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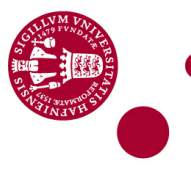
A Case of Task-Based Search in the Medical Domain

Radu Dragusin[1], Paula Petcu[2], Christina Lioma[1], Ole Winther[3]

dragusin@diku.dk   paula.petcu@findwise.com   lioma@gmail.com   owi@imm.dtu.dk

[1] Computer Science Department, University of Copenhagen, Denmark

FINDWISE
[2] Findwise Aps, Denmark

DTU
[3] DTU Informatics, Technical University of Denmark

1 Rare disease definition

affecting

<1 in 2.000 EU citizens

~30 million rare disease patients in EU

40% are misdiagnosed

25% suffer diagnostic delays of 5–30 yr.

5.000 – 8.000 known rare diseases in EU

80% of genetic origin

rare diseases are often difficult to diagnose:

- large number & variation 5–8 thousand
- low prevalence ($< \frac{1}{2000}$ affected)
- non-specific symptoms

Although clinicians will encounter many rare disease patients throughout their career they will have little practical experience with specific diseases since the likelihood of encountering them more than once is small.

2 Task-based search for rare diseases

when confronted with difficult cases, clinicians

traditionally use:
- books
- journals
- colleagues

increasingly use:
- Google
- PubMed
- Orphanet

information retrieval systems
- good at matching queries to large medical corpora

clinicians
- good at filtering unsuitable results

A task-based search engine would better fit the clinician's task-specific needs, being tailored for the workflow and time restrictions of the diagnostic process.

3 Zebra search engine

task integration
- easy-to-use interface
- generate diagnostic hypotheses
- takes patient data as free text

index
- 31,114 medical documents
- 10 online medical sources

evaluation
on 56 difficult real-life medical cases

Zebra: **67.9%** of test cases return the correct disease (MRR *0.385*)

Google: **37.5%** of test cases return the correct disease (MRR *0.206*)

system	MRR	P@10	P@20	correct answers in	
				top 10	top 20
Zebra	0.385	0.125	0.089	35 (62.5%)	38 (67.9%)
Google	0.206	0.088	0.071	16 (28.6%)	18 (32.1%)

For some of the test cases we observed the retrieval of multiple articles describing the same diseases.

regular search

4 Document clustering based on medical concepts

map documents to diseases
- most indexed documents cover one disease or group of diseases
- map document titles to diseases (using terms from UMLS Metathesaurus)

document ⟷ disease
- cluster documents by diseases
- search by disease name
- rank diseases, not documents

clustering results

5 Disease ranking

diagnosing difficult cases
- iterative process
- clinicians interested in disease hypotheses

ranking diseases
- disease score based on supporting documents
- a more natural framework for selecting hypotheses

disease ranking

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.