Zebra: Searching for Rare Diseases A Case of Task-Based Search in the Medical Domain

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

Paula Petcu
Findwise Aps.,
Valkendorfsgade 13A,
Copenhagen 1151, Denmark
paula.petcu@findwise.com

Christina Lioma
Computer Science Dept.,
University of Copenhagen,
Njalsgade 128, 2300,
Denmark
liomca@gmail.com

Ole Winther
DTU Informatics, Technical
University of Denmark
2800 Lyngby, Denmark
owi@imm.dtu.dk

ABSTRACT

Task-based search addresses situations where standard off-the-shelf Information Retrieval (IR) technology may not suffice to satisfy users in their tasks. In these situations, IR systems should be tailored to the user's task-specific needs and requirements. One such task is searching for rare disease diagnostic hypotheses in the domain of medical IR.

In this work, we build upon an existing vertical medical search engine, Zebra, that is focused on rare disease diagnosis. In previous work, Zebra has been evaluated using real-life medical cases of rare and difficult diseases, and has been found to be a useful and competitive tool for clinicians. In this work, we extend Zebra's functionalities to optimise the task of medical diagnosis through search as follows: we add the option of grouping retrieved documents into clusters based on disease name occurrence, and we offer a 'disease-ranking' option, in addition to the standard 'document-ranking' option. This paper presents and discusses these functionalities.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Algorithms, Experimentation, Performance

Keywords

Task-Based Search, Rare Diseases, Clinical Information Retrieval

1. INTRODUCTION

Rare disease definition.

Diseases with a prevalence lower than 1 case per 2000 people are classified as rare in Europe. By this classification, there are between 5000 and 8000 distinct rare diseases, collectively affecting around 30 million EU citizens [8]. Rare diseases, 80% of which have a genetic origin, are often hard to diagnose, and as a consequence patients can experience long diagnostic delays and misdiagnosis. A study in Europe shows that 40% of rare disease patients are misdiagnosed, and 25% wait between 5 to 30 years before the correct diagnosis is reached [8]. The difficulty in diagnosing rare diseases stems from their large number, low prevalence, and non-specific symptoms.

It is reasonable to assume clinicians to be unfamiliar with many of the rare diseases since, in their entire career, they will probably only encounter a few of them. Especially in difficult cases, it is common for a clinician to select at most five or six diagnostic hypotheses for further investigation [2]. If the correct disease is short-listed at this step, the diagnostic outcome will probably be successful [9].

Task-based search for rare diseases.

When confronted with difficult cases, clinicians would traditionally resort to using medical books or journals, or consult with more experienced colleagues. However, given the restrictions of the clinical setting, they are increasingly using computer systems to aid them in finding the right answers at the time and place where medical decisions are made [10]. This shift is backed by the fact that Information Retrieval (IR) systems are very good at matching queries to large corpora, whereas clinicians are good at filtering unsuitable results. Nevertheless, the use of a task-based search engine as opposed to a general purpose one would better fit the clinician's task-specific needs and requirements, being tailored for the work-flow and time restrictions of the diagnostic process.

As multiple studies involving medical personnel revealed [11, 6], the most extensively used web systems in finding medical answers are Google[1] and PubMed[2]. Google is preferred mainly because of its familiarity and ease of use, and PubMed is chosen due to its reliable content [12]. However, considering the time restrictions of the clinical setting, many of the questions still remain unanswered [5]. While having the right information is crucial when making diagnostic decisions, filtering through Google's results or formulating a PubMed query can be time consuming [7].

There are a few web systems, such as Phenomizer[3] and Orphanet[4], that allow clinicians to find rare or genetic diseases fitting a set of clinical signs. However, the input for these systems has to be selected from a predefined list of clinical signs. Moreover, both of these systems return results from a single source of medical articles.

Our work on the Zebra search engine.

