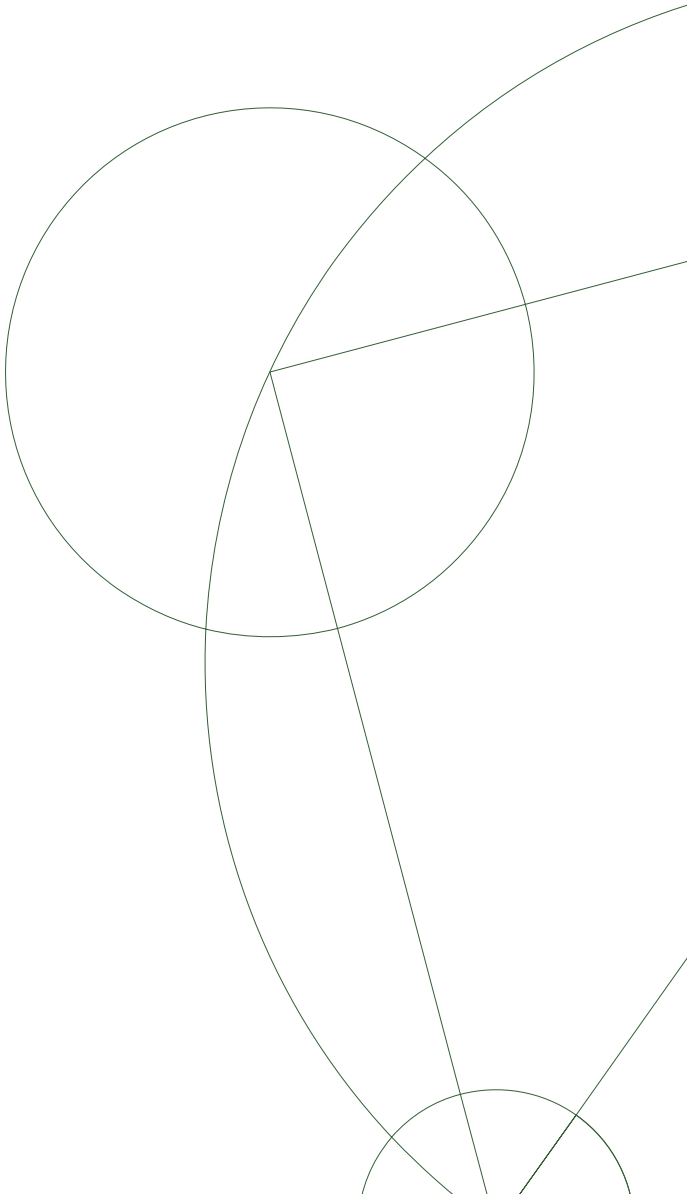
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

Master's thesis

Radu Drăgușin

Paula Petcu

A Vertical Search Engine Supporting the Diagnosis of Rare Diseases



Academic advisor: Ole Winther
Co-supervisor: Christina Lioma

Submitted: 07/08/11

Abstract

Around 30 million EU citizens suffer from a rare disease, and for many of them an early diagnosis could be lifesaving. However, rare diseases are notoriously difficult to diagnose because of their low prevalence, large number, and broad diversity of symptoms, so rare disease patients are often misdiagnosed or experience long diagnostic delays.

In this thesis we develop a search engine specifically designed for the task of diagnosing rare diseases. The retrieval is performed on a large collection of topically relevant medical articles and the user interface is optimised for generating diagnostic hypotheses.

The performance of the vertical search engine is compared to that of other web tools currently used by clinicians as aids in diagnosing difficult cases. The evaluations show that the developed search engine has overall better performance than the other tools.

Although our evaluations are promising, further studies are needed to establish if using a vertical search engine could improve the clinical process of diagnosing difficult cases and reduce diagnostic errors.

Contents

Chapter 1

Introduction

1.1 Motivation

Computer software has been shown to improve many aspects of the clinical process [1]. However, in the area of clinical diagnostics, the impact and adoption of specialized software fell short of expectations [2, 3].

1.1.1 Clinical diagnosis decision support systems

Even if clinical diagnosis decision support systems (CDDSS) have been researched and developed over the years, the proposed solutions were not adopted by the medical community, in part due to a lack of synergy between the final software products and the diagnostic process [4].

Multiple reasons were identified for the limited adoption of such systems in clinical use: the lengthy process of introducing patient data and interpreting the response, the lack of integration with the clinical workflow, and the inability to anticipate the clinical needs [5].

Although CDDSSs are not widely used in practice, the need for a support system exists. Many clinicians, when faced with difficult cases, rely on general purpose search engines or medical databases [6, 7]. Recent studies have shown that Google Search[1] is the preferred resource for searching medical information [7, 8, 9], but PubMed[2] is also widely used [7]. However, neither of these systems fits well with the task of finding a diagnosis based on patient data. Google is not optimized for this task, but rather for general web search, whereas PubMed, a medical bibliographic search engine, does not rank results by relevance, but merely sorts them by publish date or other bibliographic information.

Even if most of the clinical work is on common diseases, clinicians are most likely to search for information when they encounter diagnostic difficulties.

[1] Google Search, `http://www.google.com/`

[2] PubMed, `http://www.ncbi.nlm.nih.gov/pubmed/`

Therefore, dealing with such cases is an area where CDDSSs could improve the current clinical practice. This is especially important, since such cases often result in misdiagnosis or diagnosis delays that could negatively affect the patient's outcome [10].

1.1.2 Rare diseases

Many rare diseases are notoriously difficult to diagnose. The difficulty in diagnosing rare diseases stems from their low prevalence, large number, and broad diversity of symptoms. When encountering a rare disease patient, clinicians often have little information on the disease. This can lead to referring the patient to a specialist, performing unnecessary tests, or misdiagnosis. A study conducted by EURORDIS, the European Organization for Rare Diseases, 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 [11].

In recent years, rare diseases (also known as orphan diseases) gained special status[3,4] but there is no international consensus on what defines a rare disease. Some diseases, such as malaria, are common in some areas, but have low prevalence in others. Under the EC Regulation on Orphan Medicinal Products [12], a rare disease must have a prevalence of less than 1 case in 2000 persons. Under this classification, there are close to 8000 rare diseases and around 30 million (6-8%) EU citizens affected by a rare disease [13]. About 80% of rare diseases have genetic origins.

Existing rare disease diagnostic tools are either restrictive on their input (symptoms must be selected from a predefined list), use manually constructed knowledge bases (difficult to keep up-to-date) [14, 15], or they use a general-purpose information retrieval (IR) system (not optimised for the task of diagnosing rare diseases). For example, Google's use of PageRank [16] does not make sense for rare disease retrieval since articles on rare diseases are highly specialized and not necessarily popular.

Given the high percentage of misdiagnoses, long diagnostic delays, the large number of patients suffering from rare diseases, and the costs of unnecessary tests and interventions, it can be argued that there is a need to research and develop a system with the purpose of supporting the diagnosis of rare diseases.

[3]European Commission Perspective, `http://ec.europa.eu/health/ph_information/documents/ev20040705_rd05_en.pdf`

[4]US Rare Diseases act of 2002, `http://www.gpo.gov/fdsys/pkg/PLAW-107publ280/pdf/PLAW-107publ280.pdf`

1.2 Project Goal

The overall goal is to create a freely available search engine dedicated to rare diseases, that can be used by general practitioners, as well as experts in rare diseases. The system intends to improve clinical practice by (1) providing an extensive resource of rare disease information, (2) that can be freely accessed, (3) providing a simple and intuitive search interface, and (4) displaying information meaningful for clinicians to rapidly take decisions at the time and place of the consultation.

In order to assess the possible improvements to clinical practice, the system is evaluated and compared in terms of effectiveness and time requirements to other systems used by clinicians in the diagnostic process.

On the long term, a system based on this approach could lower the misdiagnosis rate and reduce delays in the diagnosis of patients suffering from rare diseases. Ultimately, such a system could have a positive impact on patients' outcome, and lower healthcare costs.

1.3 Research Questions

RQ1 Does the experimental evaluation of our system show substantial improvements over other systems in terms of document relevance?

RQ2 Does the inclusion of a larger pool of articles on the topic of genetic diseases improve the effectiveness of the system in diagnosing rare diseases?

RQ3 Does increasing the prior probabilities of the relevance of rare disease articles in contrast to the relevance of genetic disease articles improve the effectiveness of the system in diagnosing rare diseases?

RQ4 Does the use of our system, in comparison with other systems, decrease the search time spent by clinicians looking for rare disease diagnostic hypotheses?

1.4 Contributions

The main contributions of this thesis are the following:

(a) Gathered a large collection of articles on rare and genetic diseases

(b) Developed a vertical search engine for the task of diagnosing rare diseases

(c) Developed a web UI and an API to interact with the search engine

(d) Delivered an alternative to the existing systems supporting rare disease diagnosis

(e) Established an evaluation methodology tailored for clinical diagnosis on the web

(f) Created a query collection for rare disease diagnosis systems evaluation

(g) Evaluated the developed system and other systems currently used by clinicians as aids in the diagnostic process

The information resources used in the vertical search engine were collected from various sources, providing rare and genetic disease articles heterogeneous in quality, length, and authority. A collection of around 30,000 topical documents was retrieved from eight online medical resources and two medical database resources. Additional general medical databases, collections and classifications were retrieved and analysed.

The developed vertical search engine takes as input any textual patient data, such as symptoms, test results, demographic information, and returns a ranked list of potentially relevant documents on the topic of rare diseases. Alternatively, the user can request a ranked list of disease names instead of documents. The engine was developed using the open-source Lemur Project[5] and is licensed under the GNU General Public License v2[6].

The system provides a simple-to-use web user interface (UI). Additionally, we provide PDF output capability summarizing the results for later analysis by clinicians. The system provides a web application programming interface (API) for third-party applications to submit queries and receive results in either HTML, XML, JSON, or PDF formats.

The design and development of the vertical search engine was backed by a previous literature review on CDDSSs [17], discussions with a clinician and a group of rare and genetic disease specialists, as well as input from information retrieval experts.

In order to assess the performance of the system when compared to current products used by clinicians, an evaluation methodology was devised specifically for the task of diagnosing rare diseases based on textual patient data. An evaluation based on this methodology was applied on two query collections: a query collection constructed in collaboration with a medical doctor[7] consisting of 30 cases of rare disease patients, and another set of 26 queries from a previous study [18]. All of the queries are based on case descriptions published in medical journals, as there is no dataset associating patient data to rare diseases.

[5]The Lemur Project, `http://www.lemurproject.org`

[6]GNU General Public License v2, `http://www.gnu.org/licenses/gpl-2.0.html`

[7]Henrik L. Jørgensen, MD, PhD, Department of Clinical Biochemistry, Bispebjerg University Hospital, Denmark

1.5 Thesis Outline

The rest of the thesis is organized as follows. Chapter 2 describes the clinical process of diagnosing diseases, the difficulties that are encountered by clinicians, the current trends in assisting them in the diagnostic process, and the available medical information resources. Chapter 3 discusses the design of the vertical search engine and the methodology devised for evaluating it and other systems used in diagnosis. The vertical search engine's efficiency and effectiveness test results are presented in Chapter 4. Chapter 5 summarizes the work done in the thesis, analyses the limitations of the current system, and provides future extension ideas. Finally, Chapter 6 concludes the thesis and restates the contributions of this work.

Chapter 2

Background

2.1 Supporting the Diagnostic Process

In order to develop a system to improve the diagnostic process, it is important to understand how this process works, what the difficulties are, and where are the diagnostic errors most likely to occur. Understanding these issues is crucial in successfully integrating the CDDSSs into the clinical workflow and being accepted by the medical community [19].

2.1.1 The diagnostic process

The definition of diagnosis is not limited to a single concept, and ranges from simply associating a disease to the symptoms presented by the patient [20], to the analysis of the course of a disease from patient details (medical history, symptoms, signs) [21]. Disease diagnosis involves a sequential testing of hypotheses that are often drawn from additional history, symptoms, physical exams, and laboratory tests, and that are verified by trials to see if the patient responds to a specific treatment [21].

Our focus is on associating diseases to patient data. Given this definition, the process of eliciting the correct disease (Figure 2.1) consists of generating several hypotheses and, after a process of selection and elimination, reaching a diagnostic decision. Finally, the clinician selects the best way to manage the disease. However, the process is not necessarily that linear and sometimes a hypothesis is selected after a therapeutic trial is administered to see if the patient responds to treatment [21].

Both the clinician's knowledge and experience play an important role in the diagnostic process. When generating hypotheses, clinicians use two levels of medical knowledge: a low level one, comprised of medical facts, and a high level one, obtained through professional experience [22]. It was suggested that clinicians acquire approximately two million medical facts during their studies and career [22].

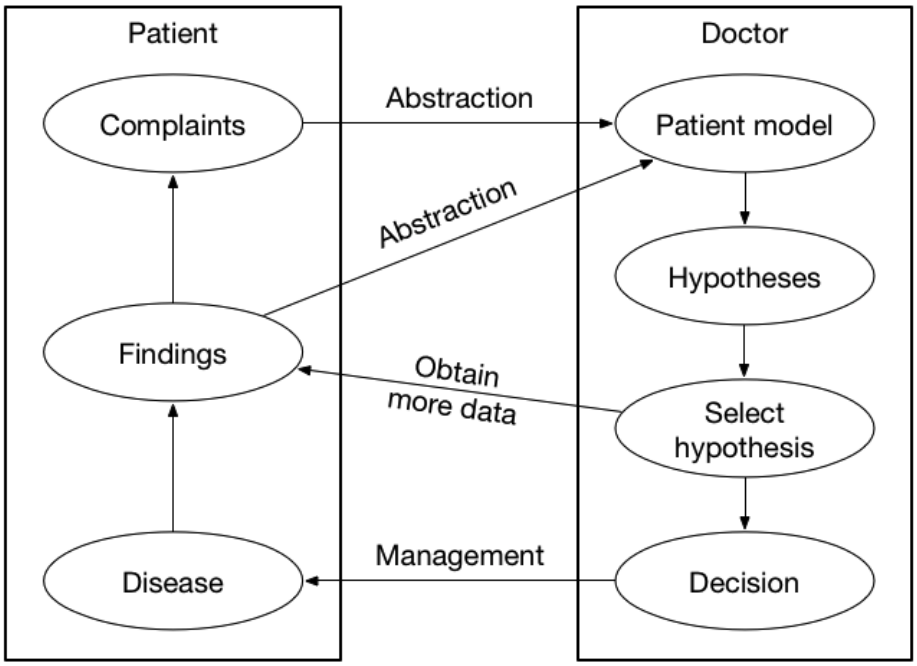


Figure 2.1: **The diagnostic process**, after [22]. A patient suffering from a disease arrives at the clinician with some complaints. Together with the findings, the clinician forms a patient model from which several hypotheses are derived. In order to verify a hypothesis, more findings could be necessary. Once a decision is reached, the clinician chooses the best way to manage the disease.

