Specialised tools are needed when searching the web for rare disease diagnoses

Dragusin, Radu; Petcu, Paula; Lioma, Christina; Larsen, Birger; Jørgensen, Henrik L.; Cox, Ingemar; Hansen, Lars Kai; Ingwersen, Peter; Winther, Ole

Published in:
Rare Diseases

DOI:
10.4161/rdis.25001

Publication date:
2013

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC

Citation for published version (APA):
Dragusin, R., Petcu, P., Lioma, C., Larsen, B., Jørgensen, H. L., Cox, I., ... Winther, O. (2013). Specialised tools are needed when searching the web for rare disease diagnoses. Rare Diseases, 1(1), [e25001]. https://doi.org/10.4161/rdis.25001
Specialized tools are needed when searching the web for rare disease diagnoses

Radu Dragusin,1,2 Paula Petcu,1,3 Christina Lioma,1,2 Birger Larsen,4 Henrik L. Jørgensen,5 Ingemar J. Cox,1,6 Lars Kai Hansen,1 Peter Ingwersen* and Ole Winther†,*
1DTU Compute; Technical University of Denmark; Lyngby, Denmark; 2Department of Computer Science; University of Copenhagen; Copenhagen, Denmark; 3FindWise; Copenhagen, Denmark; 4Information Systems and Interaction Design; Royal School of Library and Information Science; Copenhagen, Denmark; 5Department of Clinical Biochemistry; Bispebjerg Hospital; Copenhagen, Denmark; 6Department of Computer Science; University College London; London, UK

Keywords: search engines, rare diseases, rare diagnoses, information technology within medicine

Submitted: 04/22/13
Accepted: 05/10/13

Citation: Dragusin R, Petcu P, Lioma C, Larsen B, Jørgensen HL, Cox IJ, et al. Specialized tools are needed when searching the web for rare disease diagnoses. Rare Diseases 2013; 1:e25001; http://dx.doi.org/10.4161/rdis.25001
*Correspondence to: Ole Winther; Email: owi@imm.dtu.dk

In our recent paper, we study web search as an aid in the process of diagnosing rare diseases. To answer the question of how well Google Search and PubMed perform, we created an evaluation framework with 56 diagnostic cases and made our own specialized search engine, FindZebra (findzebra.com). FindZebra uses a set of publicly available curated sources on rare diseases and an open-source information retrieval system, Indri. Our evaluation and the feedback received after the publication of our paper both show that FindZebra outperforms Google Search and PubMed. In this paper, we summarize the original findings and the response to FindZebra, discuss why Google Search is not designed for specialized tasks and outline some of the current trends in using web resources and social media for medical diagnosis.

Web Search for Diagnoses: Making the Case for Specialized Search Engines

When collaborating with physicians, one soon realizes that the web is an important resource for medical information.1 Google Search and PubMed are arguably the most popular web interfaces for physicians, although specialized resources are also widely used. Google indexes (collects, parses and stores) web data more thoroughly than any other search engine, and PubMed provides the search interface to the largest database of medical abstracts in the world. So if the medical information that the physician is looking for is available online, then one would imagine that at least one of these would have indexed it and would be able to retrieve it. Unfortunately, it turns out to be only partly true when used for medical diagnosis on rare diseases. In our recent study,2 we queried these tools with a list of symptoms and patient information for cases in which the final diagnosis was known. Documents associated with the correct diagnosis turned up among the first 20 Google Search results in roughly only one-third of the cases. Meanwhile, FindZebra, our specialized search engine, was able to retrieve relevant documents in around two-thirds of the cases. In the following, we discuss the shortcomings of Google Search for the task of searching for rare disease diagnostic hypotheses and the ingredients in FindZebra that make it more useful (in a statistical sense) for this task.

The ranking algorithm used by FindZebra matches indexed medical resources, such as web pages and documents, to the query terms and retrieves a ranked list of the documents that best match the query. The match between query terms and documents is computed using a query likelihood model that estimates the probability of the query being randomly sampled from a document model. The details of the Google Search ranking algorithm are not public, as this...
algorithm is central to Google’s business. However, it is known that it uses personalized information beyond the query, adjusts for page popularity (using PageRank) and has around 200 adjustable parameters that are optimized based on large-scale experimentation with users’ queries (that is, monitoring whether users, on average, click on a link ranked closer to or further from the top after parameter adjustment). In everyday life, we all experience Google’s effectiveness at finding what we are looking for to such a degree that we take it for granted. That a specialized search engine—tailored to a specific application domain—may still be superior can be explained by the following two points.