This work is an extension of our continuous research and development efforts to build a state-of-the-art medical search engine for rare disease diagnosis by clinicians [4, 3]. We have developed Zebra, a task-based search engine for rare disease diagnosis that takes free text as input, and returns diagnostic results from multiple specialised and expertly curated sources of medical articles. This work describes an extension to Zebra's functionalities that optimises the task of medical diagnosis through search as follows: we add the option of grouping retrieved documents into clusters based on disease name occurrence, and we offer a 'disease-ranking' option, in addition to the standard 'document-ranking' option.

The rest of the paper is organised as follows. Section 2 provides an overview of the Zebra search engine and presents the new functionalities. Section 3 summarises and concludes this work.

2. ZEBRA SEARCH ENGINE

Given a medical patient case involving a rare or very difficult disease, we consider the task of facilitating clinicians in generating relevant diagnostic hypotheses. To this end, we have developed Zebra[5], a search engine optimised for the task of diagnosing difficult cases, to be used by clinicians at the time and place where diagnostic decisions are made [4].

2.1 Overview

Zebra provides an easy-to-use interface to generate a high number of diagnostic hypotheses given patient data as free text input. Clinicians can search for diagnoses by typing in symptoms, test results, or any textual data, and then filter through the documents that the system returns on the basis of their estimated relevance to the query terms.

Currently, Zebra's index contains 33,114 medical documents on rare and genetic diseases. A custom component is designed for crawling documents from 10 online medical resources. The indexing and retrieval is performed using the default settings of the open source search engine Indri[6].

In previous work we experimentally evaluated Zebra on more than 50 real-life medical cases of very difficult and rare diseases, where the query terms consisted of a list of patient symptoms [3]. Our findings showed that Zebra succeeded in finding the correct diagnosis, on average at ranks 1-3, in 67.9% of the test cases. However, for some of the cases, we observed the retrieval of multiple articles describing the same diseases.

2.2 Ranking diseases as opposed to documents

Diagnosing difficult cases is often an iterative process, where several fitting diagnostic hypotheses are selected by the clinician for further investigation. Returning a list of relevant documents given clinical data can be useful, but we argue that clinicians are more interested at this step primarily in the actual diseases to be considered, and secondarily in documents supporting these diseases. Therefore, this work focuses on enhancing Zebra with features such as clustering documents that treat the same disease, or retrieving and displaying the top diseases covered by the retrieved documents.

Annotating documents with medical concepts.

The medical documents in our index are very focused, most of them describing a particular disease. This high topical focus allows us to map documents to a disease or group of diseases. In order to map documents to diseases, we have used a subset of the Unified Medical Language System (UMLS) Metathesaurus[7] and the MetaMap[8] tool, both of which are recommended as standard by the U.S. National Library of Medicine [1].

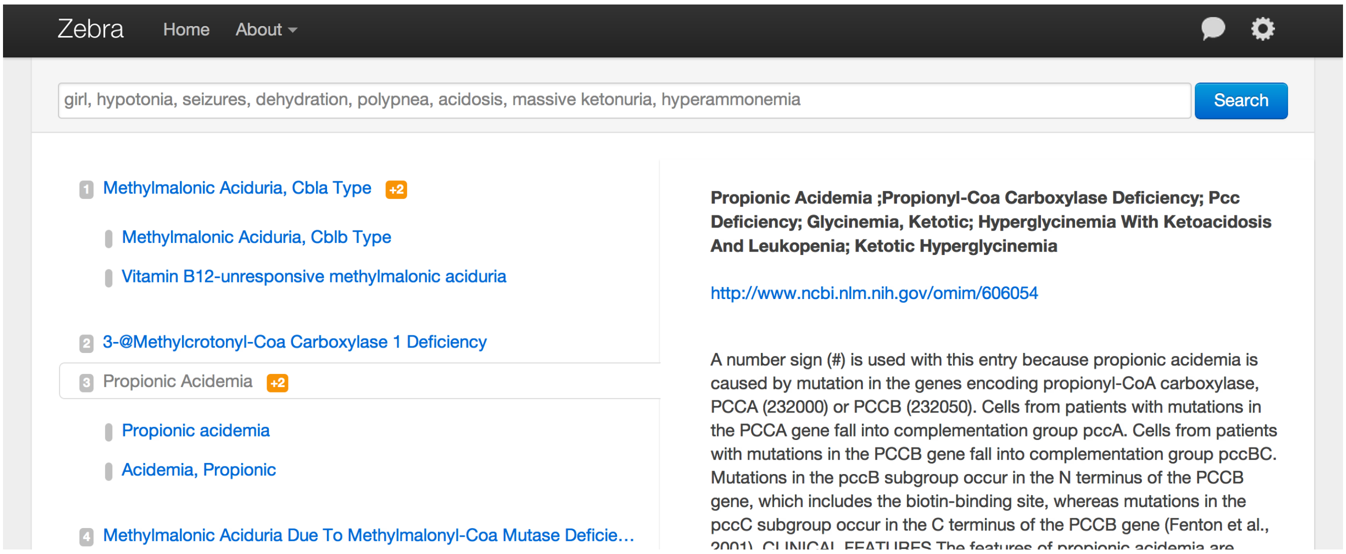
The UMLS Metathesaurus is a compendium of biomedical controlled vocabularies containing more that 3.5 million medical concept names (i.e. diseases) in English. We have selected a subset of knowledge sources from the 2011AA version of the Metathesaurus that are specifically focused on disease names and thus of interest for the task of annotating our medical documents: ICD10CM, OMIM, Disease Database, DXP, QMR and RAM. Altogether, these include 170,728 concept names.