By pattern matching their medical knowledge with patient data, clinicians generate up to six or seven hypotheses, sometimes consisting of classes of diseases [23]. This ability to rapidly generate hypotheses increases with clinical experience [24].

As the volume of medical knowledge is constantly increasing, clinicians find it hard to keep the pace with the medical literature. MEDLINE, the leading medical bibliographic resource, adds between 2,000 and 4,000 citations each month to its existing 18 million references[1]. Even if many medical institutions have guidelines in place, there is a significant delay between these guidelines being published and being adopted in clinical practice [25].

2.1.1.1 Diagnostic difficulty and error

For ninety percent of the patients, the first contact with the medical environment is through the general practitioner [26]. From early on in the

[1]MEDLINE Fact Sheet, `http://www.nlm.nih.gov/pubs/factsheets/medline.html`

consultation, clinicians are able to identify a few diagnostic hypotheses, however, with experience, they tend to rapidly recognize patterns instead of exhaustively considering alternative hypotheses [27]. While this rapid pattern-matching approach saves time and reduces testing costs in most of the cases, for unusual presentations of common diseases or cases of rare diseases this could lead to misdiagnoses.

Studies have shown that the key to avoiding misdiagnoses is having a good set of diagnostic hypotheses [24, 27]. It was reported that in most of the misdiagnoses, the correct hypothesis was not considered in the differential diagnosis [27]. Moreover, in the case of rare diseases, general practitioners may not be familiar with the pathology of many of the rare diseases, and thus may not consider them in the differential diagnosis. If this is the case, the diagnosis could be delayed or the patient may be misdiagnosed [24].

Another issue that may be indicative of misdiagnosis relates to the unanswered questions that clinicians face during consultations. Studies show that up to half of the questions clinicians raise at the time and place where diagnostic decisions are made remain unanswered [28, 29]. Although most of these questions do not necessarily affect the final diagnosis, many of the medical errors are caused by delayed or erroneous decisions [30]. Time constraints and lack of adequate resources are the main obstacles in pursuing an answer [29].

Difficult cases increase the likelihood of diagnostic errors. Subsequently, it is important to provide general practitioners with the best possible support to avoid misdiagnosing difficult cases. A computer-assisted diagnostic system that generates alternative hypotheses given patient data could improve the diagnostic process for difficult cases, reducing delays and misdiagnosing rates. The challenge is to understand how to support difficult cases diagnosis without undermining the clinician's experience, lengthening the diagnostic process, or obstructing the clinician's reasoning.

2.1.2 Previous efforts on supporting diagnosis

Early efforts to use computer diagnostic aids date to more than five decades ago [19], but health care institutions have been slow in incorporating them into the clinical workflow. It has been repeatedly asserted in literature that these systems have the potential to reduce diagnostic errors and improve quality of care [31, 32, 26, 33], and the utility of some of them was even demonstrated through laboratory evaluation studies [31], but few were tested in the field or developed further than the prototype stage [34], and none of them is in widespread use today.

These systems have been previously categorized in literature along several axes: based on their timing (before, during, or after consultation), setting (inpatient or ambulatory care), scope (general or specialized), and in terms of integration with other systems (with, for example, electronic health

records EHR) [35, 36].

Early CDDSSs used predefined sets of rules, applied Bayesian inference to calculate disease probabilities, or used machine learning to recognise patterns between patient symptoms and diseases, to arrive at a list of possible diagnoses. This first generation of diagnosis support systems included MYCIN[2], QMR[3], Iliad[4] or DXplain[5], and despite proven utility in experimental settings [37], they encountered acceptance difficulties by the medical community - mainly due to the amount of time needed to introduce clinical data and the lack of high-quality clinical diagnostic knowledge content [21]. Of these, DXplain displayed rare diseases separately from common diseases [38]. Specialised on genetic disorders, Phenomizer[6] is a tool based on the Human Phenotype Ontology (HPO)[7] that correlates phenotypic abnormalities with genetic disorders (OMIM entries) and contains around 9,900 features and 5,020 diseases [39]. Regardless of implementation, these systems usually take as input some patient data through predefined drop-down lists or by repeatedly asking clinicians for specific patient details, which is time-consuming and cumbersome to use.

With the goal of facilitating the storage and searching of medical information, a wide variety of medical data has been aggregated into databases. One such example is the OMIM database system[8], specialized on human genes and genetic phenotypes, containing information for all mendelian disorders and over 12,000 genes [40]. On the topic of rare diseases, the Orphanet database[9] contains information on more than 5,000 rare diseases, and provides a service for retrieving data for about 2,000 rare diseases based on clinical signs [14]. Other databases on topics associated with rare diseases include the London Dysmorphology Database[10], which is focused on photographic information for rare dysmorphic syndromes [41], and Possum[11], which is a dysmorphology database that contains textual and photographic information on more than 3,000 syndromes [42].

The search by clinical signs service provided by both Orphanet and Phenomizer is done using a controlled vocabulary (thesaurus). To search for a diagnosis in Orphanet, the user has to go through multiple steps. Going through a thesaurus and finding the right match can be a complex process that lengthens the diagnostic time, negatively impacts the usability, and limits integration in the clinical environment. Similarly, in Phenomizer, the

[2]MYCIN, `http://www.computing.surrey.ac.uk/ai/PROFILE/mycin.html`
[3]Quick Medical Reference, `http://www.openclinical.org/aisp_qmr.html`
[4]Iliad, `http://www.openclinical.org/aisp_iliad.html`
[5]DXplain, `http://dxplain.org/dxp/dxp.pl`
[6]Phenomizer, `http://compbio.charite.de/Phenomizer/Phenomizer.html`
[7]HPO, `http://www.human-phenotype-ontology.org/index.php/hpo_home.html`
[8]OMIM, `http://www.ncbi.nlm.nih.gov/omim`
[9]Orphanet, `http://www.orpha.net/`
[10]London Dysmorphology Database, `http://www.lmdatabases.com/`
[11]Possum, `http://www.possum.net.au/`

patient symptoms and signs must be selected from a predefined list compiled from the HPO ontology.

Another system that is being used by medical doctors for answering clinical questions is PubMed [7], which is a medical citation search engine that indexes over 20 million citations for biomedical literature from MEDLINE, life science journals, and online books. However, PubMed's main drawback when searching for a diagnosis is the fact that the results are not ranked based on query relevance, but only on publish date, author name or other article meta-information that is not necessarily relevant in the search for a diagnosis. Moreover, when submitting a query without additional boolean operators, only articles containing all query terms are retrieved, dramatically reducing the number of retrieved documents.

2.1.3 Current trends in computer-assisted diagnosis

Web IR systems are becoming increasingly popular for the task of diagnosing difficult cases [10, 18, 32]. These systems are easy to use, fast, accessible, and their databases are continuously updated.

The two main differences between web IR systems and medical database systems are: the method of entering patient data, and the matching algorithms they use. While most of the medical database systems take as input complex structured queries requiring expert training, web IR systems simply accept free-text queries. Moreover, medical database systems often return only results that exactly match the user query, whereas web IR systems use approximate matching algorithms. This is especially important for difficult cases where symptoms can be missing or misleading. For example, searching to solve a difficult case using PubMed usually requires the use of boolean operators, as by default the results must match all query terms.

Currently, the most popular web systems used by clinicians are general search engines such as Google, medical websites such as UpToDate, Medscape, or WebMD, and medical database search tools such as PubMed [7, 8, 43]. A recent study reported that the majority of medical personnel used electronic medical resources in their day-to-day work, and that Google was the preferred resource, with 82% of the physicians using it, followed by PubMed, with 74% [7].

Despite the existence of specialized systems such as Orphanet, OMIM, Phenomizer or Possum, the general web search engine Google is repeatedly mentioned in literature as a valuable tool for diagnosing difficult and rare disease cases [6, 10, 18, 44]. Among the advantages of using Google in this setting are its comprehensive index[12], its ease of use, and medical personnel's familiarity with it. Its main disadvantage in the scope of clinical diagnosis is that the results contain noise, many of the results being non-relevant (e.g.

[12] `http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`

pages from forums and personal blogs).

The problem with general search engines in the context of clinical diagnosis is that they are designed and optimised for web search. For example, the list of patient symptoms can become very long in some cases, while web search engines are optimised for short queries of two or three terms. Popularity boosting (e.g. hyperlinking, PageRank, user visit rates) is, again, not appropriate in the case of rare diseases, where rare disease articles are sparse and not popular. Moreover, these systems are designed for optimal matching, where documents containing all search terms are ranked higher. This is not necessarily appropriate for the task of diagnosing, as symptoms may be misleading or some patient data may be irrelevant for solving the case.

Even if popular, searching for diagnoses in Google or PubMed is still time-consuming, so a specialized search engine could decrease search time and improve performance.

2.2 State of the Art in Information Retrieval

A *search engine* application is grounded on theoretical *information retrieval* (IR) concepts that deal with information analysis, storage, and retrieval. Today, the most widespread use of search engines is in the web space, where general purpose search engines have become to define the way people access information. Beside these general purpose web search engines, a wide variety of other IR systems exist: engines for vertical search, enterprise search, bibliographic search, desktop search.

2.2.1 Information retrieval and document ranking

The vast majority of IR systems deal exclusively with text documents (e.g. web pages, papers, books, ontologies), but increasingly also involve other types of documents (e.g. images, videos, or audio material). For the task of diagnosing rare diseases, the primary sources of information are text-based resources such as published articles describing cases of rare disease patients, web pages discussing the phenotype of rare diseases, or rare disease databases maintained by medical professionals and organizations. However, given that most of the rare diseases have a genetic origin (80%) and that these often cause dysmorphological features, it is reasonable to assume that an additional database of photographs showing the main dysmorphic features of syndromes[13] can be used in searching for a rare disease diagnosis.

IR systems solve tasks such as ad-hoc search, classification, or question answering [45]. Ad-hoc search pertains to systems that take user queries

[13]London Medical Databases (LMD), Winter-Baraitser Dysmorphology Database (WBDD), `http://www.lmdatabases.com/`

as input, classification systems group items according to their content or attributes, and question answering systems take user queries formulated as questions and use natural language processing (NLP) to interpret them and return answers.

IR systems use a data structure called *index* to store the document collection and improve the speed of search. For fast full-text searches, the *inverted index* stores an *inverted list* for each word consisting of references to documents and the positions of each word in each of these documents. Because it transforms document-word into word-document information (thus the name inverted), the system can quickly evaluate the search query by directly locating the documents containing the search terms and then ranking the identified documents accordingly. To increase the likelihood of matching query terms to terms from documents, *stemming* is often used. A stemmer basically replaces members of a group of words to the base word (stem), for example, the words "disease", "diseases", and "diseased" are all stemmed to "diseas".

In IR, the goal is to retrieve *relevant* documents, that is, documents that are deemed of interest for the submitted search query. To address this, several metrics are used in measuring the relevance of the retrieved documents. *Precision and recall* are the most common. Precision refers to the proportion of retrieved documents that are relevant, and recall measures the proportion of relevant documents that are retrieved. To measure these scores, experimental evaluations use *test collections* that consist of a document collection, a sample of queries and, if available, a list of relevant documents for each of these queries (called relevance judgements). In web search, measuring recall is more problematic, as there is usually no knowledge of all relevant documents that could be retrieved for a given query.

The process of matching documents to queries is formalized by the *retrieval models*. *Ranking algorithms* are built on top of retrieval models and are used by the search engines to rank documents and return the list of the highest ranking documents for a query. Historically, the Boolean and the vector space models were used [45], but today the state of the art is represented by the probabilistic models, which replaced the use of other models in practice.

The state of the art in retrieval models is ranking based on *language modelling*, which is also a probabilistic model although it is classified separately. In a language modelling setting, given a query q, each document d is ranked according to the probability of generating the query terms from the document's language model D ($P(q|D)$). This is also known as the *query likelihood retrieval model*. [45]

2.2.2 Vertical search engines

Vertical search engines are specifically designed for retrieval on a particular topic. As they are narrower in scope than the general purpose search engines, the document collection is highly focused on a topic, their interface is tailored for the tasks associated with that topic, and they can take advantage of domain-specific knowledge. Therefore, they usually provide better precision and perform better on user tasks than general purpose search engines. On the topic of rare diseases, there are a limited number of vertical search engines, such as the Rare Disease Communities' Custom Search Engine[14], or the Raredisease.org search engine[15]. These search engines are constructed using the customization tools offered by the major web search engine providers. Even if they are limited to a number of topically relevant web resources, the core search technology used for the customized search engines is the same as for the general-purpose products offered by the same providers. As a result, these vertical search engines are not tailored for use in the diagnostic process.

2.2.2.1 Custom search providers

As part of their efforts to provide customized search solutions, Google developed the Google Custom Search Engine (Google CSE) product. This product allows web developers to select a set of web resources indexed by Google from which to limit the retrieval of documents. Although custom search engines are easy to create, they are also limiting in many aspects, for example, the user cannot supplement the index with additional materials not indexed by Google, and cannot modify the ranking algorithm, or rerank the results returned by the search[16]. However, the user has the option of limiting the retrieval to the small set of selected web resources, or perform retrieval on all web resources indexed by Google but emphasize the documents from the set of selected resources. On the topic of rare diseases, such a Google CSE exists (raredisease.org) and it is restricted to 17 websites with content related to rare diseases.

Besides Google, both Yahoo![17] and Microsoft Bing[18] provide APIs for the customization of their search engines. Unlike Google CSE, these APIs do not restrict the reranking of the results their return.

