Medical Information Retrieval: Introduction to the Special Issue

5 authors, including:

Lorraine Goeuriot
University of Grenoble
63 PUBLICATIONS **746** CITATIONS

SEE PROFILE

Gareth J F Jones
Dublin City University
461 PUBLICATIONS **5,222** CITATIONS

SEE PROFILE

Henning Müller
HES-SO Valais-Wallis
803 PUBLICATIONS **12,452** CITATIONS

SEE PROFILE

Justin Zobel
University of Melbourne
425 PUBLICATIONS **13,507** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Video Mail Retrieval Using Voice View project

⬚ MD-Paedigree - Model-Driven European Paediatric Digital Repository View project

Medical Information Retrieval: Introduction to the Special Issue

Lorraine Goeuriot · Gareth J.F. Jones · Liadh Kelly · Henning Müller · Justin Zobel

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 IR research. Medical information is of interest to a wide variety of users, including patients and their families, researchers, general practitioners and clinicians, and practitioners with specific expertise such as radiologists. There are several dedicated services that seek to make this information more easily accessible, such as the 'Health on the Net' system for the general public and medical practitioners (`http://www.hon.ch/`). However, despite the popularity of the medical domain for users of search engines, and current interest in this topic within the

L. Goeuriot
LIG - Université Grenoble Alpes, 41 rue des mathématiques, 38400 Saint Martin d'Hères, France
E-mail: lorraine.goeuriot@imag.fr

G. Jones
School of Computing - Dublin City University, Glasnevin, Dublin 11, Ireland
E-mail: gareth.jones@computing.dcu.ie

L. Kelly
Trinity College Dublin - Dublin 2, Ireland
E-mail: liadh.kelly@scss.tcd.ie

H. Müller
HES-SO - TechnoArk 3, 3960 Sierre, Switzerland E-mail: henning.mueller@hevs.ch

J. Zobel
Department of Computing & Information Systems, University of Melbourne, Parkville 3010, Australia
E-mail: jzobel@unimelb.edu.au

IR research community, development of search and access technologies remains particularly challenging.

A central issue in medical IR is the diversity of the users of these services. In particular, they will have varying categories of information needs, varying levels of medical knowledge, and varying language skills. These challenges can be summarized as follows:

– 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. Similarly, a general practitioner might require basic information quickly while advising a patient, but more detailed information if deciding a course of treatment, and 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 of various types of users and their information needs is one of the cornerstones of medical IR; development of effective, potentially personalized systems that address these needs is one of the greatest challenges.
– Varying medical knowledge: The different categories of users of medical IR 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 which should be returned to them or the type of support in understanding of retrieved material which should be provided.
– Varying language skills: Given that much of medical content is written in the English language, research to date in medical information search has predominantly focused on monolingual English retrieval. However, given the large number of non-English speakers on the Internet and the lack of content in their native language, effective support for them to search English sources is highly desirable.

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 histories, and may be linked by a physician to research papers. The importance of health and medical topics and their impact on people's everyday 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.

The Special Issue

There are several regular workshops on medical natural language processing and text mining. For example, the BioNLP workshop is associated with ACL and has been running since 2002. It targets foundational research in language processing for the biological and medical domains. A shared task track is collocated with the workshop and organises text mining and fine-grained information tasks (such as genia event extraction for NfkB knowledge base construction and cancer genetics discovery). The Louhi workshop offers another forum for researchers to explore health text mining and information analysis. While these two workshops are concerned with technologies and methodologies for medical information extraction and natural language processing, the LREC workshop has focused on building and evaluating resources for biomedical text mining, which is concerned with work on resources and issues relating to their usability.

The IR community has tackled issues related to the medical domain via shared tasks over several years: the TREC genomics (2003-2007), medical records track (2011-2012), and clinical decision support (2014) tracks, Image-CLEF (2003-2015), and CLEF eHealth (2013-2015). While these tracks mainly focus on medical IR, they are each concerned with a single specific task rather than on the field more broadly.

In 2013, the SIGIR workshop on health search and discovery (helping users and advancing medicine) sought to investigate ways to make medical and health information more accessible to laypeople, and ways to discover new medical facts and phenomena from information sought online [10]. This exploratory workshop showed that the IR community has many links with the medical domains and highlighted many directions to explore. A key direction for further exploration was IR techniques for medical search. As a result, in 2014 we organised a workshop, entitled Medical Information Retrieval [2], focusing on this direction, with emphasis on its multimodal and multilingual aspects. This workshop gave an overview of the current research directions taken in the medical IR domain. The lively discussions at the workshop showed us how active and rich the domain was, and led us to set up this special issue of the Information Retrieval journal.

Overview of Papers

We received eighteen submissions for the special issue, of which eight papers were accepted. Authors of accepted papers come from Australia, Austria, France, Switzerland, and the USA. One was written by an international consortium. We now summarise these papers.
— Koopman et al. present a graph inference retrieval model aiming at improving biomedical IR [3]. Their technique aims at solving issues often encountered in the medical domain such as vocabulary mismatch, or granularity mismatch [1]. To do so, their system integrates structured knowledge re-

sources, statistical IR methods, and inference, in a system where corpora are represented by graphs and retrieval is driven by inference mechanism over the graphs. The evaluation of their system shows that inference can be effective, and identifies relevant documents that would not be retrieved by classical approaches. It also shows that inference should not be applied in all cases, but can help improve retrieval in the case of hard or complex queries.

