Interactive Exploration of Healthcare Queries

Alexandros Bampoulidis
TU Wien
Favoriten Strasse 9-11/188
Vienna, AT
bampoulidis@ifs.tuwien.ac.at

João Palotti
TU Wien
Favoriten Strasse 9-11/188
Vienna, AT
palotti@ifs.tuwien.ac.at

Jon Brassey
TRIP Database Ltd.
Little Maristow, Glasllwch Lane
Newport, UK
jon.brassey@tripdatabase.com

Mihai Lupu
TU Wien
Favoriten Strasse 9-11/188
Vienna, AT
lupu@ifs.tuwien.ac.at

Sokratis Metallidis
TU Wien
Favoriten Strasse 9-11/188
Vienna, AT
metallidis@ifs.tuwien.ac.at

Allan Hanbury
TU Wien
Favoriten Strasse 9-11/188
Vienna, AT
hanbury@ifs.tuwien.ac.at

***Abstract*—Healthcare related queries are a treasure trove of information about the information needs of domain users, be they patients or doctors. However, unlike general queries, in order to make the most out of the information therein, such queries have to be processed within a medical terminology annotation pipeline. We show how this has been done in the context of the KConnect project and demonstrate an interactive query log exploration interface that allows data analysts and search engineers to better understand their users and design a better search experience.**

I. Introduction

The healthcare sector consists of many stakeholders, including the pharmaceutical and medical products industries, healthcare providers, health insurers, clinicians and patients. Each of them has specific information needs. What these specifically are, in specific circumstances and for specific purposes, is impossible to identify by exhaustive user studies. Instead, as in the case of the general web domain, we have to rely on log data: the traces left behind by users searching for healthcare related information on the web.

The analysis of search logs of medical search engines allows assumptions to be made about the user, such as whether the user has a good knowledge of the medical domain (e.g. a medical professional or expert patient) or has less knowledge in the medical domain (e.g. a patient that has just received a diagnosis) [1]. There is a higher probability that users with less knowledge in the medical domain will become overly concerned by what they find through their searches, which can also be detected from the search logs [2]. The information gained from analysing a user's search session can be used to adapt the search results to better fit the user's profile. The results of analysing the search logs over a long period of time, including the links that the users have clicked on, can be used to improve the search results. At present, search log analysis is conducted in one language at a time, and the majority of the published results are about analyses of search logs in English. For a website that allows queries in multiple languages (as Health on the Net[1] currently does and TRIP Database plans to do), this is very limiting, as the overall view of the user behaviour can only be obtained for each language separately, and maintaining the analysis software for each language separately is an unnecessary overhead. The medical domain has the advantage that the majority of search terms are in one of the many terminologies available. This is why, in the healthcare domain, before analysing logs, it is useful to enrich them. This demo will showcase the backend annotation architecture and frontend analysis interface developed in the context of the KConnect project[2].

II. Related Work

As soon as modern search engines appeared, the first studies on query logs started. For instance, Jansen et al. [3] and Silverstein et al. [4] analysed the logs from Excite and Altavista respectively, popular search engines at that time. These early works were important to help guiding the further development of search systems. For example, they suggested that the vast majority of users issue only a single query per session and rarely access any result page beyond the first one.

From the early 2000's, we highlight the work of Spink et al. [5], 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.

More recently, White and Horvitz have a series of papers analysing medical query logs, e.g. [6], [7]. They studied how users start looking for a simple symptom and end up searching for a serious disease, a phenomenon they named

[1] https://www.healthonnet.org/
[2] http://www.kconnect.eu

978-1-4673-8695-1/16/$31.00

cyberchondria. They used the logs of the Windows Live Toolbar to obtain their data and lists of keywords, such as the list of diseases from ICD-10 [3], to annotate symptoms and diseases in queries, while we used a GATE pipeline (see Section III-B) to do the same.

Specialised medical search engines are also often studied. Herskovic et al. [8] 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. [9] 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. Meats et al. [10] 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, under-utilising the search engine features.

We present here an interface that allows users, either researchers or search engine providers, to easily analyse query logs, finding trends, patterns and comparisons with other search engines. The tool described in the next section can be integrated in many of the use cases of query logs in search systems, such as query expansion, query suggestion or spelling correction. A complete overview of the previous 20 years of research on query log analysis and its potential applications is well described by Silvestri [11].

III. Data and System

Before showing the results, let us briefly describe the input data and the processing architecture.

A. Query Logs

In this work, we explore the search query logs provided by the TRIP Database medical search engine [4]. These search logs contain about 1.02 million entries made by users of the search engine between January 2014 and February 2015.