[14]Rare Disease Communities, Custom Rare Disease Search Engine, `http://www.rarediseasecommunities.org/en/search`, searches the following four websites: `eurordis.org`, `orpha.net`, `rarediseases.org` and `rarediseases.info.nih.gov`

[15]Rare Disease Search Engine that uses Google CSE, `http://www.raredisease.org/`

[16]The user is not allowed to "edit, modify, truncate, filter or change the order of the information contained in any Results", Google CSE Terms of service, 1.4 Appropriate Conduct, `http://www.google.com/cse/docs/tos.html`

[17]Yahoo! Search BOSS, `http://developer.yahoo.com/search/boss/`

[18]Bing API 2.0, `http://www.bing.com/developers/`

2.2.3 The Lemur Project

Although there is no standard toolkit for developing IR research projects, there are several viable options that could have been used for developing the vertical search engine[46]. We have chosen the Lemur Project because of its permissive open source license (BSD), because it is actively developed[19], scales up for tens of millions of documents[20], and provides competitive efficiency and effectiveness results [47]. Other state of the art IR systems include Lucene[21], Ivory[22], Terrier[23], Zettair[24] or MG [48].

The Lemur Project develops an open source search engine called Indri, that was designed for building IR systems that use state of the art probabilistic models and language modelling functions [49].

2.3 Medical Information Resources

Several medical information sources that are of interest for this work were identified. Although varied in size and scope, many of the medical resources are interconnected through medical classifications and ontologies.

2.3.1 Resources on rare diseases

When searching for rare diseases information, resources can be divided into web databases targeted for use by medical professionals, and more patient-oriented resources such as websites and blogs providing support for patients suffering from a rare disease or their relatives and friends.

Patient support

The European Organisation for Rare Diseases (EURORDIS) is the largest European network of patient organizations active in the area of rare diseases, representing more than 470 rare disease organizations in 45 countries. Its objective is to raise awareness of the impact of rare diseases and to improve the quality of life of those people suffering from a rare disease[25]. While providing the patients with access to specialized knowledge, EURORDIS also coordinates and makes available the undergoing research efforts for rare disease conditions, frequently releasing studies concerning the status of rare diseases in Europe. Similarly, in the US, the National Organization for Rare

[19] Lemur development, `http://sourceforge.net/projects/lemur`
[20] Lemur Project: `http://www.lemurproject.org/indri.php`
[21] Apache Lucene: `http://lucene.apache.org`
[22] Ivory, `http://www.umiacs.umd.edu/~jimmylin/ivory/docs/index.html`
[23] Terrier IR Platform, `http://terrier.org/`
[24] Zettair, `http://www.seg.rmit.edu.au/zettair/`
[25] EURORDIS, `http://www.eurordis.org/who-we-are`

Disorders (NORD) is dedicated to assisting rare disease patients, patient organizations and medical health care providers[26], and in Canada, there is the Canadian Organization for Rare Disorders (CORD)[27]. NORD, however, also provides rare disease descriptions aggregated in a 1,200 diseases database, while on the European side, Orphanet is the major rare diseases information resource provider. Moreover, many European countries have developed specific policies on rare diseases and opened local information and assistance centres, and some have constructed rare disease databases in their national language[28]. Other forms of support information include blogs (rarediseaseblogs.net, a joint EURORDIS-NORD project), European research projects websites (dyscerne.org), national clinics and patient groups (Rare Disorders Denmark, sjaeldnediagnoser.dk), or committees reports (Rare Diseases Task Force, rdtf.org).

Rare and genetic disease databases

The largest web databases focused on rare diseases are the ones provided by NORD and Orphanet[29], but if genetic diseases are also considered (80% of the rare diseases have genetic origin), an important resource is the database maintained by the Genetic and Rare Diseases Information Center (GARD)[30]. Other high-quality information resources focused on rare and genetic diseases are described in Section 3.1, as they were used in the development of the vertical search engine. Many rare diseases or subgroups and types of diseases have dedicated webpages that explain their phenotype and management, and are maintained by specialized patient organizations, medical doctors, or by patients suffering from a rare disease[31]. Resources related to this specialized topic are not limited to textual information: the Winter-Baraitser dysmorphology database includes photographs showing dysmorphic features of syndromes, and the Birth Defects Encyclopedia (BDE) has over 1700 illustrations for articles on a variety of syndromes.

2.3.2 Medical databases

One of the largest medical database is the MEDLINE/PubMed database provided by the National Library of Medicine (NLM) of the National Institutes of Health (NIH)[32], and includes around 20 million citations and ab-

[26]NORD, `http://www.rarediseases.org/about/vision-mission`

[27]CORD, `http://www.raredisorders.ca/aboutUs.html`

[28]EURORDIS News, National Reference Centre for Rare Diseases, `http://www.eurordis.org/content/learning-each-other-across-europe`

[29]Orphanet Alphabetical Disease Search List, `http://www.orpha.net/consor/cgi-bin/Disease_Search_List.php?lng=EN`

[30]GARD, `http://rarediseases.info.nih.gov/GARD/AboutGARD.aspx`

[31]Abetalipoproteinemia Foundation, `http://www.abetalipoproteinemia.org`

[32]NLM, `http://www.nlm.nih.gov/`

stracts from MEDLINE and other biomedical and life science journals. Of these, the full text of 2,2 million articles is freely accessible through PubMed Central (PMC). While users such as clinicians can access the information through the provided web user interface, the bibliographic information can also be fetched through the Entrez programming utilities[33] or downloaded through FTP[34]. However, the full text cannot be downloaded for all articles, but only for a subset of around 230,000 articles contained in the PMC Open Access Subset[35]. NLM also provides a range of other medical or biological related databases[36], such as Bookshelf, a collection of full-text online biomedical books, the Database of Genomic Structural Variation (dbVar) containing genomic variations information, the Genetics Home Reference (GHR), and Online Mendelian Inheritance in Man (OMIM).

2.3.3 Medical classifications and ontologies

Each bibliographic reference in MEDLINE is indexed with NLM's Medical Subject Headings (MeSH) controlled vocabulary thesaurus[37]. The articles are manually associated with a set of MeSH terms describing their content, and, when searching on the MEDLINE/PubMed database, the query terms are expanded using this vocabulary. However, the hierarchical structure of the 26,140 descriptors in MeSH is not the single classification that can be used for medical text annotation. The Unified Medical Language System (UMLS) Metathesaurus is a large (around 9 million distinct concept names), multi-lingual (21 languages) vocabulary database describing the relationships between biomedical and health related concepts[38]. UMLS also gives access to a comprehensive clinical terminology, Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT)[39], and to mappings into the International Classification of Diseases, editions 9 and 10, Clinical Modifications (ICD-9-CM and ICD-10-CM)[40].

Orphanet has created a clinical poly-hierarchical classification of rare diseases based on the medical speciality managing the different aspects of rare diseases[41]. For example, a rare disease can be categorized using this Or-

[33] EFetch for NLM Databases, `http://eutils.ncbi.nlm.nih.gov/corehtml/query/static/efetchlit_help.html`

[34] Access Instructions for NLM Data, `http://www.nlm.nih.gov/bsd/licensee/access/`

[35] PMC open access articles, `http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/#XML_for_Data_Mining`

[36] NLM databases, `http://www.nlm.nih.gov/databases/`

[37] MeSH, `http://www.nlm.nih.gov/pubs/factsheets/mesh.html`

[38] UMLS Metathesaurus, `http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html`

[39] SNOMEDCT, `http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html`

[40] ICD-10-CM, `http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ICD10CM/`

[41] Orphanet Classification of rare diseases, `http://www.orpha.net/data/patho/Pro/en/OrphanetClassificationRareDiseases.pdf`

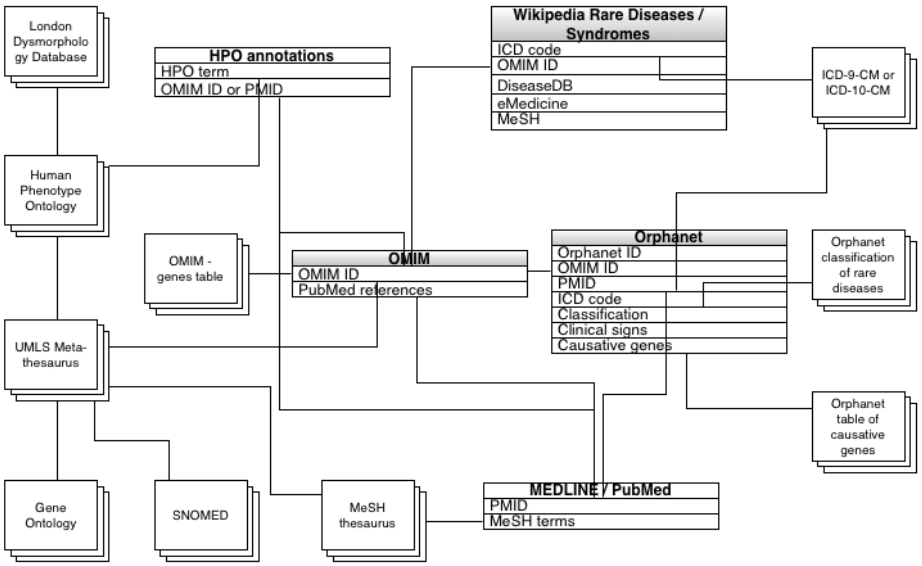


Figure 2.2: **Authors' depiction of the interconnections between several medical databases, ontologies and classifications**. Marked with grey are those resources that were indexed in the vertical search engine.

phanet classification both as a rare neurologic disease and a rare hematological disease. With a focus on genetic diseases, the Human Phenotype Ontology (HPO) maps phenotypic abnormalities to OMIM records, genes, and entries from the London Dysmorphology Database[42]. HPO is used by Phenomizer, which is a tool designed for clinical diagnosis in human genetics that matches HPO terms to diseases corresponding to OMIM entries.

These classifications, as well as the medical databases discussed in the previous section, are interconnected through identification references, such as OMIM numbers, ICD codes, or UMLS Metathesaurus identifiers (as seen in Figure 2.2).

[42]HPO, `http://www.human-phenotype-ontology.org/index.php/hpo_home.html`

Chapter 3

Methodology and Design

3.1 Rare Disease Information Resources

On the Internet, one can find numerous resources on rare diseases, but care must be taken to prevent the selection of possibly low quality material such as patient blogs, web forums, and low-quality commercial sites. The following websites have been identified by the authors to provide alphabetically-sorted lists of rare and genetic diseases information and were subsequently used in the IR system. Each disease entry contains one or more of the following fields of information: disease name synonyms, symptoms, diagnostic process, treatment, number of cases, organizations related to the disease, research studies conducted for the disease, related articles in medical journals and more. See Table 3.1 for details on what type of information is provided by each of the resources.

Orphanet The portal for rare diseases and orphan drugs
`http://orpha.net`;
the leading resources on rare diseases in Europe; the information is based on published scientific articles and updated on a regular basis; the disease reports are peer-reviewed; database of around 6,000 rare diseases.

NORD National Organization for Rare Disorders
`http://rarediseases.org`;
the disease reports are written by NORD medical writers and reviewed by physicians (in some cases, the reports are written directly by the physician); database of more than 1,200 diseases.

GARD Genetic and Rare Diseases Information Center, National Institutes of Health
`http://rarediseases.info.nih.gov/GARD/`;
a collaborative effort of two agencies of the National Institutes of

Health, The Office of Rare Diseases Research (ORDR) and the National Human Genome Research Institute (NHGRI) to help people find useful information about genetic conditions and rare diseases; contains information for about 7,100 rare and genetic diseases[1].

Socialstyrelsen The Swedish National Board of Health and Welfare
`http://www.socialstyrelsen.se/rarediseases`;
a government agency under the Ministry of Health and Social Affairs; 265 diagnoses in Swedish and 88 diagnoses in English; the reports are made by medical specialists in cooperation with patient organizations[2].

About.com Rare Diseases Portal
`http://rarediseases.about.com/`;
contains around 550 rare disease pages; the content is reviewed by a medical review board[3]; although all articles are targeted to patients, many of them describe diseases that could be useful in the IR system.

GHR Genetics Home Reference
`http://ghr.nlm.nih.gov/BrowseConditions`;
genetic conditions database; more than 550 health conditions, diseases and syndromes; the information contained in the database is developed by genetic counsellors, biologists, and information scientists[4].

OMIM Online Mendelian Inheritance in Man
`http://www.ncbi.nlm.nih.gov/omim`;
a database of human genes and genetic phenotypes; updated daily; includes around 20,700 entries; edited at Johns Hopkins University School of Medicine.

HON Health on the Net Foundation List of Rare Diseases
`http://www.hon.ch/HONselect/RareDiseases/index.html`;
database of around 180 rare diseases; includes description of diseases and accepted synonyms; provides links to multiple web resources.

Wikipedia Category Rare Diseases
`http://en.wikipedia.org/wiki/CategoryRarediseases`;
provides description for around 400 rare diseases; pages include links to classifications such as ICD-9 or OMIM or other web resources; includes a sub-category for rare cancers.

Wikipedia Category Syndromes
`http://en.wikipedia.org/wiki/CategorySyndromes`;

[1] Source of GARD data, `http://rarediseases.info.nih.gov/GARD/AboutGARD.aspx`
[2] About the Socialstyrelsen database, `http://www.socialstyrelsen.se/rarediseases/aboutrarediseases`
[3] About.com Medical Review Board, `http://www.about.com/health/review.htm`
[4] GHR content, `http://ghr.nlm.nih.gov/about#curation`

	Article title / Disease name	General disease description	Synonyms / Alternative names	Prevalence section / Age of onset	Symptoms section	Diagnosis section	Treatment section	Prognosis section	References	Links to PubMed or OMIM	Links to other web resources	Conferences	Support groups	Clinical trial and research	Inheritance / Genes	Subdivision / Classification	Visible update date	Well-defined structure	Multiple languages
Orphanet	x	x	x*	x*	x*				x	x*	x*		x*	x*	x*	x*		Yes	Yes
NORD	x	x	x										x*			x		Yes	No
GARD	x	x	x						x	x*	x*	x*	x*	x*				Yes	No
Social.	x	x	x	x	x	x	x		x	x	x		x	x	x		x	Yes	Yes
About.com	x	x			x	x	x		x		x*				x		x	No	No
GHR	x	x	x	x					x*	x*	x*		x*		x		x	Yes	No
OMIM	x	x	x		x				x						x		x*	No	No
HON	x	x	x							x*	x*	x*		x*		x*		Yes	Yes
Wiki. RDis.	x	x	x		x	x	x	x	x	x	x				x		x	No	Yes
Wiki. Synd.	x	x	x		x	x	x		x	x	x				x		x	No	Yes
Madisons	x	x	x		x	x	x	x	x		x*				x		x	Yes	No

Table 3.1: **Summary of the information provided by each resource**. Fields marked with * contain the specific information provided by the resource, but this information is not indexed by the developed vertical search engine.

provides description for around 490 medical syndromes, many of which lead to rare diseases; provides links to external medical classifications and resources.

