Considerations for the Development of Task-Based Search Engines

Paula Petcu
Findwise Aps.
Frederiksborggade 32
Copenhagen 1360 Denmark
paula.petcu@findwise.com

Radu Dragusin
Computer Science Dept.
University of Copenhagen
Universitetsparken 1, 2100 Denmark
dragusin@diku.dk

Abstract

Based on previous experience from working on a task-based search engine, we present a list of suggestions and ideas for an Information Retrieval (IR) framework that could inform the development of next generation professional search systems. The specific task that we start from is the clinicians' information need in finding rare disease diagnostic hypotheses at the time and place where medical decisions are made. Our experience from the development of a search engine focused on supporting clinicians in completing this task has provided us valuable insights in what aspects should be considered by the developers of vertical search engines.

1 Background

In task-based search scenarios, general search engines might not suffice in satisfying user information needs and information seeking behaviour. Searching for rare disease diagnostic hypotheses is one such case where a task-oriented IR system has been developed to adapt to the task-specific user needs.

1.1 The cognitive process of disease diagnosis

The diagnostic process is a complex, often non-linear sequence of actions taken by the clinician towards reaching the final diagnosis. However, simply viewed, the process of finding the correct disease consists of generating several diagnostic hypotheses matching the patient case, followed by an iterative process of selection, testing and elimination, after which the final diagnosis is made and corresponding treatment is identified. The clinicians select up to around 6-7 diagnostic hypotheses based on pattern matching the patient data with their medical knowledge and experience [Cam87]. Studies have shown that having a good list of diagnostic hypotheses is key to reaching correct diagnosis. It was further suggested that in many of the cases of misdiagnosis the correct disease was not included in the list of potential diagnoses [KDM08].

1.2 The task of diagnosing rare diseases

The particular difficulty of diagnosing rare diseases stems from their low prevalence (less than 1 in 2000 people being affected), large number (between 5000-8000 distinct diseases), and non-specific symptoms. Therefore, clin-

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

icians seldom encounter patients suffering from rare diseases, and considering the rate of biomedical publishing[1], might not be familiar with the latest findings. It was shown that around 40% of the rare disease patients are misdiagnosed and 25% experience diagnostic delays between 5-30 years [EUR04]. Furthermore, statistics show that there are around 30 million European citizens affected by a rare disease, and treatment is critical in many of the cases [EUR04]. This makes the diagnosing of rare diseases a difficult task for clinicians and at the same time it provides a promising ground for IR research.

1.3 Rare diseases IR

For the medical field, several professional IR systems have been developed over the years with the goal of helping medical personnel in completing their tasks at the time and place where clinical decisions are made. However, the acceptance rate of such systems is low and some studies suggest that medical personnel would rather use a familiar web search engine instead. [CF12]

Existing IR systems that are focused on rare disease retrieval have several shortcomings. General purpose search engines, such as Google, although popular amongst clinicians, cannot model the specific task, and maybe more damaging, include low quality data in their indexes. Specialised medical systems, such as Phenomizer[2] or Orphanet[3], restrict the input to a limited set of medical concepts (ICD codes or MEDLINE terminology), usually selected from a drop-down list. Such limitations could make it difficult and time consuming for clinicians to accurately input patient data. Specialised information databases with search interfaces such as PubMed[4] exhibit similar limitations, requiring the use of complex queries with boolean operators.

This work asks what design considerations are needed for the development of an IR system to support clinicians in the process of diagnosing rare diseases. To address this question, the current work looks at previous research done by the authors towards the task of diagnosing rare diseases with the help of IR technologies, and based on the findings, provides suggestions on developing similar systems for different tasks.

The rest of the paper is organised as follows. Section 2 provides an overview of previous work related to the described task. Section 3 discusses considerations for developing vertical search engines based on the authors' experience. Section 4 summarises and concludes this work.

2 Previous work

Previous work conducted by the authors consisted in the design, development and evaluation of an IR system specialised on retrieving rare and genetic diseases[5] information based on queries consisting of patient data, with the goal of helping clinicians in finding diagnostic hypotheses for difficult to diagnose cases [DPL+11].