Around 93 thousand ($\sim 9\%$) of the queries were made by users who had logged in, from which we have access to their profile, and we could extract information about their medical specialty and country of origin. Some examples of entries are shown in Table I. The *Timestamp* column is the timestamp of when the user (identified by the *UID* column) clicked the search result (*URL* column) after issuing a query (*Query* column). The URL refers to a document or a web page (the full URL was omitted from the table) with its title being the column *Title*. The *Specialty* and *Country* are information available only for registered users who are logged in.

[3]International Classification of Diseases 10th Edition

[4]https://www.tripdatabase.com/

B. Enriching the Query Logs

Mapping the user queries to known controlled vocabularies can provide a better understanding of the user intent behind a query. In the medical domain, the most well established of these is the Unified Medical Language System Metathesaurus (UMLS [12]). For each entry in the query logs, we mapped both user query and document title to concepts in UMLS using the GATE [13] annotation pipeline especially developed for the KConnect project [14].

The GATE pipeline parses the text of the query and the title and enriches the text that it recognizes as medical with the following annotations:

- Annotation class: Either Anatomy, Disease, Drug or Investigation
- CUI: Concept Unique Identifier [15]
- TUI: Semantic Type Unique Identifier [5]

We show in Table II the annotations extracted from the examples of Table I.

C. Architecture

We present the system's architecture in Figure 1. A Spring[6] application was developed using the Spring Boot suite with an embedded Apache Tomcat web server.

The user logs into the system, through a security system which isolates access only to her queries. She chooses a comma separated values file (CSV) that contains the query logs for upload, specifies its structure and the date format, and uploads the file.

The Spring application takes care of the uploading process and creating of all internal configuration files for the next steps. Next, the query logs are processed by the GATE annotation pipeline and enriched as explained in Section III-B We then use the Elasticsearch/Logstash/Kibana (ELK) stack [7] for parsing (Logstash), indexing (Elasticsearch) and analysis (Kibana). On the top of that, a user friendly frontend was developed using HTML/CSS/Javascript.

IV. Results

Generally, any of the fields of the supplied data can be visualized, with the rest being adjustable parameters. Possible visualizations are area, line, pie and vertical bar charts. The following are some interesting examples.

Figure 2 depicts the occurrence of certain keywords over the period of 01/2014 – 02/2015. From this example, we can identify some trends. For instance, we may assume that there was an Ebola virus epidemic going on this period, peaking in October 2014. This is in fact true. Also, the keyword "influenza" occurs in the autumn and winter months more than any other months.

However, these might not be actual trends, because the majority of the queries might have been made by one specific

[5]The list of TUIs is available at https://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt

[6]http://spring.io/

[7]https://www.elastic.co/

TABLE I
SOME EXAMPLES OF ENTRIES EXTRACTED FROM THE QUERY LOGS (ACTUAL VALUES MAY BE CHANGED FOR PRIVACY REASONS)

Timestamp	UID	URL	Title	Query	Specialty	Country
2014-01-03 09:19:41	1	www.allergy.org.au	Position paper on hereditary angioedema	cinryze	Pharmacist	GB
2014-01-05 20:28:45	2	dx.doi.org	Interventions for female pattern hair loss	alopecia	Physician	SE
2014-01-07 07:16:33	3	www.ncbi.nlm.nih.gov	The spine surgery waiting place.	discectomy	Researcher	US
2014-01-10 13:33:31	4	www.wemerec.org	Eye health	viral conjunctivitis	Pharmacist	ES
2014-01-16 07:14:38	5	www.youtube.com	Prevention of Pressure Ulcers	decubitus	Nurse	GR
2014-01-20 19:51:19	6	emedicine.medscape.com	Agranulocytosis (Overview)	thenalidine	Pharmacist	CH
2014-01-23 08:47:55	7	www.ncbi.nlm.nih.gov	Cardiovascular effects of gliptins.	gliptins	Physician	IN
2014-03-14 02:03:29	8	pathways.nice.org.uk	Dementia	steroid psychosis	-	-
2014-08-05 16:24:10	9	emedicine.medscape.com	Hand, Anatomy	Flexor digiti minimi longus	-	-
2014-09-20 22:55:07	10	www.ncbi.nlm.nih.gov	Obesity	victoza	-	-
2014-12-11 16:27:04	11	www.ncbi.nlm.nih.gov	Nephritis, Lupus (Overview)	Mycophenolic acid	-	-

TABLE II
ANNOTATIONS EXTRACTED FROM THE EXAMPLES OF TABLE I

Anatomy	**Disease**	**Drug**	**Investigation**
spine;C0037949;T023	hereditary angioedema;C0019243;T047	cinryze;C2366371;T121	surgery;C0543467;T060
eye;C0015392;T023	alopecia;C2748784;T033	thenalidine;C0076382;T121	
cardiovascular;C0007226;T022	hair loss;C0002170;T033	gliptins;C1827106;T121	
flexor;C1879367;T029	viral conjunctivitis;C0009774;T047	victoza;C2732208;T121	
hand;C0018563;T023	pressure ulcers;C0011127;T047	mycophenolic acid;C0026933;T195	
	agranulocytosis;C0001824;T047		
	steroid psychosis;C2363722;T048		
	dementia;C3277719;T033		
	obesity;C3280128;T033		
	nephritis;C0027697;T047		
	lupus;C0409974;T047		

