM.

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, *33*(1), 6–12.

Silvestri, F. (2010). Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, *4*(1:2), 1–174.

Spink, A., Yang, Y., Jansen, J., Nykanen, P., Lorence, D. P., Ozmutlu, S., et al. (2004). A study of medical and health queries to web search engines. *Health Information and Libraries Journal*, *21*(1), 44–51.

Tsikrika, T., Müller, H., & Kahn, C., Jr. (2012). Log analysis to understand medical professionals' image searching behaviour. In *Medical Informatics Europe*.

Walsh, T. M., & Volsko, T. A. (2008). Readability assessment of internet-based consumer health information. *Respiratory Care*, *53*(10), 1310–1315.

Wang, L., Wang, J., Wang, M., Li, Y., Liang, Y., & Xu, D. (2012). Using Internet search engines to obtain medical information: A comparative study. *Journal of Medical Internet Research*, *14*(3), e74.

Weeber, M., Klein, H., Aronson, A. R., Mork, J. G., de Jong van den Berg, L. T. W., & Vos, R. (2000). Text-based discovery in biomedicine: The architecture of the dad-system. In *Proceedings of the AMIA symposium* (pp. 903–907).

White, R. W. & Horvitz, E. (2012). Studies of the onset and persistence of medical concerns in search logs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 265–274), SIGIR '12, New York, NY, USA, ACM.

White, R. W. , Dumais, S. T., & Teevan, J. (2009) Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining* (pp. 132–141), WSDM '09, New York, NY, USA, ACM.

White, R. W., & Horvitz, E. (2009). Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems*, *27*(4), 23:1–23:37.

Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the Association for Information Science and Technology*, *55*(3), 246–258.

Yan, X., Lau, R. Y., Song, D., Li, X., & Ma, J. (2011). Toward a semantic granularity model for domain-specific information retrieval. *ACM Transactions on Information Systems*, *29*(3), 151–1546.

Younger, P. (2010). Internet-based information-seeking behaviour amongst doctors and nurses: A short review of the literature. *Health Information and Libraries Journal*, *27*(1), 2–10.

Zhang, X., Cole, M., Belkin, N. (2011). Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 1225–1226), SIGIR '11, ACM.

Zhang, Y. (2014). Searching for specific health-related information in MedlinePlus: Behavioral patterns and user experience. *Journal of the Association for Information Science and Technology*, *65*(1), 53–68.
Zuccon, G., Koopman, B., Palotti, J. (2015) Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In *Advances in information retrieval* (pp. 562–567). Springer.