The system, called FindZebra[6], receives symptoms, test results, or any textual data as input, and ranks medical documents based on their estimated relevance to the query. The medical documents are indexed and retrieved using the Indri open-source search engine[7]. The system provides a unified interface for searching medical documents crawled from 10 different content sources which were chosen based on specific criteria: medical articles discussing rare or genetic diseases that are curated and maintained by medical professionals or institutions.

Further work [DPLW12], intended to optimise the IR systems for the task of rare disease diagnosis, focused on mapping entities from UMLS Metathesaurus[8] medical ontologies and classifications to the indexed medical articles. With this annotation in place, the search engine has received the additional functionalities of grouping and ranking diseases rather than documents, where diseases are concepts in the ontologies of the UMLS Metathesaurus.

A task-oriented evaluation strategy has been considered. On queries consisting of patient symptoms extracted from real medical cases, we have experimentally evaluated FindZebra against the search results provided by PubMed and Google. Our findings showed that FindZebra is more suited for the task-specific requirements in terms of precision and time spent searching [DPL+11]. Google is not optimised for the characteristics of this task, and the quality of some of the content it indexes is a potential issue in professional search. On the other

[1] Around 2000-4000 medical references are published daily on MEDLINE, www.nlm.nih.gov/pubs/factsheets/medline.html
[2] compbio.charite.de/phenomizer
[3] www.orpha.net
[4] www.ncbi.nlm.nih.gov/pubmed
[5] The motivation behind including articles about genetic diseases is that around 80% of the rare diseases are of genetic origin.
[6] Zebra is sometimes used in medicine to denote a surprising diagnosis. FindZebra can be accessed at findzebra.com
[7] lemurproject.org/indri
[8] www.nlm.nih.gov/research/umls

hand, PubMed contains high-quality articles, but the interaction design is poor when it comes to solving the task of diagnosing rare diseases.

The conclusion from previous work on FindZebra is that rare disease retrieval can be seen as a distinct IR task, with its specific characteristics. (i) Clinicians' queries can consist of a long list of patient symptoms, usually longer than the average length of queries sent to general web search engines. (ii) Sometimes symptoms specific to a disease might be missing from the query, or symptoms in the queries might not be specific to the correct disease. (iii) Popularity based metrics and index pruning do not necessarily benefit the retrieval of rare disease articles.

To optimise the retrieval of rare disease information, previous work has consisted mainly in the analysis and selection of high-quality information, integration of the different content sources in one location, the design of a new vertical search engine, the analysis and selection of existing medical ontologies, the fusion of medical articles with medical concepts from ontologies, and the proposal of a task oriented evaluation.

3 Considerations in developing vertical search engines for knowledge workers

For knowledge workers, that is professionals whose work requires extensive domain-specific knowledge, access to information is of crucial importance. Medical doctors, lawyers, scientists, engineers are several such professionals. In all of their specific fields, they are expected to keep up with the increasing amounts of information. Vertical IR systems are an approach that could help knowledge workers in keeping up with the changes and take informed decisions.

Much of the existing information is already organised, classified, tagged in a domain-specific fashion. For instance, MeSH[9] is an hierarchical set of medical subject terms which can also serve as a thesaurus, and MEDLINE is a bibliographic database that includes most of the medical journals articles, many of these being manually tagged with MeSH terms. Nevertheless, medical information from different sources is mapped to several or different classification systems, such as ICD[10] or SNOMED-CT[11], making it difficult to use these mappings in a consistent way. This problem is somewhat mitigated by the UMLS Metathesaurus, which tries to map terms between over 100 medical classification systems.

Vertical search engines should accommodate the existing workflows of the knowledge workers and avoid being disruptive. In the medical domain there are various constraints related to the time and place where diagnostic decisions are made. Clinicians usually have only a few minutes for each patient and can only afford a short time to searching for diagnostic hypotheses or answers to medical questions, which need to be evaluated on the spot. Clustering the retrieved documents using domain-specific classifications and returning such clusters as results is one possible strategy for shortening search time.

