Report on the SIGIR 2016 Workshop on Medical Information Retrieval (MedIR)

Lorraine Goeuriot
Univ. Grenoble Alpes
CNRS, LIG
F-38000 Grenoble France
lorraine.goeuriot@imag.fr

Steven Bedrick
Oregon Health & Sci. Univ.
Portland, OR, USA
bedricks@ohsu.edu

Gareth J.F. Jones
Dublin City University
Dublin 9, Ireland
gjones@dcu.ie

Anastasia Krithara
NCSR "Demokritos"
Athens, Greece
akrithara@iit.demokritos.gr

Henning Müller
HES-SO Valais
Sierre, Switzerland
henning.mueller@hevs.ch

George Paliouras
NCSR "Demokritos"
Athens, Greece
paliourg@iit.demokritos.gr

January 3, 2017

Abstract

The workshop on Medical Information Retrieval took place at SIGIR 2016 in Pisa, Italy on July 21. The workshop programme included seven oral presentations of refereed papers, four posters and an invited keynote presentation. This allowed time for lively discussions among the 27 participants. These made clear the significant and diverse challenges in the area of medical information retrieval and the significant interest in developing mechanisms to go about tackling them. Successfully addressing them will give added value to the wide variety of users that can profit from medical information search, including patients, general health professionals and specialist groups such as radiologists who mainly search for images and image-related information.

1   Introduction

Medical information search refers to methodologies and technologies that seek to improve access to medical information archives via a process of information retrieval (IR). Such information is now potentially accessible from many sources including the general web, social

media, journal articles, and hospital records. Health-related content is one of the most searched-for topics on the internet, and as such this is an important domain for research in information retrieval[1].

Medical information is of interest to a wide variety of users, including patients and their families, researchers, general practitioners and clinicians, and clinicians with specific expertise such as radiologists. There are several dedicated services that seek to make this information more easily accessible, such as *Health on the Net*'s medical search systems for the general public and medical practitioners[2]. Despite the popularity of the medical domain for users of search engines, and current interest in this topic within the IR research community, development of search and access technologies remains particularly challenging. One of the central challenges in medical information search is the diversity of the users of these services. In particular, users have varying categories of information needs, varying levels of medical knowledge, and varying language skills. These challenges can be summarized as follows:

1. Varying information needs: While a patient with a recently diagnosed condition will generally benefit most from simple or introductory information on the disease and its treatment, a patient living with or managing a condition over a longer term will generally be looking for more advanced information, or perhaps support groups and forums. In a similar way, a general practitioner might require basic information quickly while advising a patient, but more detailed information if deciding a course of treatment. By contrast, a specialist clinician might look for an exhaustive list of similar cases or research papers relating to the condition of a patient that they are currently seeking to advise. Understanding various types of users and their information needs is one of the cornerstones of medical information search, while adapting IR to best address these needs to develop effective, potentially personalized systems is one of its greatest challenges.

2. Varying medical knowledge: The different categories of users of medical information search systems have different levels of medical knowledge, and indeed the medical knowledge of different individuals within a category can also vary greatly. This affects the way in which individuals pose search queries to systems and also the level of complexity of information that should be returned to them or the type of support in understanding / disambiguating returned material that is required.

3. Varying language skills: Given that much of medical content is written in English, research in medical information search to date has predominantly focused on monolingual English retrieval. However, given the large number of non-English speakers on the Internet and the limited content in their native language, effective support for them to search the English sources is highly desirable. The Internet in particular has affected the patient-physician relationship and providing relevant, reliable information to patients in their own language is a key to alleviate such challenging situations and lower phenomenons such as cyberchondria.

In addition, the format, reliability, and quality of biomedical and medical information varies greatly. A single health record can contain clinical notes, technical pathology data, images, and patient-contributed anamnesis, and may be linked by a physician to research papers. The importance of health and medical topics and their impact on people's everyday

---

[1]http://www.pewinternet.org/fact-sheets/health-fact-sheet/
[2]http://www.hon.ch/

lives makes the need for retrieval of accurate and reliable information especially important. Determining the likely reliability of available information is challenging.

Finally, as with IR in general, the evaluation of medical search tools is vital and challenging. For example, there are no established or standardized baselines or evaluation metrics, and limited availability of test collections. Further discussion and progression on this topic would be beneficial to the community.

Theme and Purpose of the Workshop

The objective of the workshop was to provide a forum to enable the progression of research in medical IR to provide enhanced search services for all users with interest in medical information search. The workshop aimed to bring together researchers interested in medical information search, with the goal of identifying specific research challenges that need to be addressed to advance the state-of-the-art and to foster interdisciplinary collaborations towards solving these challenges. To enable this, we encouraged participation from researchers in all fields related to medical information search including not only IR but also natural language processing (NLP), multilingual text processing, and medical image analysis.