MetaMap[9] is a tool that returns the most relevant concepts from the Metathesaurus, given some text as input. The titles of most of the documents we index consist of disease names. We use these titles as input for MetaMap and from the mapped concepts we keep only those with the highest matching score. For example, the title ”Vitamin B12-responsive methylmalonic acidemia” is mapped to the disease concept ”Methylmalonic Acidemia”.

After mapping the document titles with UMLS Metathesaurus concepts, we create a new index that includes the new meta-data associated with the documents. In our case, 99.75% of the unique titles have been mapped with at least one such medical concept, i.e. a disease name. A random inspection indicated that 93% of documents were correctly mapped.

Annotating documents with diseases as described above allows us to develop functionalities for clustering documents by diseases, searching for diseases, and ranking diseases.

[1] http://google.com
[2] http://pubmed.gov
[3] http://compbio.charite.de/phenomizer/
[4] http://www.orpha.net
[5] http://findzebra.com
[6] http://lemurproject.org/indri/
[7] http://www.nlm.nih.gov/research/umls/
[8] http://metamap.nlm.nih.gov
[9] http://skr.nlm.nih.gov/papers/



girl, hypotonia, seizures, dehydration, polypnea, acidosis, massive ketonuria, hyperammonemia

1 Methylmalonic Aciduria, Cbla Type +2

- Methylmalonic Aciduria, Cblb Type
- Vitamin B12-unresponsive methylmalonic aciduria