Madisons Madisons Foundation M-Power Rare Pedriatic Disease Database `http://www.madisonsfoundation.org/`;
around 520 diseases with symptoms, prevalence, available treatments, possible causes, prognosis, and links to other resources; all entries have references; the organization has a medical advisory board that oversees the information contained in the database.

3.2 Data Acquisition

The information resources used in the IR system were collected from various sources, as presented in the previous section, and provide rare disease articles that are heterogeneous in quality, length, and authority. In total, the corpus contains around 31,590 documents that were retrieved from eight online medical resources and two medical databases.

These resources were selected to be used based on several factors. First of all, the topic of the articles must be focused on rare diseases or genetic disorders. Secondly, each article must describe, more or less, one particular

disease. However, resources discussing only one disease or a restricted category of diseases were not included. Moreover, the publisher of the documents must be an authority in the medical field (this would exclude blogs, forums and other web pages with unverified content). The selected web sources are maintained by governmental organizations, patient groups, medical specialists, or other trusted parties. Overall, the websites had to contain high quality articles on rare or genetic diseases, with original content, written by specialists or properly referenced.

A web scraper was developed for retrieving part of the collection of medical documents; specifically, for scraping the articles included in NORD, GARD, Socialstyrelsen, About.com Rare Diseases, GHR, HON, and Madisons collections. The hierarchical structure of the web resource was identified and given to the scraper together with a set of rules to restrict scrapping to the relevant articles from the hierarchy. All articles matching the restrictions were saved for later processing.

For retrieving the Wikipedia articles, the MediaWiki API[5] was used in order to extract the XML files for the articles under the wiki categories Rare Diseases and Syndromes.

For the two other resources, OMIM and Orphanet, the collection of articles was downloaded from the server, and received by email on request, respectively. The articles stored in OMIM were provided in flat text format, and the ones from Orphanet were stored in spreadsheets.

3.3 Data Transformation

The files retrieved as a result of the text acquisition process were further transformed into a standardized format for indexing - the TREC format. The textual information from all sources was tagged with document number, article title, URL, and article text (Listing 3.1). These entities are used by other components of the IR system for rapidly accessing the information contained in the documents.

As many of the documents were extracted from web pages, they were all structured differently and needed to have their structure extracted and converted to the TREC format. Web scrapping was performed by specifying which structured elements (HTML tags) mapped to the desired TREC tags.

Documents where important structure elements were missing, such as the title element, were ignored. For each resource, we constructed two kinds of structural rules. First, the mandatory structural elements rule, specified a set of elements of which none should be missing. Secondly, the optional structural elements rule, where at least one structural element should be present. Documents that did not comply with both rules were discarded.

[5]MediaWiki API, `http://www.mediawiki.org/wiki/API:Query`

```
<DOC>
<DOCNO>18921</DOCNO>
<TEXT>
<URL>rarediseases.info.nih.gov/GARD/Condition/5787/
Alstrom_syndrome.aspx</URL>
<TITLE>Alstrom syndrome</TITLE>
<DESCRIPTION>Alstrom syndrome is a rare disorder
characterized by ...
</DESCRIPTION>
</TEXT>
</DOC>
```

Listing 3.1: **Snippet of a TREC-formatted document**

3.4 Index Creation

The output of the text transformation component was stemmed and then indexed for document retrieval. The Krovetz stemmer was used for grouping words derived from the same stem, by converting plural form to single form (e.g. -s, -es), converting from past to present tense (e.g. -ed), and by removing the -ing suffixes [50]. The index was created on the transformed TREC-formatted documents using the built-in functions provided by the Lemur Project.

Two indexes were constructed based on the processed documents. The first index, named *Rare* uses the sources that are mostly focused on rare diseases, and excludes the resources focused only on genetic diseases. The second index, named *RareGenet* uses all resources. The first index includes 10,263 documents, while the second comprises of 31,590 documents (Table 3.2). The reasons for creating two indexes are to evaluate the variations in performance given different index sizes and coverage of information.

3.5 User Interaction

The specialized search engine takes as input some textual patient data, such as symptoms, test results, demographic information, and returns a ranked list of potentially relevant documents on the topic of rare diseases. The patient data is entered as a query in the search engine interface, the query is then processed and transformed into index terms, and finally the ranked results are returned to the user through the interface. Figures 3.1 and 3.2 show how the user interaction with the system works.

To facilitate the usage of the system, the interface design is simple and straightforward, similarly to popular search engines with which most clinicians are familiar. It provides a search box that gains focus on page load and auto-expands on long inputs if the clinician decides to enter more detailed

	Rare	RareGenet
Vocabulary		
Term Count	2484358	25778103
Unique Terms	57450	319681
Document Count	10263	31590
Resources (article count)		
NORD (1230)	Yes	Yes
Orphanet (2830)	Yes	Yes
GARD (4578)	Yes	Yes
Socialstyrelsen (114)	Yes	Yes
About.com (316)	Yes	Yes
HON (183)	Yes	Yes
Wiki. Rare Diseases (500)	Yes	Yes
Madisons (522)	Yes	Yes
Wiki. Syndromes (334)	No	Yes
GHR (626)	No	Yes
OMIM (20369)	No	Yes
Storage		
Raw Size	543 MB	719 MB
TREC Size	17 MB	162 MB
Index Size	28 MB	227 MB

Table 3.2: **Repository statistics**. Vocabulary, resources and storage statistics for the two collections indexed in the IR system.

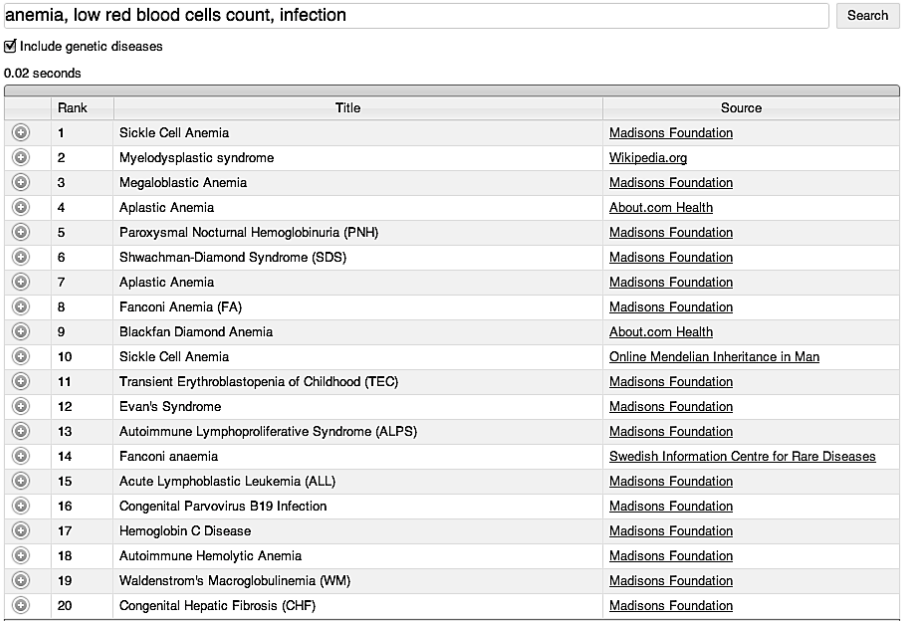
patient data. The search is initiated by either pressing ⏎ or by clicking the search button, and the query results are typically generated in under 0.1 seconds.

The list of results is presented inside a flexible widget which initially only lists the rank, article title, and source, the latter of which is a clickable link that opens the original article in a new browser window. In this simple view, a clinician would get an overview of the most relevant diseases for the given query (Figure 3.1).

If a more detailed analysis is required, a clinician could click on any of the results or on the associated plus button, situated at the left side of each entry, to display the entry's details (Figure 3.2). The details consist of the article's complete title, full URL, and the first 400 words of the article content. Multiple entries can be simultaneously opened for details.

The user can select which index is used for retrieval. By default, the retrieval is performed on the *Rare* index, but the user can enable a check-box to perform it on the *RareGenet* index. When the check-box state changes, the search is automatically performed with the new settings.

An alternative experimental variant of the search engine allows users to receive ranked disease names instead of ranked documents as results for search queries (Figure 3.4). The disease ranking is based on the frequency



	Rank	Title	Source
⊕	1	Sickle Cell Anemia	Madisons Foundation
⊕	2	Myelodysplastic syndrome	Wikipedia.org
⊕	3	Megaloblastic Anemia	Madisons Foundation
⊕	4	Aplastic Anemia	About.com Health
⊕	5	Paroxysmal Nocturnal Hemoglobinuria (PNH)	Madisons Foundation
⊕	6	Shwachman-Diamond Syndrome (SDS)	Madisons Foundation
⊕	7	Aplastic Anemia	Madisons Foundation
⊕	8	Fanconi Anemia (FA)	Madisons Foundation
⊕	9	Blackfan Diamond Anemia	About.com Health
⊕	10	Sickle Cell Anemia	Online Mendelian Inheritance in Man
⊕	11	Transient Erythroblastopenia of Childhood (TEC)	Madisons Foundation
⊕	12	Evan's Syndrome	Madisons Foundation
⊕	13	Autoimmune Lymphoproliferative Syndrome (ALPS)	Madisons Foundation
⊕	14	Fanconi anaemia	Swedish Information Centre for Rare Diseases
⊕	15	Acute Lymphoblastic Leukemia (ALL)	Madisons Foundation
⊕	16	Congenital Parvovirus B19 Infection	Madisons Foundation
⊕	17	Hemoglobin C Disease	Madisons Foundation
⊕	18	Autoimmune Hemolytic Anemia	Madisons Foundation
⊕	19	Waldenstrom's Macroglobulinemia (WM)	Madisons Foundation
⊕	20	Congenital Hepatic Fibrosis (CHF)	Madisons Foundation

Figure 3.1: **Ranked list of documents**. Search interface screenshot with the results for the example query "anemia, low red blood cells count, infection" on the *RareGenet* index. The most relevant 20 articles are listed. Each result has a rank, an article title, and a source (e.g. Wikipedia.org). Clicking on the source redirects the user to the originating article. Clicking on the plus sign or the list item itself shows the article's details.

of disease name occurrences in the documents retrieved for the same query.

The results can be saved for later referral or analysis as a PDF file. This file could also be used to print the results.

Every interaction the user has with the system is logged together with the retrieved results. All logged data is aggregated at a session level, so we have an overview of the entire set of actions performed by the user. Additionally, a feedback box is provided at the bottom of the page (Figure 3.3).

3.5.1 Patient data as queries

As clinicians become more and more familiar to using Google and PubMed as search interfaces for medical information retrieval, it may be argued that they are becoming proficient in summarizing a clinical case in just a few keywords.

It has been studied that the patient-centred queries submitted by clinicians using PubMed consist on average of only 2.5 terms [51]. More specifically, 56% of the queries consisted of only 1 or 2 terms, while 98% of them consisted of fewer than 6 terms. The number of terms in the query par-

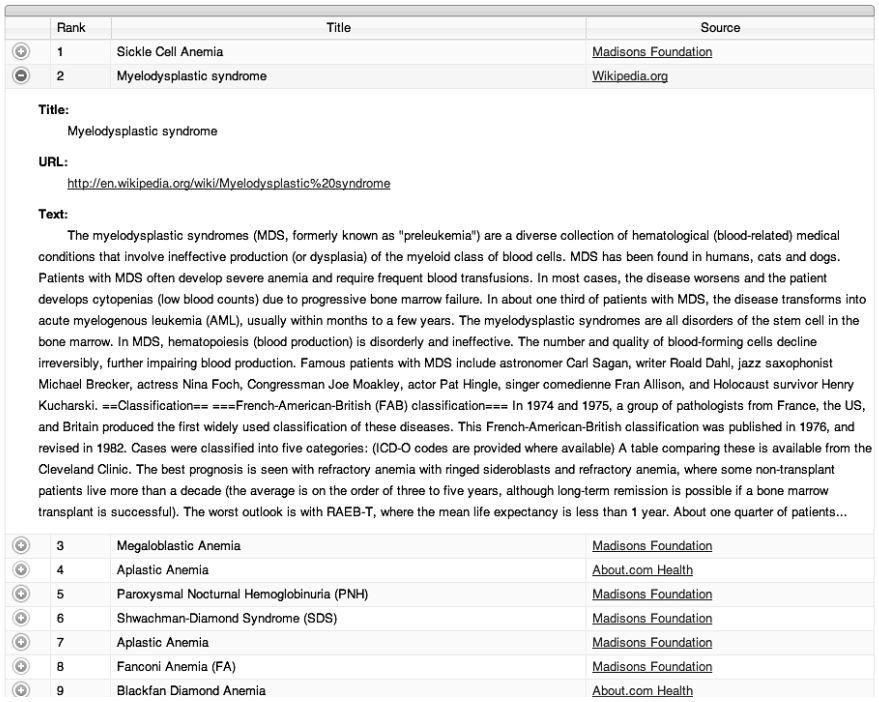


Figure 3.2: **Viewing more details about a document**. Search interface screenshot with the details for one of the results for the example query "anemia, low red blood cells count, infection". Clicking on one of the results will present more detailed information for the selected item. In this example, the second of the most relevant 20 articles was selected. Now, beside the rank, article title, and source, a snippet (the first 400 words) of the article is visible.

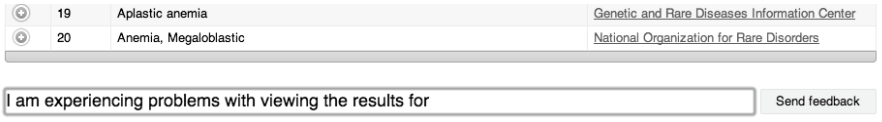


Figure 3.3: **The feedback box**. Positioned at the bottom of the page, under the table listing the ranked results.

