Rare Diseases 1, e25001; 2013; © 2013 Landes Bioscience

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[4] and Ole Winther[1,*]

[1]DTU Compute; Technical University of Denmark; Lyngby, Denmark; [2]Department of Computer Science; University of Copenhagen; Copenhagen, Denmark; [3]FindWise; Copenhagen, Denmark; [4]Information Systems and Interaction Design; Royal School of Library and Information Science; Copenhagen, Denmark; [5]Department of Clinical Biochemistry; Bispebjerg Hospital; Copenhagen, Denmark; [6]Department of Computer Science; University College London; London, UK

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.

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

Submitted: 04/22/13

Accepted: 05/10/13

Published Online: 06/05/13

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

Addendum to: Dragusin R, Petcu P, Lioma C, Larsen B, Jørgensen HL, Cox IJ, et al. FindZebra: A search engine for rare diseases. Int J Med Inform 2013; 82:528-38; PMID:23462700; http://dx.doi.org/10.1016/j.ijmedinf.2013.01.005

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

Table 1. Overview of the rare disease resources used by FindZebra

Resource	Entries
Online Mendelian Inheritance in Man (OMIM) http://www.ncbi.nlm.nih.gov/omim	20,369
Genetic and Rare Diseases Information Center (GARD) http://rarediseases.info.nih.gov/GARD	4,578
Orphanet http://www.orpha.net	2,967
Wikipedia http://www.wikipedia.org/	2,239
National Organization for Rare Disorders (NORD) http://rarediseases.org	1,230
Genetics Home Reference http://ghr.nlm.nih.gov	626
Madisons Foundation Rare Pedriatic Disease Database http://www.madisonsfoundation.org	522
About.com Rare Disease Database http://rarediseases.about.com	316
Health on the Net Foundation Rare Disease Database http://www.hon.ch	183
Swedish National Board of Health and Welfare www.socialstyrelsen.se/rarediseases	114

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 chosen link appears). If a person conducting a search works within a field that generates a relatively small volume of queries, such as rare diseases, then Google's ranking optimization might result in worse retrieval performance for this topic. For example, ranking according to page popularity may risk retrieving popular documents with only minor matches to the query. Second, if one can focus the search on websites with high quality content on the specialized topic, then one can eliminate noise coming from the overwhelming amount of irrelevant documents indexed by Google.

Can we tell which of these two reasons is the most important? Yes, because Google's advanced search option allows the user to specify which domains to search. For FindZebra we selected ten sites with highly curated information on rare diseases that represent more than 90% of Orphanet's list of about 7,000 rare diseases.[3] This yields a total of of roughly 33,400 documents (see **Table 1** for details). By restricting Google Search to the same domains used for FindZebra, 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, which, roughly speaking, ranks a document according to how frequent the query terms occur in that document.[4] As previously mentioned, using this simple ranking algorithm, FindZebra finds relevant documents in top 20 search results in two-thirds of the cases. Google Search's ranking algorithm definitely contains similarities to the query likelihood of 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*.[5] 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 FindZebra does not return relevant results. Two of these cases correspond to diseases that are not present in FindZebra's index, and four of the cases are queries from the BMJ article. Overall, we observe that Google Search handles long queries worse than FindZebra 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 FindZebra's index point to the fact that FindZebra can be improved by including more data. In general, one can expect that multiple documents on the same disease will better capture the diversity of symptoms.

These cases represent fairly realistic examples 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 finalized 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 FindZebra paper, the search engine has received widespread attention and use. Over five weeks, from March 17th to April 21st, FindZebra delivered more than 1,000,000 diagnostic hypotheses to more than 30,000 unique visitors. Anecdotally, users appear to agree that FindZebra offers improved search results over existing alternatives. For example, one blog reports that for the query "purple urine" FindZebra 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 diagnosis.[7] Another case story reported on the web[8] 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, tingling 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 FindZebra. 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 association with the correct diagnosis.

Using Web Search and Social Media for Diagnosis

FindZebra provides a simple and easy-to-use interface, which streamlines the diagnostic process. Most of FindZebra'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 examples given above, Google Search often returns documents with no direct association 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 FindZebra. It would not take the patient long to go through the first 20 suggestions given by FindZebra, ruling out quite a few and taking a shortlist of potential diagnostic hypotheses back to her physician.

The main target users for FindZebra are general practitioners and specialists within fields where rare diagnoses 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 FindZebra can facilitate the correct diagnosis. Lack of awareness about the specific diagnosis is definitely the main reason for rare diseases being mis- and late-diagnosed. For example, a study[9] 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 technology can aid physicians. They are rare, and there are many of them. The medical community will collectively have the needed experience and knowledge to deal properly with rare diseases, whereas this is not possible for an individual physician. Information technology, such as FindZebra, 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 relationship between search engines and current social media approaches as potential aids for diagnosis.

FindZebra 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 indicates that there is a strong support from the medical community to facilitate this. Decision support (test and treatment options) is only a part of FindZebra to the degree that the indexed documents contain such information. One could definitely 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 surrounding patient privacy. However, it has the advantage of being unbiased (symptoms are reported before a diagnosis is reached) and is a much richer information 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.

References

1. Cartright MA, White RW, Horvitz E. Intentions and attention in exploratory health search. Edited by Ma WY, Nie JY, Baeza-Yates RA, Chua TS, Croft WB. New York, NY: ACM 2011:65-74.
2. Dragusin R, Petcu P, Lioma C, Larsen B, Jørgensen HL, Cox IJ, et al. FindZebra: A search engine for rare diseases. Int J Med Inform 2013; 82:528-38; PMID:23462700; http://dx.doi.org/10.1016/j.ijmedinf.2013.01.005
3. Orphanet Report Series: List of rare diseases [Internet]. Orphanet: c2012 [cited 2013 April 22]. Available from: http://www.orpha.net/orphacom/cahiers/docs/GB/List_of_rare_diseases_in_alphabetical_order.pdf contains a list of around 7.000 rare diseases.
4. Strohman T, Metzler D, Turtle H, Croft WB. Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligence Analysis (ICIA) 2005.
5. Tang H, Ng JH. Googling for a diagnosis--use of Google as a diagnostic aid: internet based study. BMJ 2006; 333:1143-5; PMID:17098763; http://dx.doi.org/10.1136/bmj.39003.640567.AE
6. FindZebra: Rare Diseases Search Engine [Internet]. Oak Park (IL): Sirensong: c2013 March 22 [cited 2013 April 22]. Available from: http://sirensong.sireninteractive.com/search-engine-marketing/find-zebra-rare-disease-search-engine/
7. An Interview with the Developers of FindZebra – The Search Engine for Rare Diseases [Internet]. Rare Disease Report: c2013 April 4 [cited 2013 April 22]. Available from: http://raredr.com/medicine/articles/interview-developers-findzebra-search-engine-rare-diseases
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
9. EurordisCare2: Survey of diagnostic delays, 8 diseases, Europe (2004) [Internet]. Paris, France: EURODIS: c2007 March [cited 2013 April 22]. Available from: http://archive.eurordis.org/article.php3?id_article=454
10. The Medical Futurist [Internet]. c2013. Available from: http://themedicalfuturist.com
11. CrowdMed [Internet]. c2012. Available from: https://www.crowdmed.com