As general-purpose web search engines are ubiquitous in their personal lives, knowledge workers are probably using them in the workplace as well. Indeed, especially younger clinicians are already using Google to seek answers for medical questions [FNM+09]. Nevertheless, given the impact of medical decisions, clinicians are sensitive to the reliability of the retrieved information. For this reason, when faced with medical dilemmas, many clinicians still prefer to use authoritative sources such as PubMed or consult specialised books or colleagues [CM06]. Consequently, in order to factor in the reliability of the source when making decisions, search systems should provide knowledge workers access to supporting evidence.

Some cross-language support is possible for domains where certain classifications are available in multiple languages. In the medical domain, for the ICD-10 disease classification, the concepts are available in 42 languages. Such cross-language classifications could for instance facilitate query formulation for non-native speakers.

An important aspect of working with data from numerous sources is the difficulty of linking all the information and analysing it in a unified manner. Clinicians usually have access to a wide range of health information resources through their medical libraries, but the tools they have at their disposal might not provide for a unified analysis. Given the limited time available in the clinical setting, a strategy consisting in querying multiple sources and aggregating the resulting information while taking advantage of the domain knowledge could allow for a rapid analysis of relevant information.

Finally, domain-dependent factors could inform the ranking. Number of citations, date of publishing, license, could all be features impacting the ranking of results. Furthermore, filtering on such features could save time and

[9] www.ncbi.nlm.nih.gov/mesh
[10] www.who.int/classifications/icd/
[11] www.ihtsdo.org/snomed-ct

improve task performance. For example, clinicians specialising on rare diseases might be interested in limiting the retrieval to sources that only cover rare or genetic disorders.

4 Conclusion

Diagnosing rare diseases is a task that can be potentially improved with the support of a professional IR system, and can have an important impact on the outcome of patients. Curated, high-quality information for this specific case is available. The need of a well-designed user interface is still an open research area for IR, human-computer interaction and related fields. This work however has focused on design considerations for integrating IR technologies that support the diagnostic task.

There are certainly other verticals, apart from the medical domain, that can benefit from a similar approach to the one taken for this vertical. There is an increasing amount of information made digitally available for various professions, but off-the-shelf search engines are not optimised to the specific needs of each knowledge worker. Vertical search engines can be adapted to satisfy specific information needs and support difficult tasks.

Thus, we see the potential of a framework that could simplify and speed up the development of such specialised solutions. A framework providing reusable components for vertical search engines could lower the barrier for deploying customised task-based search engines, helping knowledge workers to cope with the complexity of finding and analysing domain-specific information.

References

[Cam87] E.J. Campbell. The diagnosing mind. *Lancet*, 1(8537):849, 1987.

[CF12] R. Chisholm and J.T. Finnell. Emergency department physician internet use during clinical encounters. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1176. American Medical Informatics Association, 2012.

[CM06] H.C.H. Coumou and F.J. Meijman. How do primary care physicians seek answers to clinical questions? a literature review. *Journal of the Medical Library Association*, 94(1):55, 2006.

[DPL+11] R. Dragusin, P. Petcu, C. Lioma, B. Larsen, H. Jørgensen, and O. Winther. Rare disease diagnosis as an information retrieval task. *Advances in Information Retrieval Theory*, pages 356–359, 2011.

[DPLW12] R. Dragusin, P. Petcu, C. Lioma, and O. Winther. Zebra: Searching for rare diseases a case of task-based search in the medical domain. *Proceedings of the ECIR 2012 Workshop on Task-Based and Aggregated Search (TBAS2012)*, page 36, 2012.

[EUR04] EURORDIS. Eurordiscare2: Survey of diagnostic delays, 8 diseases, Europe, 2004.

[FNM+09] M.E. Falagas, F. Ntziora, G.C. Makris, G.A. Malietzis, and P.I. Rafailidis. Do PubMed and Google searches help medical students and young doctors reach the correct diagnosis? a pilot study. *European journal of internal medicine*, 20(8):788–790, 2009.

[KDM08] O. Kostopoulou, B.C. Delaney, and C.W. Munro. Diagnostic difficulty and error in primary care - a systematic review. *Family practice*, 25(6):400–413, 2008.