Zebra: searching for rare diseases

A case of task-based search

Radu Dragusin
Department of Computer Science, University of Copenhagen, Denmark
raddr@imm.dtu.dk

Paula Petcu
Findwise Aps.,
Copenhagen, Denmark
paula.petcu@findwise.com

Christina Lioma
DTU Informatics, Technical University of Denmark
2800 Lyngby, Denmark
camli@imm.dtu.dk

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 IR may not suffice to satisfy the users in their tasks. In these situations, IR should be tailored to the user's task-specific needs and requirements. One such task is searching for rare disease diagnostic hypotheses.

In this work, we build upon an existing vertical search engine focused on rare disease diagnosis with good overall retrieval performance. We optimise the task of diagnosing such difficult cases by providing a more natural framework for the clinicians to select diagnostic hypotheses, and we illustrate this by adding the functionalities of grouping documents into clusters based on disease name occurrence, and by ranking diseases instead of documents.

Keywords

rare diseases, clinical information retrieval

1. INTRODUCTION

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 [7]. 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 [7]. 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 [1]. If the correct disease is short-listed at this step, the diagnostic outcome will probably be successful [8].

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 [9]. This shift is backed by the fact that 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 [10, 5], 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 [11].

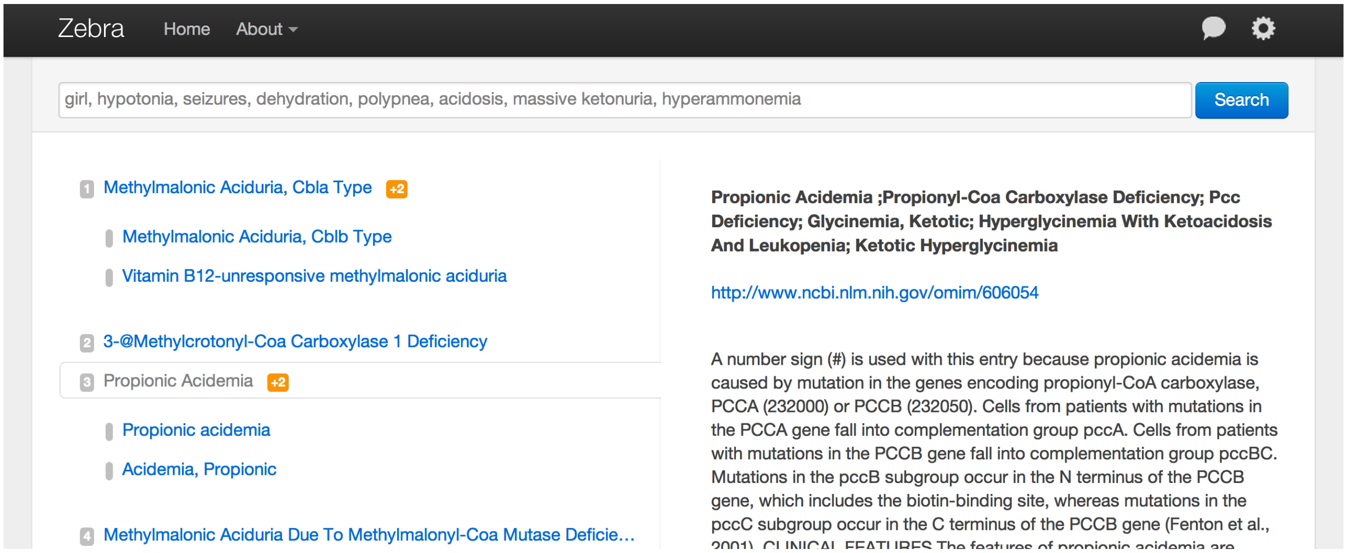
However, considering the time restrictions of the clinical setting, many of the questions still remain unanswered [4]. While having the right information is crucial when making diagnostic decisions, filtering through Google's results or formulating a PubMed query can be time consuming [6].

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.

In our previous work, we have developed a task-based search engine for rare disease diagnosis that takes free text

Task Based and Aggregated Search Workshop 2012, Barcelona, Spain
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$10.00.

[1] http://google.com
[2] http://pubmed.gov
[3] http://compbio.charite.de/phenomizer
[4] http://www.orpha.net



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

as input, and returns results from multiple sources of medical articles. This paper illustrates additional functionality designed to facilitate the diagnosis of rare diseases. Specifically we describe the grouping of documents into clusters based on disease name occurrence, and the ranking of diseases as opposed to documents.

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

2. ZEBRA SEARCH ENGINE

In an effort to tackle the problem of generating relevant rare disease diagnostic hypotheses for a given patient case, we 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 [3].

2.1 Overview

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

The index of the search engine 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].