The first is the ranking algorithm. Google Search optimizes the average retrieval performance (that is, how close to the top the relevant documents are retrieved in the top 20 search results in around only one-third of the cases. The ranking algorithm used in FindZebra is the one from Indri, but we can conclude that all other elements of their ranking algorithm make the overall results inferior to FindZebra for this particular task.

### Test and Feedback on FindZebra

In the following section, we discuss in more detail the setup used to test the performance of the search engines for the task of finding the correct rare disease diagnosis. We also discuss the feedback we received from users in the month or so after the FindZebra search engine started attracting public attention and the degree to which that feedback confirms our results.

The 56 test queries used for evaluating the performance of the search engines were collected in three different ways. Five of these queries were constructed by the physician in the team, H.L.J., based upon his knowledge about the symptoms associated with specific rare diseases. For example, “Jewish boy age 16, monthly seizures, sleep deficiency, aggressive and irritable when woken, highly increased sexual appetite and hunger” corresponds to a diagnosis of Kleine Levin Syndrome.

Another 25 test queries were extracted by the authors from case stories published in the Orphanet Journal of Rare Diseases. For example, the symptoms “six year old, girl, weight length head circumference below the third percentile, atrophic and hyperpigmented skin lesions, pointed nose, aberrant thumbs with diminished flexion, bilateral glue ears, purulent rhinitis” correspond to a diagnosis of Rothmund-Thomson Syndrome. The last 26 queries were taken from a paper by Tang and Ng from the British Medical Journal. These authors tailored case descriptions to web searches and then investigated how well a group of medical experts equipped with the symptom list and Google Search could identify the correct diagnosis. These latter descriptions are, in general, shorter than the rest. For example, the symptoms “acute aortic regurgitation, depression, abscess” correspond to a diagnosis of Infective Endocarditis.

Going into a bit more detail with the results of the evaluation presented in our recent study, we identify that there are five of the total 56 cases for which some version
of Google Search or PubMed returns relevant results and for which FindZeba
does not return relevant results. Two of
these cases correspond to diseases that
are not present in FindZeba’s index, and
four of the cases are queries from the BMJ
article. Overall, we observe that Google
Search handles long queries worse than
FindZeba and that on shorter queries the
performance of the two search engines
is comparable. The two cases for which
the correct diagnosis is missing from
FindZeba’s index point to the fact that
FindZeba can be improved by including
more data. In general, one can expect
that multiple documents on the same dis-
ease will better capture the diversity of
symptoms.

These cases represent fairly realistic
elements of how a list of symptoms made
by medical experts will look at a well-informed stage of the diagnostic process.
It is doubtful that the lists of symptoms
are truly blind—that is to say final-
ized before the final diagnoses have been
reached. This means that there may be a
slight bias toward emphasizing symptoms
known to be associated with the disease.

Since the publication of the FindZeba
paper, the search engine has received wide-
spread attention and use. Over five weeks,
from March 17th to April 21st, FindZeba
delivered more than 1,000,000 diagnostic
hypotheses to more than 30,000 unique
visitors. Anecdotally, users appear to
agree that FindZeba offers improved
search results over existing alternatives.

For example, one blog reports that for the
query “purple urine” FindZeba suggests
the likely diagnosis (Porphyrias) at rank
two. In contrast, Google does not return
documents relevant for the diagnosis on
at least the first three pages of results.6

Another example is the query “osteopenia, hepatomegaly, anemia, fatigue, thrombocytopenia, nosebleed, Jewish”
with Gaucher disease as the likely diag-
nosis.7 Another case story reported on
the web4 describes the laborious process
of finding the correct diagnosis. Typing
in the symptoms listed in that case story,
“muscle cramps, intense headaches, rapid
weight gain, fatigue, edema, intolerance
to heat, excessive sweating, joint pain, ting-
ing in her hands and feet, frequent bone
fractures, acid reflux, intense anxiety and
panic attacks, high blood pressure, high
cholesterol, high blood sugar, sleep apnea,
menstrual irregularities, peripheral vision
loss and double vision” results in the
correct diagnosis (Cushing’s syndrome)
being retrieved at rank 15 in FindZeba.