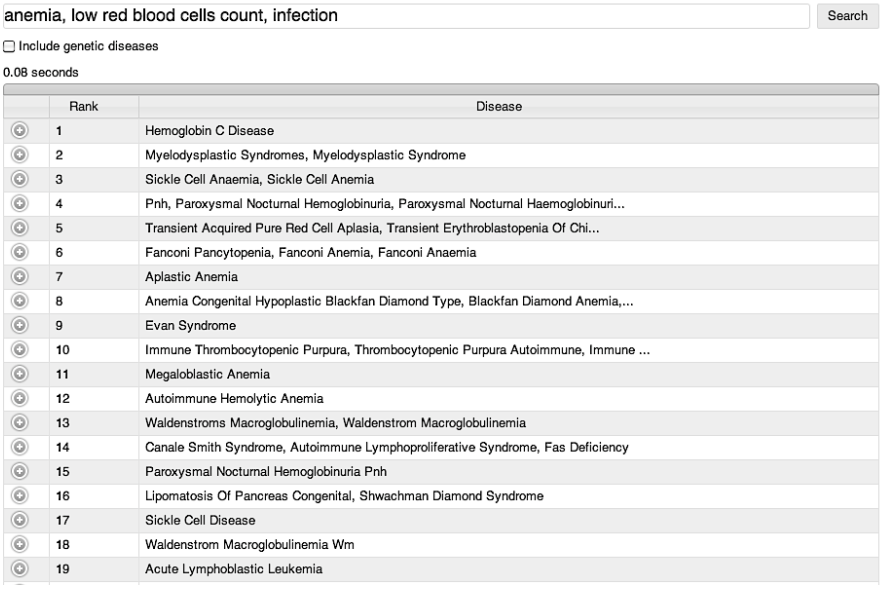


Figure 3.4: **Ranked list of diseases**. Search interface screenshot with the disease names retrieved for the example query "anemia, low red blood cells count, infection". The returned list of diseases provides access to the corresponding list of relevant documents.

tially determines the number of articles retrieved. For a query of only a few terms, a large number of articles are expected to be returned, whereas for queries consisting of more terms, the number of retrieved articles is expected to decrease [51]. This means that using more terms increases the risk of finding no articles at all, but it could be that it also increases the chance of evaluating more relevant articles (as the query might be more accurate). Although this study indicates that PubMed queries in a clinical environment have an average of 2.5 terms, it should be noted that this covered all queries provided to PubMed. It is likely that when looking for a list of diagnostic hypotheses, the clinician would provide more information than for other clinical questions (e.g. medication dosage).

Because the developed vertical search engine accepts free-text input, the patient-related questions that are to be summarized in queries for the search interface can consist of any patient information. This is one of the advantages of using free-text input over using predefined symptoms that need to be selected from a list. The queries can include patient gender, demographic information, symptoms, evidence of diseases, test results, previous diagnoses, and other information that the clinician might find relevant in the differential diagnosis.

3.5.2 Ranked results

Of the 3205 PubMed queries collected in the study mentioned in the previous section, for 81.9% of them only the first ten titles were viewed, and no successive page was selected [51]. We can therefore conclude that 20 should be an adequate number of results that could be reasonably taken into consideration by the clinician. Indeed, in a discussion with a clinician[6], it was confirmed that 20 results are enough given the time constraints in the clinical setting. Popular search engines usually display 10 search results by default.

3.5.2.1 Ranked articles

For each of the maximum 20 results returned for a query, the following information is provided: rank (based on the ranking algorithm described in Section 3.6.1), article title, source (organization or website), URL of the original article, and a snippet of article text (the first 400 words). The purpose of the snippet is to give to the clinician a preview of what the article (hence, the disease) is about, the quality of the source, and to assist in filtering the results. If the user is interested in the full article content, the original document is one click away.

3.5.2.2 Ranked diseases

For the experimental version of the search engine that returns ranked diseases as results for query searches, each of the results provides the following information: rank (based on the ranking algorithm described in Section 3.6.2), the disease name and its synonyms, and the list of titles for those articles returned by the document ranking algorithm that mention the disease.

3.5.3 Programming interface

The system allows third party applications to submit queries and receive the same information provided by the web interface. Currently, XML, HTML and JSON responses are provided, together with the ability to directly request responses as PDF files.

Therefore, the system could be integrated into the existing electronic health record (EHR) systems deployed in many hospitals. One possible scenario would be that the doctor could request a list of probable diseases for a patient from inside the EHR system. In this way, the EHR could automatically input patient data as a query and receive the results in XML, JSON, HTML or PDF format.

[6]Henrik L. Jørgensen, chief physician at Bispebjerg Hospital

3.6 Ranking Process

The ranking component is the core part of the search engine. The documents are ranked according to a ranking algorithm that matches the terms from the query with the terms from the indexed documents.

The user submits a query through the web interface or API to the server. The server initiates the search for the query using the Lemur built-in functions on the selected index. The result of the search is a list of up to 20 internal document reference numbers. These references are used to fetch additional document information, which is then processed for output in the requested output format.

In addition to ranking documents, the search engine provides an experimental disease ranking feature. The list of disease names is computed by first performing document ranking, extracting the disease names mentioned in the first 20 document results, and then sort the extracted disease names as described in Section 3.6.2.

3.6.1 Document ranking algorithm

To rank documents with respect to queries, a probabilistic model is applied to the task of rare disease diagnosis. Given a query (q) consisting of patient data, we would like to compute the probability of the document model (D) being generated by the query ($P(D|q)$) and rank documents based on the probability of generating the terms of the query from the article's language model ($P(q|D)$) [45]. Using Bayes' theorem:

$$P(D|q) = \frac{P(q|D)P(D)}{P(q)} \tag{3.1}$$

which is rank equivalent to:

$$P(D|q) = P(q|D)P(D) \tag{3.2}$$

since $P(q)$ is constant. $P(D)$ is the prior probability of a document, and we assume at this stage that these priors are uniform and thus can ignore them. Hence:

$$P(D|q) \propto P(q|D) \tag{3.3}$$

Estimating the probability of a query being generated by a document corresponds to estimating how likely it is for a document to be about the correct disease from which the patient described in the query suffers.

As we use an unigram language model,

$$P(q|D) = \prod_{i=1}^{n} P(q_i|D) \tag{3.4}$$

where n is the number of query terms, and q_i is a query term, and:

$$P(q_i|D) = f_{q_i,D} \tag{3.5}$$

could be an estimate for $P(q_i|D)$, where $f_{q_i,D}$ is the frequency of query term q_i in document D. However, if a document does not contain at least one query term, $P(q_i|D)$ will become zero. To avoid this problem, we use a smoothing technique. If $P(q_i|C)$ is the probability of the query term q_i in the document collection model C, then $\alpha_D P(q_i|C)$ is the probability estimate of unseen words, where α_D is a parameter. The probability for words that occur is given by

$$(1 - \alpha_D)P(q_i|D) + \alpha_D P(q_i|C) \tag{3.6}$$

Using Dirichlet smoothing, we have:

$$\alpha_D = \frac{\mu}{|D| + \mu} \tag{3.7}$$

where $|D|$ is the length of the document, and μ is tunable. By replacing α_D in Equation 3.6, we obtain an estimate for $P(q_i|D)$:

$$P(q_i|D) = \frac{f_{q_i,D} + \mu P(q_i|C)}{|D| + \mu} \tag{3.8}$$

which leads to the document score being computed by the formula:

$$P(q|D) = \sum_{i=1}^{n} \log \frac{f_{q_i,D} + \mu P(q_i|C)}{|D| + \mu} \tag{3.9}$$

Furthermore, evidence about the collected medical articles can be used to revise the ranking. More specifically, the article's origin can be used to adjust the prior probability of article relevance. Thus, the articles on rare diseases get a positive boost in their relevance prior probability, compared to articles on genetic diseases. This is based on the intuition that documents about rare diseases are more relevant when searching for a rare disease diagnosis.

3.6.2 Disease ranking algorithm

For the experimental version of the vertical search engine that ranks diseases instead of documents, the ranked diseases are extracted from the most relevant 20 documents returned by the document ranking algorithm. The first step is the information extraction process, which consists of matching rare disease names and synonyms from the Orphanet collection[7] with article titles. In the next step, a matrix of candidate diseases and their score is computed using Equation 3.10:

[7]Orphanet rare diseases collection, `http://www.orpha.net/consor/cgi-bin/Disease_Search_List.php?lng=EN`

$$S_i = \sum_{j=1}^{n} f_{ij}, \tag{3.10}$$

where n is the number of documents returned by the IR system (in this case, 20), S_i denotes the score of disease i, and f_{ij} is the frequency of disease i in document j.

3.6.2.1 Step A: Constructing the disease-document frequency matrix for each of the indexes

(1) Extract all document titles (i.e. disease names) from an index. Keep corresponding document numbers. Remove the duplicates of existing titles.

(2) Use the Orphanet synonyms database to add synonyms to the collection of disease names. (If a disease name and its synonyms from orphanet database was not found in the existing collection of titles, it will not be added)

(3) For each disease (and its synonyms), get their frequency in documents.

3.6.2.2 Step B: Ranking diseases

(1) From the top 20 ranked documents, give each document a weight of 1 to share among all occurrences of rare disease names inside that document, and sum the weights attributed to each of the diseases over all 20 documents (Equation 3.10).

(2) Sort all diseases with $S_i > 0$ by their score (S_i).

3.7 Query Extraction

Two query collections were used in the evaluation of the system, one extracted by the authors, and the other extracted by a previous study [18]. In both query collections, the selection of query terms was based on previously published patient case reports. The use of case reports is motivated by the lack of a publicly available test collection for rare diseases.

3.7.1 Rare diseases query collection

Twenty-five queries, each consisting of a series of symptoms and medical observations describing the phenotype of a rare disease, were extracted from 20 case reports published in Orphanet Journal of Rare Diseases (OJRD)[8],

[8]Orphanet Journal of Rare Diseases (OJRD), `http://ojrd.com/`

an online open access journal providing full text articles on rare diseases and drugs to treat them. At the time when the selection of case reports was made, 22 case reports were available in the journal. Two of the reports were excluded because one presented symptoms of multiple patients, and the other presented the usage of a drug in an atypical presentation of a common disease. From the 20 case reports, five of them discussed two cases of patients presenting the same rare disease, thus the total number of queries extracted from the journal reached 25.

The 25 queries were formulated by two non experts (the authors) and validated by a clinician. The queries consist of the initial findings and the symptoms that the patients presented before any genetic or more elaborate mean of testing (that could directly point to the rare disease) was performed. Physician Henrik L. Jørgensen provided feedback on the initial query formulations, guiding in adding or removing some of the medical observations in two of the queries, based on what knowledge of the patient a physician would have before making some tests or further assumptions in the differential diagnosis of the rare disease. This collection of queries will be referred to as the *OJRD query collection* from now on.

Another five queries were previously extracted as described in [52] from five case descriptions proposed by the same physician. Each case is associated with a different rare disease. The queries were extracted by two non experts but were not validated by a physician. These five cases will be referred to as the *5-cases query collection*. Together, the OJRD and 5-cases query collections form the *rare diseases query collection*.

For each case, the query terms and the associated correct diagnosis were entered on a sheet. Synonyms for the correct diagnosis were also added to the sheet if they appeared in the Orphanet dataset. The average query length for the rare diseases query collection was 22.17 terms. Appendix A provides the full table of queries, corresponding diagnoses and their synonyms, and the source of the rare disease cases used in the evaluation.

3.7.2 Difficult cases query collection

Twenty-six cases extracted from the New England Journal of Medicine (NEJM) as described in the British Medical Journal (BMJ) [18] were also used in evaluating the performance of our system in comparison with systems that were previously described to be useful aids in the diagnosis of difficult cases (such as Google [18] or PubMed). The queries, extracted by a physician together with a rheumatologist, include three to five terms for each case record. The cases are supposed to be difficult to diagnose, as they were deemed interesting enough to be included in NEJM.

For each of the 26 cases, from now on referred to as *difficult cases query collection*, we entered the case synopsis, query terms and the assigned correct diagnosis on a sheet, based on the table provided by the original BMJ

article. Similarly to the rare disease cases described above, we searched for alternative disease names for each final diagnosis and added them in the sheet.

Moreover, we verified whether the final diagnoses of these difficult cases are actually rare disease by searching their disease names on Orphanet. Indeed, in 84.62% of the cases (22 cases), the correct diagnosis was a rare disease. The average query length was of 5.0 terms.

3.8 Evaluation Methodology

In order to compare our system to other systems currently used by clinicians, an evaluation methodology was established. Ideally, the experimental evaluation for such a system should involve clinicians, but as this was not possible as part of this work, we focused on creating an evaluation methodology that could be easily replicated.

3.8.1 Relevance assessment

Two non-expert evaluators (the authors) queried the vertical search engine with each of the query collections, and collected the results in a sheet for later analysis. The correct diagnosis was not included in the search terms.

For a document to be considered relevant it must predominantly cover the correct disease or one of its synonyms. A document is also considered relevant if it predominantly covers a different form or type of the correct disease, such as 'Loeys-Dietz type 1A' instead of the correct 'Loeys-Dietz syndrome type II', or the 'X-linked Adrenoleukodystrophy' instead of 'Autosomal Neonatal Form of Adrenoleukodystrophy'.

In contrast, a non-relevant document is one where the correct disease cannot be identified after a couple of minutes of reading. That is, the correct disease should be mentioned in the beginning of the article, its title, or in the snippet[9]. The first 400 words threshold was established. If an article mentions the correct disease after the threshold, given the time constraints in the clinical setting, it is reasonable to presume that there is a good chance the clinician will not see it. In the case of articles that list multiple diseases, a threshold of the first ten listed diseases was established. Moreover, a document with restricted access is deemed not relevant if it can not be considered relevant given only the freely available information.

An intermediary class of relevance was considered. A marginally relevant document must predominantly cover other diseases or class of diseases than the correct one, but it must mention the correct disease as an alternative diagnosis or point to it in the beginning (title, snippet, first 400 words, or in the first 10 members of a list).

[9]The snippet returned by our system contains the first 400 words of the article.