Our previous work showed that Zebra succeeds in finding the correct diagnosis, on average before rank 3, in 67.9% of the test cases, where the query terms consisted of a list of patient symptoms corresponding to real patients suffering from rare diseases [2].

2.2 Annotating documents with concepts

Diagnosing difficult cases is often an iterative process, where several fitting diagnostic hypotheses are selected 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 diseases to be considered, and secondarily in supporting documents. Therefore, this work focuses on features such as clustering similar documents, or listing the top diseases covered by the retrieved documents.

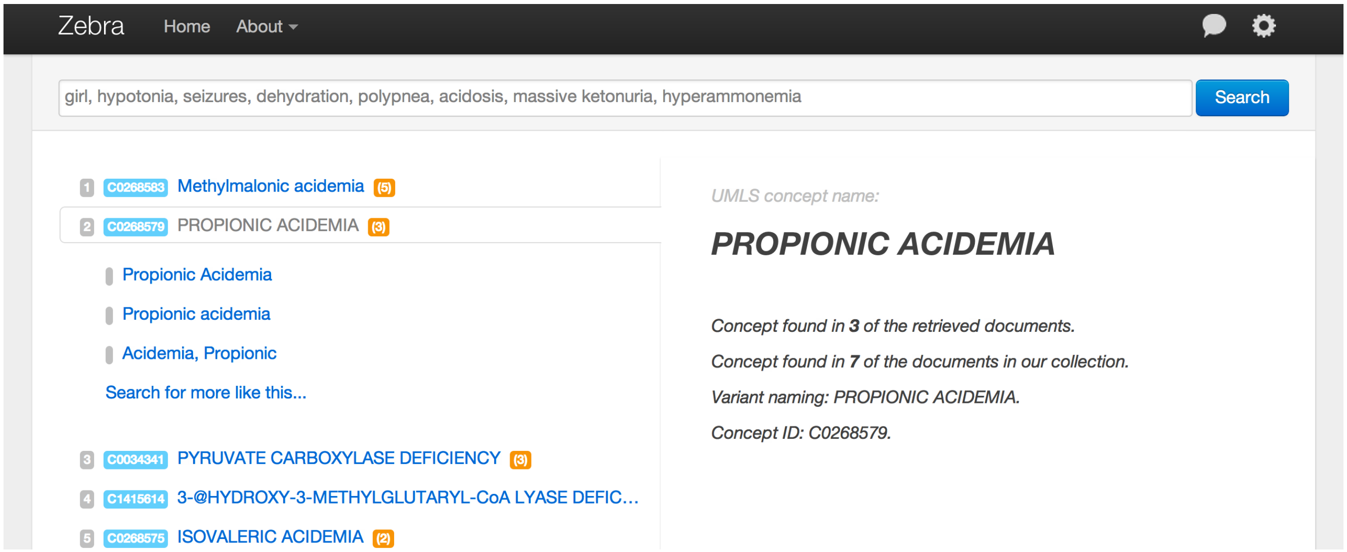
As the medical documents in our index are very focused, most of them describing a particular disease, we can map documents to a disease or group of diseases. In order to map documents to medical concepts, we have used a subset of the Unified Medical Language System (UMLS) Metathesaurus[7] and the MetaMap[8] tool.

The UMLS Metathesaurus is a compendium of biomedical controlled vocabularies containing more that 3.5 million concept names 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. 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

[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 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

one such concept. A random inspection indicated that 93% of documents were correctly mapped.

With the goal of improving the task of generating diagnostic hypotheses, we use the medical concepts associated to documents in order to add the functionalities of clustering documents by concepts, searching for concepts, and ranking concepts. With clustering, retrieved documents associated with the same concepts are grouped and hidden under the highest ranking document from the cluster (Figure 1). When ranking concepts, for each concept mapped to the retrieved documents, a concept score is computed taking into consideration the number of associated documents and the rank at which these documents were retrieved (Figure 2).

3. CONCLUSION

Searching for rare disease diagnostic hypotheses is a highly specialised IR task that is not well served by general purpose search engines. A specialised search engine tailored for this task would better integrate with the work-flow and time restrictions associated with the rare disease diagnostic process.

Previous work shows that we have good results for the task of diagnosing difficult cases. In this work, by automatically annotating documents with medical concepts, we streamline the 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.

Especially by ranking diseases, we provide a more natural framework for clinicians to select hypotheses, since they will now select diseases and get supporting documents instead of selecting documents and elucidate the disease these cover.

Future work will go beyond concept search and ranking. Presenting the 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 and the UMLS Metathesaurus data, we are able to explore this synergy in order to improve the usefulness of the system for clinicians and hopefully the outcome of the diagnostic process.

4. REFERENCES

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

[2] 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. *to be submitted for publishing*, 2012.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.