2 3-@Methylcrotonyl-Coa Carboxylase 1 Deficiency

3 Propionic Acidemia +2

- Propionic acidemia
- Acidemia, Propionic

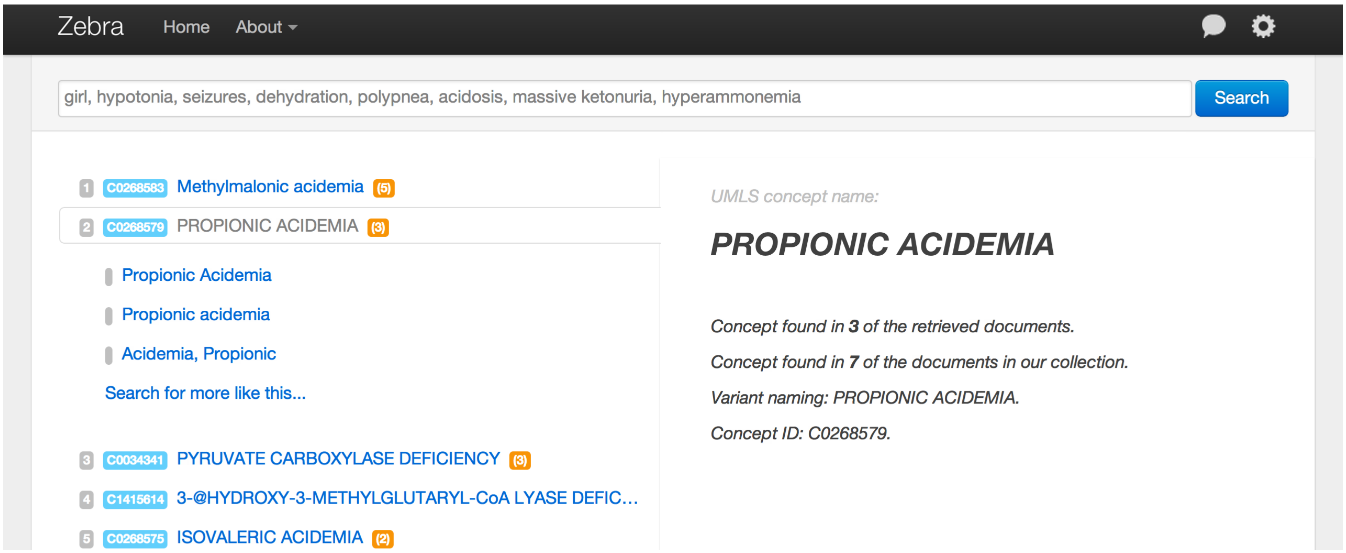
4 Methylmalonic Aciduria Due To Methylmalonyl-Coa Mutase Deficie…

Propionic Acidemia ;Propionyl-Coa Carboxylase Deficiency; Pcc Deficiency; Glycinemia, Ketotic; Hyperglycinemia With Ketoacidosis And Leukopenia; Ketotic Hyperglycinemia

http://www.ncbi.nlm.nih.gov/omim/606054

A number sign (#) is used with this entry because propionic acidemia is caused by mutation in the genes encoding propionyl-CoA carboxylase, PCCA (232000) or PCCB (232050). Cells from patients with mutations in the PCCA gene fall into complementation group pccA. Cells from patients with mutations in the PCCB gene fall into complementation group pccBC. Mutations in the pccB subgroup occur in the N terminus of the PCCB gene, which includes the biotin-binding site, whereas mutations in the pccC subgroup occur in the C terminus of the PCCB gene (Fenton et al., 2001). CLINICAL FEATURES The features of propionic acidemia are

Figure 1: Clustering documents annotated with the same UMLS Metathesaurus concepts



girl, hypotonia, seizures, dehydration, polypnea, acidosis, massive ketonuria, hyperammonemia

1 C0268583 Methylmalonic acidemia (5)

2 C0268579 PROPIONIC ACIDEMIA (3)

- Propionic Acidemia
- Propionic acidemia
- Acidemia, Propionic

Search for more like this...

3 C0034341 PYRUVATE CARBOXYLASE DEFICIENCY (3)

4 C1415614 3-@HYDROXY-3-METHYLGLUTARYL-CoA LYASE DEFIC…

5 C0268575 ISOVALERIC ACIDEMIA (2)

UMLS concept name:

PROPIONIC ACIDEMIA

*Concept found in **3** of the retrieved documents.*

*Concept found in **7** of the documents in our collection.*

Variant naming: PROPIONIC ACIDEMIA.

Concept ID: C0268579.

Figure 2: Ranking UMLS Metathesaurus concepts instead of documents

Figure 1 illustrates the clustering functionality of our system. Retrieved documents associated with the same medical concept (i.e. disease) are grouped and hidden under the highest ranking document from the cluster. Figure 2 illustrates the 'disease-ranking' functionality of our system. When ranking diseases, for each disease mapped to the retrieved documents, a disease score is computed taking into consideration the number of associated documents and the rank at which these documents were retrieved. By ranking diseases, we provide a more natural framework for clinicians to select diagnostic hypotheses: they can select diseases and get supporting documents, instead of selecting documents and having to elucidate the disease these documents cover.

The above two functionalities have not been evaluated in a principled way yet - this is something we are working on at the moment. However, we can report that we have received a very positive initial feedback from clinicians who have had access to Zebra.