For the results returned by the system using the experimental disease ranking algorithm, relevant results should include the correct disease name, mention one of the disease's synonyms, or name a different form or type of the correct disease.

The performance of the developed system was measured against Google Search[10], two instances of Google Custom Search[11], and PubMed on the same set of queries, using the same evaluation methodology. To avoid previous searches influencing the results, all searches on Google were made after clearing browsing data and logging off any Google account.

The first three pages of results for each query, with 10 results per page, were saved and evaluated for all Google searches. Three pages of results were saved because many of the results pointed to the original articles from which the queries were extracted. These articles were eliminated from the results, and the new top 20 documents were evaluated.

Google Search imposes a 32-word limit (after the elimination of stopwords) for the search query and truncates queries exceeding this limit. Only 2 out of the 30 queries in the rare diseases query collection were truncated. For all experiments, the system was evaluated on the truncated versions of those two queries.

On PubMed, the first 50 results were saved for each query and, as in the case of Google, the original articles from which the queries were extracted were discarded. The remaining top 20 articles were evaluated. The search was performed using the default settings. Notably, by default, PubMed displays the retrieved documents in reverse chronological order of their publish date.

3.8.2 Measuring search time

It is difficult to reliably assess the time it would take a clinician to find the correct diagnostic hypothesis using a web search tool without the help of medical personnel. However, we tried to overcome this problem by counting the number of words to read and clicks to press until arriving at the correct diagnosis. To simplify the evaluation, we assumed that the user would go through the results page from start to finish, stopping when the correct disease is read. If the correct disease is not mentioned in the results page, then the reader would start following the links associated to the results, in the order they were ranked. We consider that each click to an article would take at least one minute of the user's time before an assessment of relevance is made.

[10]Google Search, `http://www.google.com/`

[11]Google CSE, `http://www.google.com/cse/`

Chapter 4

Results

4.1 Efficiency and Effectiveness

As part of the evaluation process of the system discussed in this work, we performed both efficiency tests and effectiveness evaluation experiments. The scope of efficiency tests is to elucidate the system's behaviour with respect to time and space constraints. On the other hand, effectiveness measures evaluate the system with respect to its ability to find relevant information.

4.1.1 Efficiency scores

The efficiency of the system can be assessed by measuring the time and space requirements of the data acquisition and indexing processes, and more importantly, the throughput and latency of the querying process.

For all measurements, the testing machine was a Xen[1] virtual instance running the x86-64 version of CentOS 5.5 Linux distribution[2]. The virtual instance was allocated with 1 GB RAM, and 100 GB of disk space, and ran on an Intel Xeon E5530 clocked at 2.40 GHz.

In order to create the index, we first need to download the raw datasets from their respective sources and preprocess them. Measuring this process is important because it is the most time and space consuming. The time dimension of data acquisition is mostly determined by the constraints imposed by the data owners. For web resources, we limit the scrapper to download at most three resources per source every second. To speed up the acquisition time, data downloading and preprocessing are performed in parallel, by creating a separate thread for each resource provider.

The indexing process is performed on a set of TREC-formatted files representing the processed raw data. For this evaluation, the index creation

[1] Xen Hypervisor, `http://www.xen.org`

[2] CentOS, `http://www.centos.org`

	Rare	RareGenet
Querying		
Query throughput	142 queries / min	121 queries / min
Query latency	0.42 s	0.49
Indexing		
Indexing Time	10 s	57 s
Index Size	28 MB	227 MB
Data acquisition		
Download + Transform Time	6949 s (115.8 min)	6940 s (115.7 min)
Download Size	613 MB	791 MB
TREC Transform Size	17.19 MB	162.2 MB

Table 4.1: **Efficiency scores**, measured on the two indexes *Rare* and *RareGenet*.

process was allocated 500 MB of RAM. Stemming was enabled (the Krovetz stemmer algorithm was used).

In order to measure the query throughput and latency, we used the queries from both query collections, the *rare diseases query collection* and the difficult cases query collections, totalling 56 queries. The system was queried 10 times using the 56 queries and the average latency and throughput measures were recorded (Table 4.1).

4.1.2 Effectiveness scores

For the task of diagnosing, medical literature suggests that it is crucial to have the correct disease considered in the set of diagnostic hypotheses (Section 2.1.1.1). As the goal of this system is to generate hypothesis ideas, we consider the presence of the correct disease in the top 20 results as the primary effectiveness measure. We also consider that the rank would influence the chances for a disease to be considered as a hypothesis.

Therefore, in the evaluation of the system, we also employ the following effectiveness scores: mean reciprocal rank (MRR), average precision at ranks 10 and 20 (P@10 and P@20), and the normalized discounted cumulative gain at ranks 10 and 20 (NDCG@10 and NDCG@20) [45]. These scores measure different aspects of the ability of the search engine to retrieve and rank the most relevant documents given a query.

4.1.2.1 Mean reciprocal rank

The reciprocal rank score is equal to the inverse of the rank where the first relevant document was retrieved (marginally relevant documents excluded). Let RR denote the reciprocal rank for a given query, then:

$$RR = \frac{1}{rank_1}, \tag{4.1}$$

where $rank_1$ denotes the rank of the first relevant document retrieved.

The mean reciprocal rank is computed by averaging the reciprocal rank for queries where a relevant document was retrieved. Queries for which no relevant document was retrieved have a reciprocal rank equal to zero.

This relevance measure is suited for our purpose as we are mostly interested at what rank the correct diagnosis is first retrieved. It is important for the correct document to be first mentioned higher in the ranking, because that is when the doctor may consider it as a diagnostic alternative. Given the constraints of the clinical setting, the clinician may not have the time to look over all 20 results. The MRR measure severely penalises queries for which the first relevant document is not returned as the first result.

4.1.2.2 Average precision at rank n

Precision measures the proportion of retrieved documents that are relevant. In the case of a diagnostic system, good precision could be seen as a confirmation of a hypothesis.

Let P denote the precision for a given query, then:

$$P = \frac{|Rel \cap Ret|}{|Ret|}, \tag{4.2}$$

where $|Ret|$ denotes the number of retrieved documents, and $|Rel \cap Ret|$ denotes the number of retrieved documents that are relevant.

Marginally relevant documents are considered non-relevant for this measurement, as we use binary relevance in computing the precision, and are interested only in highly relevant documents.

In the evaluation, precision was measured at two ranks, considering only the topmost results returned by the system. This is called *precision at* n and denoted $P@n$. Considering this, the average precision is computed by averaging the precision values from the rank positions where a relevant document was retrieved.

In the setting of clinical diagnosis of rare diseases, we can argue that precision at rank 10 is more important, as the clinician is expected to be confronted for the first time with unknown diseases, and might need some time to reflect upon a disease and form a hypothesis.

4.1.2.3 Normalized discounted cumulative gain at rank n

Unlike the previous relevance measures, the discounted cumulative gain (DCG) uses a graded relevance. Graded relevance, as opposed to binary relevance judgements, distinguishes between different levels of relevance of a document. We chose to give grade 3 to relevant documents and grade 1 to marginally relevant documents. Marginally relevant documents are those

that, although not on the topic of the correct disease, mention the correct diagnostic as an alternative.

Let DCG_n denote the discounted cumulative gain for a given query at rank n, then:

$$DCG_n = r_1 + \sum_{i=2}^{n} \frac{r_i}{\log_2 i}, \tag{4.3}$$

where r_i denotes the relevance grade for rank i. To take into account that results at lower ranks have reduced influence, their grade is divided by the binary logarithm of their rank.

In order for queries with different number of relevant documents to be compared, we compute the ideal discounted cumulative gain (IDCG) by computing the DCG for the ideal ranking (obtained by a descending sort of the relevance grades). Then, the normalized discounted cumulative gain (NDCG) at rank n is computed by:

$$NDCG_n = \frac{DCG_n}{IDCG_n}, \tag{4.4}$$

4.2 Experimental Evaluation

In order to evaluate the research questions from Section 1.3, the effectiveness of the developed vertical search engine was tested on the two indexes *Rare* and *RareGenet*, and compared the results with the effectiveness of other systems used by clinicians: Google Search, two customized Google search engines, as well as PubMed.

4.2.1 Rare and RareGenet indexes

The evaluation of the search engine on the two indexes, *Rare* and *RareGenet*, had the goal of establishing if the inclusion of genetic disease articles would improve the effectiveness without degrading efficiency, as well as establishing whether by increasing the prior probabilities of the relevance of the rare disease articles improves the overall effectiveness.

The queries were entered as they are listed in Appendix A. For each query, the system returned the twenty top ranked documents.

4.2.1.1 Rare diseases query collection

Comparing effectiveness of Rare and RareGenet indexes

Overall, the system performs better on the *RareGenet* index, confirming the hypothesis that adding genetic diseases, which represent 80% of the rare diseases, would improve the coverage of the system. As such, on the *Rare*

	Rare	RareGenet
Total number of cases	30	30
Correct diagnosis in top 10	20 (66.67%)	21 (70%)
Correct diagnosis in top 11-20	0 (0%)	2 (6.67%)
Correct diagnosis not found	10 (33.3%)	7 (23.33%)
Mean reciprocal rank (MRR)	0.445	0.467
Average precision rank 10 (P@10)	0.123	0.157
Average precision rank 20 (P@20)	0.073	0.105
NDCG@10	0.516	0.423
NDCG@20	0.536	0.493

Table 4.2: **Effectiveness of the two indexes on the *rare disease query collection***. Including the articles on genetic diseases improves the overall performance of the system. Retrieval on the *RareGenet* index results in finding the correct diagnosis in 76.67% of the cases (23 out of 30 cases).

index, the system returns relevant results for 20 queries (66.67%). In contrast, on the *RareGenet* index, this number increases to 23 (76.67%). The results for the *rare disease query collection* are summarized in Table 4.2.

Although most of the effectiveness metrics improve, notably the NDCG scores drop for the *RareGenet* index. This indicates that, for the *RareGenet* index, the ranking of relevant documents deviates more from the ideal ranking than is the case for the *Rare* index document ranking.

It should be noted that for one of the 30 cases the correct diagnosis is not in the *Rare* index; specifically, for case 18-1-1 (correct diagnosis: Ligase 4 syndrome). Thus, it can be argued that this case should be eliminated from the evaluation process. If so, retrieval on the *Rare* index results in 68.96% (20 out of 29) cases with the correct diagnosis in the results.

The hypothesis also holds true when considering the *OJRD* and *5-cases query collections* separately. For the queries from the *OJRD query collection*, retrieval on the *RareGenet* index results in finding the correct diagnosis in 72% of the cases (18 out of 25 cases), while on the *Rare* index it finds the correct diagnosis in 60% of the cases. For these cases, on the *Rare* index, MRR was 0.394, P@10 was 0.116 and P@20 was 0.068. Similarly, for the *RareGenet* index, MRR was 0.459, P@10 was 0.160, and P@20 was 0.106. For the *5-cases query collection* proposed by a physician, the system finds all correct diagnoses (100%) for both indexes.

However, in some cases, for example for query H-5, the rank of the first relevant document drops from 1 to 15 for the *RareGenet* index. Genetic disease articles have in general a higher rank than the rare disease articles and result in lower ranks for the rare disease articles that might be more relevant. On the other hand, the intuition was that including genetic disease articles

will result in more correct diagnoses being found. This was confirmed by the fact that three cases that were not correctly diagnosed using the *Rare* index were found using the *RareGenet* index.

Assigning prior probabilities based on article topic

Motivated by these observations, documents from the *RareGenet* index were assigned prior probabilities of relevance in accordance to the type of disease they cover. Thus, documents that also appear in the *Rare* index were assigned higher relevance probabilities than the rest, the intuition being that rare disease articles are highly relevant for our task. If C denotes the index containing both rare and genetic disease documents, then:

$$P(R|C)x + P(G|C)y = 1, \tag{4.5}$$

where $x = \phi y$ (ϕ is the boosting factor), and $P(R|C)$ (resp. $P(G|C)$) denotes the probability of relevance of all rare disease (resp. genetic disease) documents in the index C.

By giving a four times higher ($\phi = 4$) relevance prior probability to those articles that are about rare diseases, the number of relevant documents in the top ten results increases from a value of P@10 of 0.157 to 0.173, as well as the NDCG@10 from 0.423 to 0.433, indicating that relevant documents are ranked higher for the first ten results. For the less important results from ranks 11 to 20, precision increases from 0.105 to 0.115, while NDCG@20 remains the same. This indicates that relevant rare disease articles that were previously not retrieved are now appearing at lower ranks, and rare disease articles that were previously retrieved at lower ranks are now closer to the top. Despite the better ranking, the number of cases for which the correct diagnosis is retrieved remains the same. Results are in Table 4.3.

These experiments were ran using the query likelihood model with Dirichlet smoothing at default settings. That is, using the Krovetz stemmer, no stop words removal, and a smoothing parameter with the value of 2500 ($\mu = 2500$).

Although we did not systematically evaluate our system on a range of values for μ (Chapter 5) to empirically establish the best smoothing parameter for our indexes, as the μ parameter was tunable in the model, we did perform two additional evaluations for μ values of 800 and 4000 (Table 4.4). Even if the effectiveness of the system is not dramatically improved, the performance with a μ value of 4000 slightly improves the overall effectiveness with the exception of P@10. Combined with the prior probabilities boost factor of 4 for the relevance of rare disease articles, the improvements in performance are even more evident (as seen in Table 4.4). As a result, for all experiments that follow, we have chosen to use a μ value of 4000 for the smoothing.

	RareGenet $\phi = 2$	RareGenet $\phi = 4$
Number of cases	30	30
Correct diagnosis in top 10	21 (70%)	21 (70%)
Correct diagnosis in top 11-20	2 (6.67%)	2 (6.67%)
Correct diagnosis not found	7 (23.33%)	7 (23.33%)
Mean reciprocal rank (MRR)	0.468	0.469
Average precision rank 10 (P@10)	0.167	0.173
Average precision rank 20 (P@20)	0.110	0.115
NDCG@10	0.431	0.433
NDCG@20	0.490	0.492

Table 4.3: **Effectiveness scores obtained by boosting the relevance of the rare disease articles** (with boosting factors ϕ of 2 and 4) on the *rare disease queries collection*. The number of correct diagnoses found in top 10 and top 20 remains the same as for retrieval from index *RareGenet* without using priors. However, the performance scores have slightly improved (the rank of the correct disease is higher and the number of relevant documents has increased).