Performing the same search in Google
returns the web page from which the
list of symptoms were taken, followed by
many pages with no immediate associ-
ation with the correct diagnosis.

Using Web Search and Social
Media for Diagnosis

FindZeba provides a simple and easy-to-
use interface, which streamlines the diag-
nostic process. Most of FindZeba’s search
results include the disease name in their
title, so it is easy to get an overview of the
potential diagnoses. The documents are
descriptions of the diseases so it is fast to
get the relevant information and rule out
possibilities. As with the anecdotal exam-
ple given above, Google Search often
returns documents with no direct associa-
tion with the specific disease, so even
though the information might be there,
it takes longer to extract. It is quite likely
that the patient in the last example above
would have been able to benefit from
FindZeba. It would not take the patient
long to go through the first 20 suggestions
given by FindZeba, ruling out quite a few
and taking a shortlist of potential diagnos-
tic hypotheses back to her physician.

The main target users for FindZeba
are general practitioners and specialists
within fields where rare diseases can
occur. Time is an important factor for
general practitioners, which makes dealing
with unusual symptoms especially challenging. General practitioners
are bound to meet diseases that they will
only encounter once in their career and
thus are very likely to miss the correct
diagnosis. It is our hope that FindZeba
can facilitate the correct diagnosis. Lack
of awareness about the specific diagnos-
isis definitely the main reason for rare
diseases being mis- and late-diagnosed.

For example, a study8 conducted by the
European Organisation for Rare Diseases
(EURORDIS) showed that 40% of rare
disease patients were wrongly diagnosed
before the correct diagnosis was given and

that 25% of patients had diagnostic delays
between 5 and 30 years.

Diagnosis of rare diseases is one of the
prime examples of how information tech-
nology can aid physicians. They are rare,
and there are many of them. The medi-
cal community will collectively have the
needed experience and knowledge to deal
properly with rare diseases, whereas this
is not possible for an individual physi-
cian. Information technology, such as
FindZeba, should enable the individual
to tap into this collective knowledge.
Social media also has the same potential,
as exemplified by recent initiatives.10,11
Elsewhere, we will discuss the relation-
ship between search engines and current
social media approaches as potential aids
for diagnosis.

FindZeba and similar systems can
have a major impact over how medical
diagnostic decisions are made. Certainly,
these systems can be improved, and the
feedback we have received so far indi-
cates that there is a strong support from
the medical community to facilitate this.

Decision support (test and treatment options) is only a part of FindZeba to
the degree that the indexed documents
contain such information. One could def-
initely streamline the presentation of the
decision support aspect and include an
option to have prevalence as a part of the
ranking algorithm. One long-term vision
is to have a truly individualized system
in which the physician registers each case
so that queries and the final diagnosis
are logged on a case-by-case basis. This
is a complex task because of the need
for user involvement, the possibly long
time-span before diagnosis and issues sur-
rrounding patient privacy. However, it has
the advantage of being unbiased (sym-
ptoms are reported before a diagnosis is
reached) and is a much richer informa-
tion source than using only the cases and
consensuses reported in literature. Such
a shared knowledge base on rare diseases
would have the potential of increasing our
understanding of rare diseases and greatly
improving diagnostics.

Disclosure of Potential Conflicts of Interest
No potential conflict of interest was
disclosed.
8. Ellen Uses UpToDate to find a diagnosis [Internet]. Waltham, MA: UpToDate: c2013 [cited 2013 April 22]. Available from: http://www.uptodate.com/home/ellen-uses-upToDate-find-diagnosis

References
8. Ellen Uses UpToDate to find a diagnosis [Internet]. Waltham, MA: UpToDate: c2013 [cited 2013 April 22]. Available from: http://www.uptodate.com/home/ellen-uses-upToDate-find-diagnosis