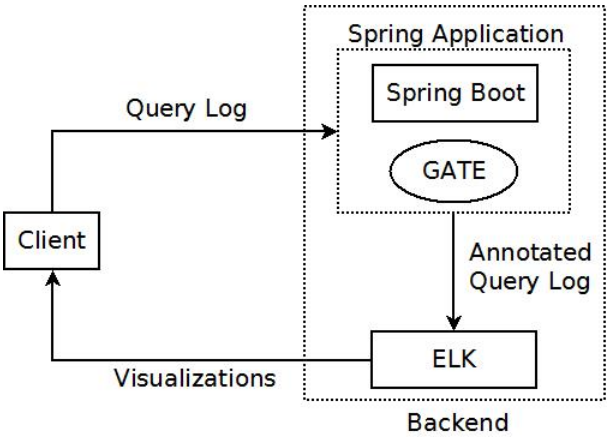

Fig. 1. System Architecture

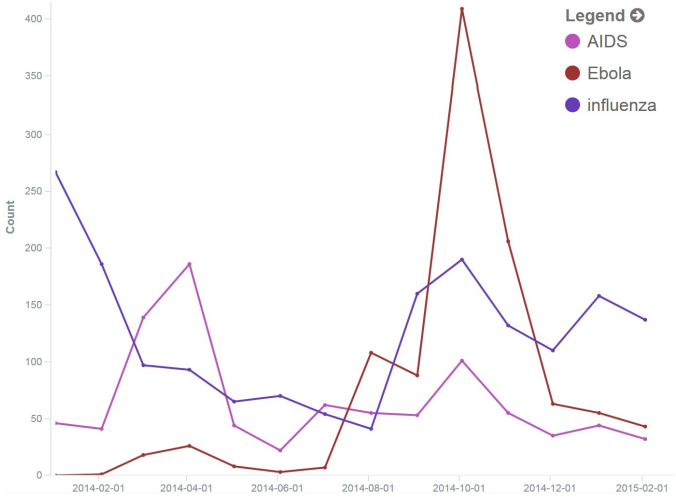

Fig. 2. Occurrence of the keywords AIDS, Ebola and influenza in the users' queries

user only. In order to find out if these are actual trends, a visualization of the user sessions or IDs using a specific keyword can be produced. Such is the case of "AIDS" in this example, as 80% of the queries were made by one specific user at the time of its peak in April 2014 (Figure 3).

Figures 4 and 5 depict the usage of medical keywords in the users' queries. These figures help us get an insight into what the users are querying about and how, better understand the queries and improve search results.

Figure 4 illustrates the most common diseases that concern the dietitian users of the search engine. Supervised learning can be used to infer user specialty, as done for user expertise in [1], [6]. Then, unregistered users can be classified into a specialty, and their search results can be adapted according to their specialty. The same approach can be used for other fields, such as country.

Figure 5 illustrates the most common drugs searched when "pneumonia" is in the users' queries. We observe the usual suspects (antibiotics, both as a generic term and as a set of specific names), but also corticosteroids, which are used in severe cases, and proton pump inhibitors, which are often

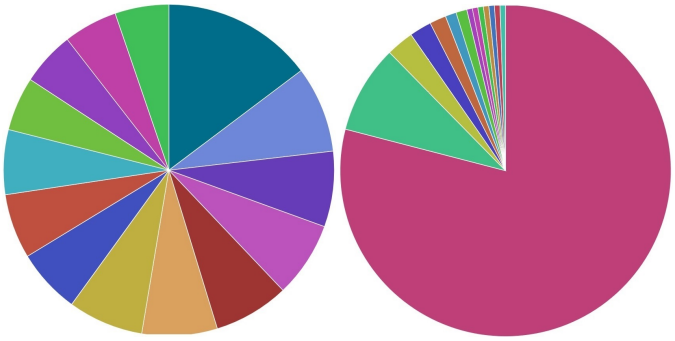

Fig. 3. Distribution of session IDs by number of queries for Ebola (left) and AIDS (right) in their peak months