4.2.1.2 Difficult cases query collection

The evaluation results for the retrieval from *Rare* and *RareGenet* for the *difficult cases query collection* are summarized in Table 4.5.

As noted in the evaluation of the *rare diseases query collection*, retrieval on the *RareGenet* index performs better than using the *Rare* index. The percentage of queries for which the correct diagnosis was found in the results using the *RareGenet* index is 50% (13 of 26 cases). Of these 13 queries, all had the correct diagnosis first mentioned in the top 10 results, and 6 of them had the correct diagnosis first mentioned in the top 5 results. Retrieval on the *Rare* index results in finding the correct diagnosis in 38.46% of the cases (10 of 26). Of these 10 queries, 9 had the correct diagnosis mentioned in first 10 results, and 6 of them in the first 5 results.

Three cases correspond to diagnoses that were not found in either of our document indexes. If we exclude these cases, and consider only the subset of 23 cases from the *rare diseases query collection* for which the correct diagnosis is found in the indexes, retrieval from the *Rare* index results in 43.47% (10 of 23) cases solved, and 56.52% (13 of 23) for retrieval from the *RareGenet* index.

Of the 26 cases included in the *difficult cases query collection*, 22 have been identified as rare disease entries in the Orphanet database. If we consider this subset of 22 rare disease cases, retrieval on the two indexes *Rare* and *RareGenet* results in 45.45% (10 of 22), respectively 59.09% (13 of 22) queries with the correct diagnosis mentioned in top 20 results. Moreover,

$\mu = 4000$	RareGenet	RareGenet $\phi = 2$	RareGenet $\phi = 4$
C.d. in top 10	22 (73.33%)	22 (73.33%)	23 (76.67%)
C.d. in 11-20	1 (3.33%)	1 (3.33%)	0 (0%)
C.d. not found	7 (23.33%)	7 (23.33%)	7 (23.33%)
MRR	0.496	0.481	0.481
P@10	0.147	0.160	0.173
P@20	0.107	0.112	0.122
NDCG@10	0.434	0.438	0.448
NDCG@20	0.505	0.504	0.503
$\mu = 2500$	RareGenet	RareGenet $\phi = 2$	RareGenet $\phi = 4$
C.d. in top 10	21 (70%)	21 (70%)	21 (70%)
C.d. in 11-20	2 (6.67%)	2 (6.67%)	2 (6.67%)
C.d. not found	7 (23.33%)	7 (23.33%)	7 (23.33%)
MRR	0.467	0.468	0.469
P@10	0.157	0.167	0.173
P@20	0.105	0.110	0.115
NDCG@10	0.423	0.431	0.433
NDCG@20	0.493	0.490	0.492
$\mu = 800$	RareGenet	RareGenet $\phi = 2$	RareGenet $\phi = 4$
C.d. in top 10	21 (70%)	21 (70%)	21 (70%)
C.d. in 11-20	0 (0%)	0 (0%)	0 (0%)
C.d. not found	9 (30%)	9 (30%)	9 (30%)
MRR	0.437	0.438	0.438
P@10	0.167	0.170	0.163
P@20	0.103	0.107	0.112
NDCG@10	0.461	0.447	0.434
NDCG@20	0.496	0.497	0.496

Table 4.4: **Effectiveness scores for the *rare disease queries collection* obtained by changing ϕ and μ.** Performance evaluation using boosting factors ϕ of 2 and 4, and μ values of 800, 2500 and 4000.

	Rare	RareGenet
Total number of cases	26	26
Correct diagnosis in top 10	9 (34.61%)	13 (50%)
Correct diagnosis in top 11-20	1 (3.84%)	0 (0%)
Correct diagnosis not found	16 (61.54%)	13 (50%)
Mean reciprocal rank (MRR)	0.158	0.186
Average precision rank 10 (P@10)	0.054	0.073
Average precision rank 20 (P@20)	0.042	0.044
NDCG@10	0.358	0.279
NDCG@20	0.390	0.325

Table 4.5: **Evaluation of *Rare* and *RareGenet* on the *difficult cases text collection*.** Including in the search the articles about genetic diseases improves the performance of the system. Retrieval on the bigger index, *RareGenet*, concludes in finding the correct diagnosis in 50% of the cases.

for *RareGenet*, the MRR score would improve to 0.219 and P@10 to 0.086.

The authors of the original BMJ article providing this query collection extracted three to five search terms for each of the 26 NEJM published case. They had the Google search engine in mind from the beginning, and thus one could argue that these queries were tailored for Google search - short queries consisting of only a few keywords that would "not return a non-specific result" [18].

However, in the clinical setting, at the time and place where diagnostic decisions are made, the clinician has access to a larger amount of patient information that could be relevant, and thus is more likely to introduce a more detailed description of the case. As the authors of the BMJ article also provided the synopses of the NEJM cases, we have designed an experiment to compare the performance of the system on these synopses. The average number of terms in a query from this *difficult cases synopses collection* is 9.38 (it was 5.0 for the *difficult cases query collection*).

For the *difficult cases synopses collection*, retrieval on *RareGenet* results in finding the correct diagnosis mentioned in 34.62% (9 of 26) cases, and on *Rare* in 38.46% (10 of 26) cases. Thus, it performs poorly when compared to the results obtained on the *difficult cases query collection*. However, some of the queries returning relevant results on the synopses did not return relevant results on the *difficult cases query collection*, indicating that a combination of synopses and keywords could perform better together than individually.

4.2.2 Google Search and Google Custom Search

As the Google search engine proved to be the most widely used web tool in the clinical setting, it was evaluated and compared in performance with

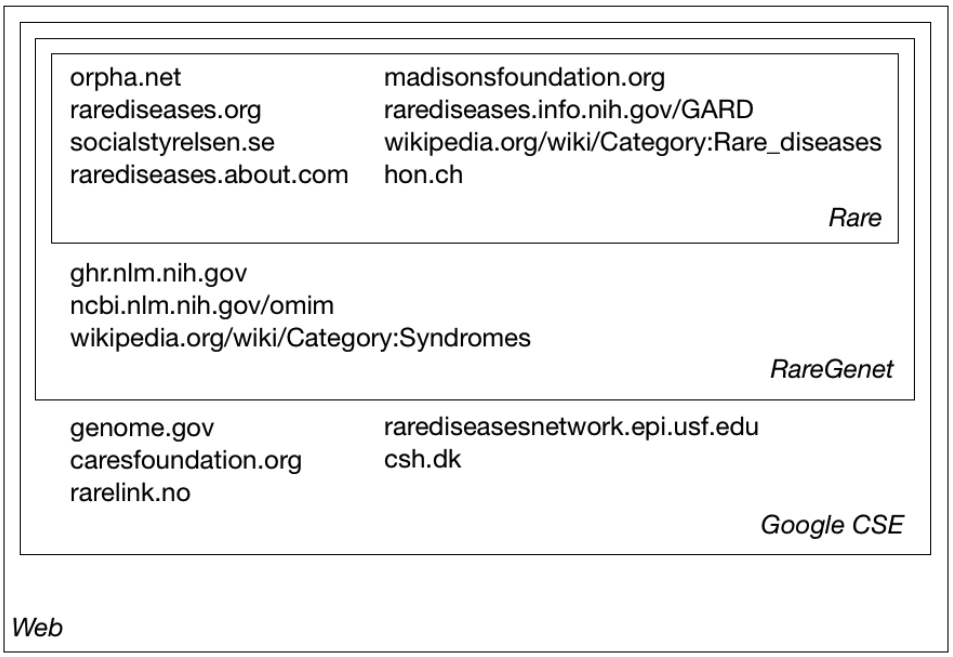


Figure 4.1: **Web resources**. The *Rare* index contains documents extracted from eight rare disease web sources, while the *RareGenet* index also adds documents extracted from three genetic disease web sources. Meanwhile, the customized search engines *Google CSE Restricted* and *Google CSE Web* were customized on these eleven web sources plus five more web pages. The standard Google Search retrieves from the entire web.

our system. Moreover, two versions of Google CSE were customized and evaluated. The first custom search, referred to as *Google CSE Restricted* for the purpose of this evaluation, was restricted on the resources used by our system and five additional web pages. The five additional web pages consisted mostly of links to other web resources and thus were not included in our indexes. The second custom search, referred to as *Google CSE Web*, was set to search the entire web, but emphasize the sources provided to *Google CSE Restricted*. See Figure 4.1 for the list of resources used in customizing the Google search.

4.2.2.1 Rare diseases query collection

As *Google CSE Restricted* was customized to retrieve from a superset of the web resources we used in our system, it is the most similar in terms of index content. However, as shown in Table 4.6, it performs the worst. This could indicate that the algorithms Google uses may be tailored for web search and not for a restricted set of resources. This is confirmed by the performance

	Rare	RareGenet	Google Search	Google CSE Rest.	Google CSE Web
No. of cases	30	30	30	30	30
Corr. diag. top 10	20(66.67%)	21 (70%)	5(16.67%)	1(3.33%)	6(20%)
Corr. diag. top 20	0 (0%)	2 (6.67%)	0 (0%)	0 (0%)	1(3.33%)
Corr. diag. NF	10 (40%)	7(23.33%)	25(83.33%)	29(96.67%)	23(76.67%)
MRR	0.445	0.467	0.056	0.033	0.173
P@10	0.123	0.157	0.023	0.003	0.030
P@20	0.073	0.105	0.013	0.002	0.017
NDCG@10	0.516	0.423	0.168	0.033	0.275
NDCG@20	0.536	0.493	0.189	0.033	0.283

Table 4.6: **Effectiveness of Google on the *rare disease query collection***. Performance comparison between the vertical search engine and Google, Google CSE Restricted, and Google CSE Web.

of *Google CSE Web* which combines Google's general search index with the given custom resources. This hybrid approach performs considerably better than both the regular Google search and *Google CSE Restricted*, finding the correct diagnosis in 23.33% (7 out of 30) of the cases. However, our system manages to find the correct diagnosis in 23 cases (76.67%).

The performance of the Google Search and the two customizations of Google CSE is similar on all query collections. For the 25 cases from the *OJRD collection*, retrieval on *Google Search* results in the correct diagnosis being found in 3 cases (12%), the *Google CSE Web* results in finding the correct diagnosis in 4 cases (16%), while *Google CSE Restricted* does not find any of the correct diagnoses (0%). Our system finds the correct diagnosis in 18 cases (72%). Similarly, for the *5-cases query collection*, retrieval on the general *Google Search* results in finding the diagnosis in 2 cases (40%), the *Google CSE Web* finds the correct diagnosis in 3 (60%), and *Google CSE Restricted* in 1 (20%). This is in comparison with our system that finds the correct diagnosis for all five cases.

One of the issues that seems to affect Google's performance is the length of the queries from our query collection. As Google is focused on general web search, it is a reasonable assumption that, in that setting, most queries will be much shorter. However, a clinical case deemed difficult will probably be described in more than a few keywords. Thus, given the relatively poor results of the Google search engines on our query collection, we can conclude that it is not tailored for the task of diagnosing.

4.2.2.2 Difficult cases query collection

A controversial study of the usage of Google Search as an aid in diagnosing difficult cases concluded that for the *difficult cases query collection* in 58% of the cases the correct diagnosis was found using Google Search [18]. However,

	Rare	RareGenet	Google Search
Total number of cases	26	26	26
Correct diagnosis in top 10	9 (34.61%)	13 (50%)	11 (42.30%)
Correct diagnosis in top 11-20	1 (3.84%)	0 (0%)	2 (7.69%)
Correct diagnosis not found	6 (61.54%)	13 (50%)	13 (50%)
Mean reciprocal rank (MRR)	0.158	0.186	0.380
Average precision rank 10 (P@10)	0.054	0.073	0.123
Average precision rank 20 (P@20)	0.042	0.044	0.106
NDCG@10	0.358	0.279	0.391
NDCG@20	0.390	0.325	0.506

Table 4.7: **Effectiveness of Google on the *difficult cases query collection***. Performance comparison between the vertical search engine and Google.

the study is hard to reproduce as not all evaluation settings were given. Nevertheless, we decided to replicate their study with our own methodology as described in Section 3.8. The evaluation results for the general *Google Search* on the *difficult cases query collection* are summarized and compared with retrieval from *Rare* and *RareGenet* in Table 4.7.

In terms of the percentage of queries for which the correct diagnosis was found, both Google Search and our vertical search engine using the *RareGenet* index succeeded in finding the correct disease in 50% of the cases. However, the MRR score of the Google search engine is considerably better, 0.380 compared to 0.186, which could be expected since this query collection was optimized for web search, as discusses in Section 4.2.1.2.

If we eliminate the four cases of diseases that are not classified as rare, the results are as follows: Google Search finds the correct diagnosis in 54.54% cases (12 of 22), while on our search engine using the *RareGenet* index, the correct diagnosis is found in 59.09% cases (13 of 22), with MRR values of 0.426 for Google Search and 0.219 for *RareGenet*.

4.2.3 PubMed

Besides Google Search, PubMed is also popular with clinicians [7, 9, 53]. Thus, we evaluated its effectiveness in retrieving relevant articles for the two query collections, and compared the results with those obtained by the developed vertical search engine.

4.2.3.1 Rare diseases query collection

Searching on PubMed for correct diagnoses on the *rare diseases query collection* returned no results, with the exception of three query searches that retrieved only the original articles from which the queries were extracted.

As the searches were performed using the queries as listed in Appendix A, that is, comma-separated patient symptoms, the parser used by PubMed automatically replaces the commas with AND operators, which forces the matching algorithm to return only those articles containing all symptoms. This is not necessarily desired, especially in the case where there is a long list of symptoms that a patient presents, as is the case with the set of queries contained in the *rare diseases query collection*.

An alternative would be to replace commas with OR operators, but this results in matching the query with all documents that contain at least one of the terms in the query's OR symptoms list. Moreover, these documents are returned based on article meta-information, such as publishing date (default choice), and cannot be ranked based on, for example, the number of symptoms that are covered in a document. Even when using a research methodology filter such as the Diagnosis category filter which limits results to the specific area of clinical diagnosis[3], the number of results is unmanageable in clinical setting (tens of thousands of results).