— Kovacs et al. examine the issue of large-scale data in hospitals, using the case of radiation dosage monitoring (requiring large quantity of data, such as scans) [4]. To deal with this issue, they propose a retrieval system to search for radiation dosage data and provide visualization and mining techniques for the results. They demonstrate in the paper the utility of such visual methods for medical professionals in practice.

— Lossio-Ventura et al. propose a novel approach for biomedical terminology extraction [5]. Terminologies and other knowledge sources are crucial for systems that mine biomedical data. Their approach extracts biomedical terms and ranks them using measures based on linguistic, statistical, and structural aspects. The evaluation of their approach shows that it outperforms state-of-the-art techniques. Furthermore, it proves to be efficient in several languages, namely English, French, and Spanish.

— Markonis et al. investigate the effectiveness of relevance feedback techniques on biomedical image retrieval [6]. Image retrieval can assist physicians as a support in their daily practice, for instance while making a diagnosis. While relevance feedback has been widely used in the text retrieval field, they propose a novel method using both text and visual information. Their results show that this combined relevance feedback approach improves retrieval results.

— Roberts et al. present an overview of the TREC Clinical Decision Support (CDS) track [8]. The goal of this evaluation track is to provide access to relevant biomedical literature in clinical settings, in order to help clinicians in their practice of evidence-based medicine. The paper describes the task, the participants' submissions, and provides an analysis of the results. This analysis shows that standard approaches such as word-based indexing and pseudo-relevance feedback gave the best results. However, as this track had only run once, the authors note that the lack of training data might have influenced results and caused classical approaches to be the most efficient. Tuning of systems using annotation, negation, and attribute extraction, or document type preference, should prove useful in the second iteration of the track, once participants get access to enough training data.

— Soldaini et al. present a query clarification approach aiming at improving medical IR by lay people [9]. This approach tackles the issue of vocabulary mismatch between lay queries and expert vocabulary, preventing patients from finding relevant or authoritative information. Query clarification is a form of expansion, where the most appropriate expression (that is, the most similar expert expression) is added to the query. Using three different synonym mappings and conducting two task-based retrieval studies, they show that users are more satisfied with the results using this approach.

— Zheng and Yu aim at assisting patients in understanding their own electronic health records (EHR) [11]. To do so, they develop a system for retrieving consumer-oriented health educational material, linked to complex sections of their EHR. They investigated techniques to extract EHR notes for specific queries, namely topic models (with Latent Dirichlet Allocation) and key concept identification (with Inverse Document Frequency or Conditional Random Fields). They show that identifying of key concepts and pseudo relevance feedback gives the best results.

— Palotti et al. investigates user behavior while searching for medical information online [7]. Using multiple query logs, they categorize users by expertise and analyze their search behavior. They show that medical professionals are more persistent and interact more with the system. They also discovered that users seem to be more interested in diseases than symptoms. As an outcome of this analysis, they developed a classifier inferring user expertise, that could be used to adapt search results to the user.

Summary

Amongst the general public, health topics are one of the most common kinds of query, while medical practitioners also make extensive use of search in their work. The ability to find relevant, informative results can be critical in determining whether people seek treatment, or whether a treatment is found – factors that are helping to drive interest in medical search in the IR, NLP, and data mining communities. This interest has led to the creation of several workshops, including our own MedIR at SIGIR in 2014, and also to this special issue.

The papers in the special issue reflect the breadth of factors that might influence successful medical IR in practice. We believe that they are a valuable illustration of the scope of the field, as well as being significant contributions in their own right. We hope that you find this work as inspiring as we do.

References

1. Ely, J., Osheroff, J., Gorman, P., Ebell, M., Chambliss, M., Pifer, E., Stavri, P.: A taxonomy of generic clinical questions: classification study. British Medical Journal

321(7258), 429–432 (2000)

2. Goeuriot, L., Jones, G.J.F., Kelly, L., Müller, H., Zobel, J. (eds.): Proceedings of the SIGIR workshop on medical information retrieval (2014)

3. Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., Lawley, M.: Information retrieval as semantic inference: A graph inference model applied to medical search. Information Retrieval (2016)

4. Kovacs, W., Weisenthal, S., Folio, L., Li, Q., Summers, R.M., Yao, J.: Retrieval, visualization, and mining of large radiation dosage data. Information Retrieval (2016)

5. Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire, M.: Biomedical term extraction: Overview and a new methodology. Information Retrieval (2016)

6. Markonis, D., Schaer, R., Müller, H.: Evaluating relevance feedback techniques for medical image retrieval. Information Retrieval (2016)

7. Palotti, J., Hanbury, A., Müller, H., Kahn, C.E.: How users search and what they search for in the medical domain. Information Retrieval (2016)

8. Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., Hersh, W.: State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS track. Information Retrieval (2016)

9. Soldaini, L., Yates, A., Yom-Tov, E., Frieder, O., Goharian, N.: Enhancing web search in the medical domain via query clarification. Information Retrieval (2016)

10. White, R., Yom-Tov, E., Horvitz, E., Agichtein, E., Hersh, B. (eds.): Proceedings of the SIGIR workshop on health search and discovery (2013)

11. Zheng, J., Yu, H.: Methods for linking EHR notes to education materials. Information Retrieval (2016)