3. CONCLUSION

Searching for rare disease diagnostic hypotheses is a highly specialised IR task for clinicians in the medical search domain. This task may not be optimally served by general purpose search engines. A specialised search engine tailored for this task could integrate better with the work-flow and time restrictions associated with the rare disease diagnostic process. To this end, we have created Zebra, a task-based vertical search engine for finding rare diseases. In previous work, Zebra has been evaluated using real-life medical cases of rare and difficult diseases, and has been found to be a useful and competitive tool for clinicians. In this work, we have extended Zebra's functionalities to optimise the task of medical diagnosis through search as follows: we added the option of grouping retrieved documents into clusters based on disease name occurrence, and we offered a 'disease-ranking' option, in addition to the standard 'document-ranking' option. In this way, by automatically annotating documents with medical concepts (i.e. diseases), we can streamline the diagnostic hypothesis generating process (1) by grouping similar documents and therefore allowing for a more diverse set of hypotheses to be presented, and (2) by ranking diseases instead of documents.

Future work will go beyond medical concept search and ranking. Presenting the search results as a network of diseases, suggesting query terms based on known medical terminology, and further integrating other knowledge sources are now viable due to the annotation of documents with medical concepts. Ultimately, by fusing two different data types, the medical web articles on one hand and the UMLS Metathesaurus data on the other, we aim to improve the overall usefulness of Zebra for clinicians and the outcome of the diagnostic process.

4. ACKNOWLEDGEMENTS

We thank MD Henrik L. Jørgensen (Department of Clinical Biochemistry, Bispebjerg Hospital, Copenhagen, Denmark), Lars Kai Hansen (DTU Informatics, Technical University of Denmark), Ingemar Cox (Department of Computer Science, University College London, London, U.K.), Birger Larsen and Peter Ingwersen (Information Systems and Interaction Design, Royal School of Library and Information Science, Copenhagen, Denmark) for their contribution to the development of the Zebra search engine.

5. REFERENCES

[1] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *JAMIA*, 17(3):229–36, May 2010.

[2] E. J. Campbell. The diagnosing mind. *Lancet*, 1(8537):849–51, Apr. 1987.

[3] R. Dragusin, P. Petcu, C. Lioma, B. Larsen, H. Jørgensen, I. Cox, L. Hansen, P. Ingwersen, and O. Winther. Zebra: a vertical search engine for rare diseases. *under review*, 2012.

[4] 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.

[5] J. W. Ely, J. A. Osheroff, M. L. Chambliss, M. H. Ebell, and M. E. Rosenbaum. Answering physicians clinical questions: obstacles and potential solutions. *JAMIA*, 12(2):217–224, 2005.

[6] P. N. Hider, G. Griffin, M. Walker, and E. Coughlan. The information-seeking behavior of clinical staff in a large health care organization. *JMLA*, 97(1):47–50, Jan. 2009.

[7] A. Hoogendam, A. F. H. Stalenhoef, P. F. D. V. Robbé, and a. J. P. M. Overbeke. Analysis of queries sent to PubMed at the point of care: observation of search behaviour in a medical teaching hospital. *BMC medical informatics and decision making*, 8:42, Jan. 2008.

[8] A. E. Kole and F. E. Faurisson. *The Voice of 12,000 Patients*. EURORDIS, 2009.

[9] O. Kostopoulou, J. Oudhoff, R. Nath, B. C. Delaney, C. W. Munro, C. Harries, and R. Holder. Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine. *Medical decision making*, 28(5):668–80, 2008.

[10] K. A. McKibbon and D. Fridsma. Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs. *JAMIA*, 13(6):653–659, 2006.

[11] M. G. Sim, E. Khong, and M. Jiwa. Does general practice Google? *Australian family physician*, 37(6):471–4, June 2008.

[12] R. H. Thiele, N. C. Poiro, D. C. Scalzo, and E. C. Nemergut. Speed, accuracy, and confidence in Google, Ovid, PubMed, and UpToDate: results of a randomised trial. *Postgraduate medical journal*, 86(1018):459–65, Aug. 2010.