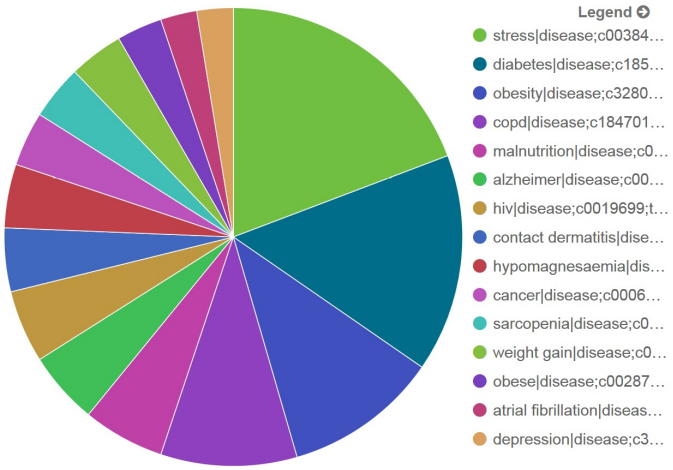

Fig. 4. Most searched diseases by dietitians

administered to counteract the effects of the antiobiotics on the gastro-intestinal tract. The information we gain by analyzing this chart can be used for a query recommendation system, in order to improve the search results and overall search experience.

V. Conclusion

Users of healthcare information have specific needs. Understanding these needs, as in the case of web search users in general, allows us to devise better algorithms, and ultimately better systems to answer those needs. Unlike general web search requests however, healthcare queries contain specific terminologies. Leveraging existing resources, we have created a healthcare query analysis pipeline and interface that allows us to understand different types of concepts related to different types of users. In turn, this helps us and the community at large to develop domain-specific search systems.

Acknowledgment

This research was partially supported by the KConnect (H2020, Grant Agreement No.: 644753) and ADmIRE (FWF, Project No.: P25905-N23) projects, funded respectively by the European Commission and the Austrian Science Fund.

pneumonia|disease;c0810014;t047

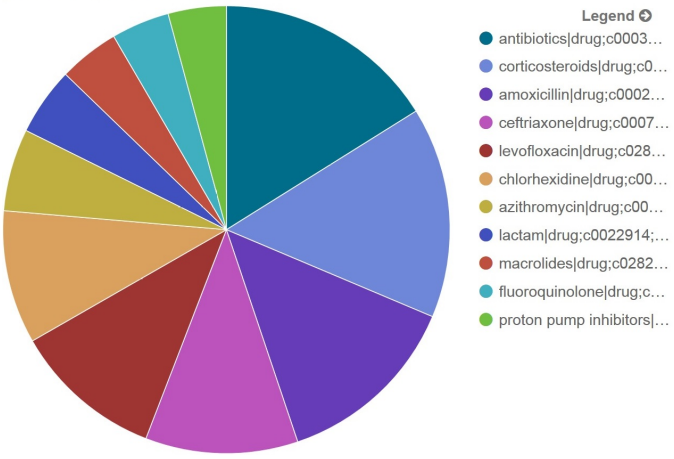

Fig. 5. Drugs related to pneumonia

References

[1] J. Palotti, A. Hanbury, H. Müller, and C. E. Kahn, "How users search and what they search for in the medical domain," *Information Retrieval Journal*, 2015.

[2] R. W. White and E. Horvitz, "On 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*, ser. SIGIR '12, 2012.

[3] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real life information retrieval: a study of user queries on the web," *SIGIR Forum*, 1998.

[4] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," *SIGIR Forum*, 1999. [Online]. Available: http://doi.acm.org/10.1145/331403.331405

[5] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D. P. Lorence, S. Ozmutlu, and H. C. Ozmutlu, "A study of medical and health queries to web search engines," *Health Information & Libraries Journal*, 2004. [Online]. Available: http://dx.doi.org/10.1111/j.1471-1842.2004.00481.x

[6] R. W. White and E. Horvitz, "Cyberchondria: Studies of the escalation of medical concerns in web search," *ACM Transactions on Information Systems*, 2009.

[7] ——, "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*, ser. SIGIR '12, 2012. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348322

[8] J. Herskovic, L. Tanaka, W. Hersh, and E. Bernstam, "A Day in the Life of PubMed: Analysis of a Typical Day's Query Log," *Journal of the American Medical Informatics Association*, 2007.

[9] R. Islamaj Dogan, G. C. Murray, A. Névéol, and Z. Lu, "Understanding PubMed ® user search behavior through log analysis," *Database*, 2009.

[10] E. Meats, J. Brassey, C. Heneghan, and P. Glasziou, "Using the Turning Research Into Practice (TRIP) database: how do clinicians really search?" *Journal of the Medical Library Association*, 2007.

[11] F. Silvestri, "Mining query logs: Turning search usage data into knowledge," *Foundations and Trends in Information Retrieval*, 2010.

[12] (2009) Umls® reference manual. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK9676/

[13] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters, *Text Processing with GATE (Version 6)*, 2011. [Online]. Available: http://tinyurl.com/gatebook

[14] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva, "Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics," *PLoS Computational Biology*, 2013.

[15] (2014, August) Unique identifiers in the metathesaurus. [Online]. Available: https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.html