FindZebra - using machine learning to aid diagnosis of rare diseases

FindZebra is a search engine for rare diseases intended to act as a diagnosis decision support system (DDSS) capable of assisting the user both during and after a search. Rare diseases are diseases that affect only a small part of the population (less than one in two thousand). Currently around seven thousand rare diseases are known and it is estimated that 6–8% of the population will be affected by a rare disease during their lifetime. Due to their rarity and large number, diagnosis of rare diseases is difficult and often associated with year long delays and diagnostic errors. These difficulties with diagnosis have a profound human and societal cost. This means that even a small increase in success rate when using a tool such as FindZebra could potentially have a great impact on society. In this dissertation we explore four lines of research for improving FindZebra using machine learning methods. The first line of research is on how to improve the retrieval performance of FindZebra. By using a combination of improved models, medical databases and corpus expansion we show that it is possible to obtain a substantial improvement in retrieval performance compared to current state-of-the-art document retrieval systems. Improving retrieval performance is important, but is not the only way of improving the success rate of a DDSS such as FindZebra. Following an unsuccessful search, the search engine should assist the user by indicating what information is likely to be missing. This idea is called Information Completion (IC) and will be explored in the second line of research. In order to represent words (and other discrete tokens) in a neural network it is necessary to transform each word to a vector form. This is typically accomplished by using a word embedding, which is an essential component in any word based neural network. The third line of research is on how to improve this basic component. Users of FindZebra who do not have English as their primary language often have difficulty expressing complex medical queries in English. Optimally, a user should be able to write a query in his or her native language and the search engine should then give a suggestion for a differential diagnosis based on all the information contained in a multilingual corpus, not only in the native corpus. Methods for performing multilingual search will be the fourth line of research explored in this dissertation.

General information
State: Submitted
Organisations: Department of Applied Mathematics and Computer Science, Cognitive Systems
Authors: Svenstrup, D. T. (Intern), Winther, O. (Intern)
Number of pages: 118
Publication date: 2017

Publication information
Publisher: DTU Compute
Original language: English
Series: DTU Compute PHD-2017
Volume: 463
ISSN: 0909-3192
Main Research Area: Technical/natural sciences

Relations
Projects:
FindZebra - using machine learning to aid diagnosis of rare diseases
Publication: Research › Ph.D. thesis – Annual report year: 2017
Method of and System for Information Retrieval

This invention relates to a system and a method (100) of searching a collection of digital information (150) comprising a number of digital documents (110), the method comprising receiving or obtaining (102) a search query, the query comprising a number of search terms, searching (103) an index (300) using the search terms thereby providing information (301) about which digital documents (110) of the collection of digital information (150) that contains a given search term and one or more search related metrics (302; 303; 304; 305; 306), ranking (105) at least a part of the search result according to one or more predetermined criteria providing a ranked search result, and providing at least a part of the ranked search result (106), wherein the ranking provides robust likelihood for low count terms by using the one or more search related metrics (302; 303; 304; 305; 306). In this way, a method of and a system for information retrieval or searching is readily provided that enhances the searching quality (i.e. the number of relevant documents retrieved and such documents being ranked high) when (also) using queries containing many search terms.

General information
State: Published
Organisations: Department of Informatics and Mathematical Modeling, Department of Applied Mathematics and Computer Science, Cognitive Systems
Authors: Winther, O. (Intern), Svenstrup, D. T. (Intern)
Publication date: 18 Jun 2015

Rare disease diagnosis: A review of web search, social media and large-scale data-mining approaches

Physicians and the general public are increasingly using web-based tools to find answers to medical questions. The field of rare diseases is especially challenging and important as shown by the long delay and many mistakes associated with diagnoses. In this paper we review recent initiatives on the use of web search, social media and data mining in data repositories for medical diagnosis. We compare the retrieval accuracy on 56 rare disease cases with known diagnosis for the web search tools google.com, pubmed.gov, omim.org and our own search tool findzebra.com. We give a detailed description of IBM's Watson system and make a rough comparison between findzebra.com and Watson on subsets of the Doctor's dilemma dataset. The recall@10 and recall@20 (fraction of cases where the correct result appears in top 10 and top 20) for the 56 cases are found to be be 29%, 16%, 27% and 59% and 32%, 18%, 34% and 64%, respectively. Thus, FindZebra has a significantly (p <0.01) higher recall than the other 3 search engines. When tested under the same conditions, Watson and FindZebra showed similar recall@10 accuracy. However, the tests were performed on different subsets of Doctors dilemma questions. Advances in technology and access to high quality data have opened new possibilities for aiding the diagnostic process. Specialized search engines, data mining tools and social media are some of the areas that hold promise.

General information
State: Published
Organisations: Department of Applied Mathematics and Computer Science, Cognitive Systems, Department of Informatics and Mathematical Modeling, Copenhagen University Hospital
Authors: Svenstrup, D. T. (Intern), Jørgensen, H. L. (Ekstern), Winther, O. (Intern)
Number of pages: 7
Publication date: 2015
Main Research Area: Technical/natural sciences
FindZebra - the search engine for difficult medical cases

General information
State: Published
Organisations: Department of Applied Mathematics and Computer Science, Cognitive Systems, Technical University of Denmark, Copenhagen University Hospital
Authors: Winther, O. (Intern), Svenstrup, D. T. (Intern), Henningsen, P. P. (Ekstern), Kristiásson, R. (Ekstern), Jørgensen, H. L. (Ekstern)
Pages: 822
Publication date: 2014
Main Research Area: Technical/natural sciences

Publication information
Journal: Orphanet Journal of Rare Diseases
Volume: 9
Issue number: 1
ISSN (Print): 1750-1172
Ratings:
BFI (2018): BFI-level 1
BFI (2017): BFI-level 1
Web of Science (2017): Indexed Yes
BFI (2016): BFI-level 1
Scopus rating (2016): SJR 2.864 SNIP 0.642 CiteScore 5
Original language: English
Electronic versions:
21675511.2015.1083145.pdf
DOIs:
10.1080/21675511.2015.1083145
Source: Findit
Source-ID: 2281670346
Publication: Research - peer-review › Journal article – Annual report year: 2015
Projects:

FindZebra - using machine learning to aid diagnosis of rare diseases

Department of Applied Mathematics and Computer Science
Period: 01/09/2014 → 31/10/2017
Number of participants: 5
Phd Student:
Svenstrup, Dan Tito (Intern)
Main Supervisor:
Winther, Ole (Intern)
Examiner:
Larsen, Jan (Intern)
Ohlsson, Mattias (Ekstern)
Søgaard, Anders (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Offentlig finansiering

Relations
Publications:
FindZebra - using machine learning to aid diagnosis of rare diseases
Project: PhD