Topics of interest included but were not limited to:

- users and information needs;
- semantics and NLP for medical IR;
- reliability and trust in medical IR;
- personalised search;
- evaluation of medical IR;
- multilingual questions in medical IR;
- multimedia technologies in medical IR;
- the role of social media in medical IR.

2 Keynote - William Hersh

The keynote talk was given by Prof. William Hersh (Oregon Health & Science University, Portland OR, USA), on "Challenges for Information Retrieval and Text Mining in Biomedicine: Imperatives for Systems and Their Evaluation".

In his talk, Prof. Hersh described the challenges faced by patient-specific and knowledge-based IR and text mining applications. The main challenges raised in the context of patient-specific applications are related to the data: its inaccuracy and incompleteness, the difficulty in recovering it for research purpose, its insufficient granularity, and more. The main challenges raised in the context of knowledge-based applications are mainly methodological, for instance with experiments on Randomized Controlled Trials, and observational studies. The publication bias, and reproducibility issues are also mentionned. Prof. Hersh concluded his talk by proposing that system development should be driven by use cases, and their evaluation should take into account users and clinical outcomes.

3 Presented Papers

Of the eighteen papers submitted to the workshop, seven were selected for oral presentation and four as posters. The papers cover a wide range of topics, in summary the details of the accepted paper is as follows. Abomelha et al. investigate the challenges and obstacles scientists and Bioinformaticians in the field of Medical Biology face when working with a variety of databases as well as having to combine heterogeneous structured and unstructured data [1]. Beloborodov and Goeuriot argue that medical IR can be impoved by personalization and report results of a preliminary study on health personalized search, using medical reports as a personalization mean source [2]. Hanbury presents a summary of current text IR evaluation campaigns in the medical domain in terms of users, documents and tasks, and then identify areas of medical IR in which systematic large-scale evaluation remains to be done [3]. Inthiran et al. describe a small scale study examining the health querying behavior of laypeople [4]. Kuhn and Eickhoff reflect on ways to improve the quality of bio-medical IR by drawing implicit negative feedback from negated information in noisy natural language search queries [5]. Lease et al. argue for research into the cross-pollination of work in Systematic Review in medicine and e-Discovery for medical search [6]. Lelong et al. describe a semantic search engine implemented for Electronic Health Records [7]. Rahman and Bhattachrya presents a classification-driven biomedical image retrieval approach which combines multiple visual and text features using a multi-response linear regression (MLR)-based meta-learner [8]. Saleh and Pecina investigate adaptation of a supervised machine learning model for reranking of query translations to new languages for cross-language IR in the medical domain [9]. Soldaini and Goharian present QuickUMLS: a fast, unsupervised, approximate dictionary matching algorithm for medical concept extraction [10]. Yilma et al. report that the source of health information used by college students from a developing country is the Internet with their family as the secondary source of health information [11].

4 Discussion Session

The presentations at the workshop led to many interesting and stimulating topics for during the discussion session of the workshop. A major emphasis in the discussions was to stress the importance of common document collections and specific tasks for these collections to really compare approaches on clearly specified scenarios. Another important aspect discussed was the lack of communication with medical professionals. According to a few researchers regularly in touch with them, one solution is to focus and collaborate of specific and concrete problems.

5 Conclusions

We have overviewed the second SIGIR workshop on Medical Information Retrieval. The amount of interest in the workshop, both through the number of paper submissions and the large number of workshop participants (27 registrants), highlight both the activity and interest in the medical information retrieval space within the IR community. The workshop provided insights into the active areas of research and helped to identify, understand and progress the many challenges facing medical information retrieval.

6 Acknowledgment

References

[1] Faris Abomelha, Ingo Frommholz, and Paul Sant. A study of challenges and obstacles in biomedical data search. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[2] Alexander Beloborodov and Lorraine Goeuriot. Improving health consumer search with contextual information. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[3] Allan Hanbury. What remains to be evaluated in multilingual medical search? In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[4] Anushia Inthiran, Saadat M Alhashmi, and Pervaiz K Ahmed. Describing health querying behavior. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[5] Lorenz Kuhn and Carsten Eickhoff. Implicit negative feedback in clinical information retrieval. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[6] Matthew Lease, Gordon V Cormack, An T Nguyen, Thomas A Trikalinos, and Byron C Wallace. Systematic review is e-discovery in doctor's clothing. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[7] Romain Lelong, Chloé Cabot, Lina F Soualmia, and Stefan Darmoni. Semantic search engine to query into electronic health records with a multiple-layer query language. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[8] Mahmudur Rahman and Prabir Bhattachrya. Multimodal biomedical image classification and retrieval with multi response linear regression (mlr)-based meta learning. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[9] Shadi Saleh and Pavel Pecina. Adapting smt query translation reranker to new languages in cross-lingual information retrieval. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[10] Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.

[11] Tesfahun Yilma, Anushia Inthiran, and Daniel Reidpath. College students from developing countries: Where do they get health information? In *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*, 2016.