Another alternative is to make use of the three boolean operators (AND, OR, NOT) provided by PubMed, and combine these with fields designations, such as "MeSH terms", or parentheses, in order to reformulate the query based on the relevance of some symptoms or query terms. However, this is rarely used by clinicians searching PubMed [9, 54]. Another issue of this advanced query formulation task is that it is necessary to have an understanding of the significance of patient symptoms to be able to reduce the number of query terms, which in the case of rare diseases might be difficult as many have common, non-specific symptoms.

The queries from the *rare diseases query collection* include exhaustive lists of patient information, and PubMed does not perform well on them. Excluding or reformulating some of the symptoms in order to shorten the search query and better summarize the case might be an alternative. For example, by summarizing query 15-1-1, "teenager, girl, hypotonia, dehydration, acidosis, massive ketonuria, hyperammonemia" which initially returned no results, to "acidosis, massive ketonuria, hyperammonemia" concludes in finding two articles, the first of which describes the correct diagnosis.

4.2.3.2 Difficult cases query collection

There is a major difference between the performance of PubMed on the two query collections. While for the *rare diseases query collection* it does not return any relevant result, on the *difficult cases query collection* PubMed finds the correct diagnosis for 34.62% of the queries (9 out of 26). We also repeated the search for the synopses of the difficult cases in this collection (i.e. searched on the *difficult queries synopses collection*) which concluded

[3]PubMed Clinical Queries, `http://www.ncbi.nlm.nih.gov/sites/pubmedutils/clinical`

	Rare	RareGenet	Google Search	PubMed
No. of cases	26	26	26	26
Corr. diag. top 10	9 (34.61%)	13 (50%)	11 (42.30%)	7 (26.92%)
Corr. diag. top 20	1 (3.84%)	0 (0%)	2 (7.69%)	2 (7.69%)
Corr. diag. NF	6 (61.54%)	13 (50%)	13 (50%)	17 (65.38%)
MRR	0.158	0.186	0.380	0.276
P@10	0.054	0.073	0.123	0.046
P@20	0.042	0.044	0.106	0.035
NDCG@10	0.358	0.279	0.391	0.265
NDCG@20	0.390	0.325	0.506	0.290

Table 4.8: **Effectiveness of PubMed on the *difficult cases query collection***. Performance comparison between the vertical search engine, Google and PubMed.

in finding the original articles from which the queries were extracted in four cases, and no other documents being retrieved. This is similar with the PubMed results for the *rare diseases query collection*.

While PubMed's MRR is higher than the MRR of the vertical search engine, the other effectiveness values are smaller. The MRR value of PubMed is in such contrast to the other metrics because 7 of the 9 queries for which PubMed finds the correct disease have the first relevant result at rank 1, the remaining 2 results having ranks 11 and 12. By comparison, our system returns the relevant results at rank 1 in only 3 cases, and the MRR score penalizes results where the relevant documents have lower ranks. As such, although all other metrics are better for our system, PubMed scores better in terms of MRR.

If the four queries for which the correct diagnoses were not found in the Orphanet database are excluded from the query collection, PubMed finds the correct diagnoses in 36.36% of the cases (8 out of 22), while on the *RareGenet* index, the developed vertical search engine finds the correct diagnosis in 59.09% of the cases (13 of 22). The MRR for PubMed would become 0.280, while for *RareGenet* it would be 0.219.

4.3 Search Time

The objective of the following experiment was to asses whether the developed vertical search engine is more effective in terms of time spent for answering diagnostic questions than the other systems used by clinicians.

This was tested on a subset of the two query collections. For the *rare diseases query collection*, only those queries that resulted in the correct diagnosis being found by both Google Search and the vertical search engine using the *RareGenet* index were selected (PubMed did not solve any of the

cases from this collection). For the *difficult cases query collection*, the selected subset consisted of those cases solved by our system, Google Search, and PubMed. In total, this experiment was performed on 11 cases (5 from the *rare diseases query collection*, and 6 from the *difficult cases query collection*).

Ideally, this experiment would have involved clinicians using each of the systems to find the correct diagnosis. However, as this was not possible, we used the methodology described in Section 3.8.2.

The measures used in the experiment consist of the number of words to be read and the number of clicks needed to arrive at the correct diagnosis. For the subset from the *rare diseases query collection*, using our system a user would have to read on average 9.6 words before the correct diagnosis is mentioned, while with Google Search, the correct diagnosis would be read after an average of 282.6 words (Appendix C). For this subset of queries, no clicks were required from either systems, as the correct diseases were mentioned in the results page. For the subset from the *difficult cases query collection*, a user would read on average 13.2 words using our system until the correct diagnosis is mentioned, no additional clicks being required. For the same query subset, using Google would require an average of 22.7 words being read and 2 clicks being pressed, with an additional average of 163 words to be read after the clicks, before reading the correct diagnosis. In the case of PubMed, an average of 49.2 words would have to be read, 1 click pressed, and 131 words read after the click for the correct diagnosis to be mentioned (Table C.9).

This experiment shows that a clinician using our system would spot the correct disease faster than by using the other two systems. Combining this with the fact that our system effectiveness is overall better than the other systems, we could argue that using our system decreases the search time spent by clinicians looking for rare disease diagnostic hypotheses.

4.4 Failure Analysis

Experimental evaluation on the *rare diseases query collection* shows that the vertical search engine using the *RareGenet* index can find the correct diagnosis for 23 of 30 rare disease cases (76.67%). Still, for seven cases of the *rare disease query collection* (see Table 4.9), the correct diagnosis is not found using this system, although documents on the diseases exist in the index. In what follows, we intend to analyse the reasons for which our system failed in these cases.

After the queries extraction process, the search queries were validated by a clinician[4]. For cases 7-1-1, 12-1-1, and 13-1-1 we have received a few comments from him regarding the difficulty of correctly diagnosing these

[4]Henrik L. Jørgensen, chief physician at Bispebjerg Hospital

Query ID	Final Diagnosis	Query
1-1-1	Rothmund-Thomson syndrome	6 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
7-1-1	Congenital hepatic fibrosis	10 year old, girl, thrombocytopenia, splenomegaly, headache, itching rubeoliform rash
9-1-1	Type I tyrosinemia	4 month old, boy, epistaxis, haematemesis, haematochezia, subconjunctival bleeding, petechiae, haematomas, haemangioma, slightly enlarged liver, elevated serum transaminases
12-1-1	Whipple's disease	64 year old, male, inflammatory back pain, flares of arthritis, multisegmental spondylitis
13-1-1	Dengue hemorrhagic fever	70 year old, male, massive hemoptysis, respiratory distress, anemia, hemodynamic instability, renal failure, intense headache, arthralgia, myalgias, ecchymoses over arms and abdomen, acidosis, pleural effusions, blood tinged secretion from lungs
18-1-1	LIG4 syndrome	girl, pronounced microcephaly, short stature, psychomotoric delay, distinctive facial appearance, thrombocytopenia, anemia, leukocytopenia, pancytopenia, growth retardation, telecanthus, epicanthal folds, ptosis, infections of the inner ear and respiratory tract, hypoplastic marrow with cellular dysplasia
20-1-1	Terminal deletion of chromosome 4q	21 year old, female, irregular menses, menorrhagia, hand and foot malformation, ovarian cyst, basic cognitive function

Table 4.9: **The seven cases from the *rare diseases query collection*** for which the correct diagnosis was not found by the vertical search engines using the *RareGenet* index, although documents on these diseases exist in the index.

cases. The comments were the following: for case 7-1-1 "these symptoms could be caused by many different diseases, including some fairly common ones", for case 12-1-1 "in a patient of 64 years, these symptoms could be caused by a multitude of diseases, most of them much more common than the rare infectious disease", and for case 13-1-1 "interesting, although not that uncommon; several other similar infections could produce a picture like this". The common feature describing these cases is that their presentation is very likely to fit a multitude of other diagnostic hypotheses much more probable to occur than the correct disease. Thus, they are even more difficult to diagnose and more likely to result in misdiagnoses or diagnosis delays caused by numerous laboratory tests and therapeutic trials. It is worth mentioning that neither Google Search nor any of the Google customized search engines found the correct diagnoses for the seven cases where our system failed.

For the *difficult cases query collection*, the system failed to retrieve the correct diagnosis for 13 of the queries (50%) using the *RareGenet* index (Table 4.10). Four of these thirteen cases are not listed as rare diseases by the Orphanet rare disease database (4 of 13, 30.76%). As a result, three of these diseases are not even part of our index, which is to be expected, as our index is focused on the topic of rare diseases.

From the remaining nine cases, one of the queries pertains to a patient simultaneously suffering from two diseases. This type of cases are obviously harder to elucidate. In our evaluation methodology, a document must cover both diseases in order to be considered relevant for such a case. This is probably a flaw since the patient could benefit from any of the diseases

BMJ Case	BMJ Synopsys	BMJ Google Search Terms	BMJ Final Diagnosis	Is rare	In Rare Genet	Ret. by Google	Ret. by PubMed
5	53 yo man with depression, Aortic regurg, heart block and acute puloedema.	Acute Aortic regurgitation, depression, abscess	Infective endocarditis	Yes	Yes	No	Yes
6	58 yo newly diagnosed oesophageal cancer, refractory hic cups and vomiting	oesophageal cancer, refractory hic cups, nausea, vomiting	Linitis plastica with bowel obstruction	Yes	Yes	No	No
8	10 yo boy with right thigh pain and CT showed lytic R hip lesion	hip lesion, older child	Osteoid osteoma	No	Yes	No	No
9	67 yo man with acute respiratory failure, exposure to bird dropping	HRCT centrilobular nodules, acute respiratory failure	Hot tub lung secondary to M avium	Yes	Yes	No	No
10	73 yo fever, thigh pain, urinary frequency, previous statin use	fever, bilateral thigh pain, weakness	Ehrlichiosis	Yes	Yes	No	No
14	38 yo man with ulcerative colitis, fever, blurred vision and dyspnoea	ulcerative colitis, blurred vision, fever	Vasculitis	Yes	Yes	No	No
15	80 yo man with dyspnoea and proteinuria	nephrotic syndrome, Bence Jones, ventricular failure	Amyloid light chain	Yes	Yes	Yes	No
16	9 yo female with headache, hypertension, visual disturbance	hypertension, papilledema, headache, renal mass, cafe au lait	Pheochromocytoma	Yes	Yes	Yes	No
17	22 yo female with back pain, pulmonary infiltrates, rapidly progressing to death	sickle cell, pulmonary infiltrates, back pain	Acute chest syndrome	No	No	Yes	Yes
18	45 yo female with painful abdo mass	fibroma, astrocytoma, tumor, leiomyoma, scoliosis	Endometriosis	Yes	Yes	No	No
19	17 yo female Tsunami survivor with respiratory distress and R hemiplegia	pulmonary infiltrates, cns lesion	Aspiration pneumonia and brain abscess (polymicrobial)	No	No	No	No
25	40 yo with wt loss, sweats and persistent fever after food poisoning.	portal vein thrombosis, cancer	Pylephlebitis	No	No	No	No
31	60 yo man with buttock purpuric rash, chronic renal failure.	buttock rash, renal failure, edema	Cryoglobulinaemia	Yes	Yes	No	No

Table 4.10: **The 13 cases from the *difficult cases query collection*** for which the correct diagnosis was not found by the vertical search engine using the *RareGenet* index.

being identified and managed.

From the 13 cases for which our system failed to find the correct diagnosis, Google Search succeeded in finding three of them (23.07%). Of these three, the correct diagnosis for one case is not listed as a rare disease in Orphanet, and is not indexed in *RareGenet* (Table 4.10). The relevant results returned by Google for the remaining two articles are mostly published case reports. This suggests that we might improve the coverage of the system by including additional medical case reports from, for example, PubMed Central Open Access Subset.

PubMed succeeded in finding the correct diagnosis in a total of 9 cases. Out of these, two were not found by our system. One was not indexed by our system and it was not listed as a rare disease in Orphanet, and the other was only retrieved by PubMed, as neither our system nor Google managed to retrieve relevant articles for the case.

While for the *rare diseases query collection*, which consists of long search queries (22.17 terms on average), our system obviously outperformed Google, on the queries from the *difficult cases query collection* (with 5 terms on average), the two systems perform similarly. We analyse the reasons why this happens in the following chapter.

Chapter 5

Discussion

5.1 Summary of the Experimental Evaluation

Effectiveness improvements over other systems

RQ1 *Does the experimental evaluation of our system show substantial improvements over other systems in terms of document relevance?*

The experimental evaluation of the vertical search engine, Google Search, two Google custom searches, and PubMed shows that for most of the measurements, the developed vertical search engine performs better, or at least similar to the other systems. From the range of experiments performed, the closest match to the overall effectiveness of our system is the Google Search engine's performance on the *difficult cases query collection*, where both systems find the correct diagnosis in 50% of the cases. On all other effectiveness experiments, our system consistently delivers better results.

The failure to perform better than Google on the *difficult cases query collection* is probably a result of the query collection's low average term count, which means that only what are considered to be the most important patient features are included in the query. It could be argued that searching with a short query is a familiar search strategy for clinicians, but on the other hand, at the time when diagnostic decisions are made, the clinician has access to a variety of patient data, including history and test results. Moreover, at this step, it is important to generate new hypothesis ideas as opposed to forcing the clinician to select what patient information is the most relevant for a diagnosis.

The effectiveness scores combined with a user interface optimised for the task of diagnosing rare diseases could translate into an improved diagnostic process by shortening search times and presenting more relevant diagnostic hypotheses.

Index coverage affecting the system's effectiveness

RQ2 *Does the inclusion of a larger pool of articles on the topic of genetic diseases improve the effectiveness of the system in diagnosing rare diseases?*

From the experiments made on the developed vertical search engine, we have identified that including in the system's index articles on both rare and genetic diseases results in better effectiveness scores than retrieving from a smaller index containing mostly rare disease articles. This observation can be explained by the fact that, by including genetic disease articles, many of which are also rare, the disease coverage increases.

On the *rare diseases query collection*, retrieval from the *RareGenet* index results in finding the correct diagnosis in 76.67% of the cases, while retrieval from the *Rare* index results in finding the correct diagnosis in 66.67% of the cases. On the *difficult cases query collection*, retrieval from the *RareGenet* index results in finding the correct diagnosis in 50% of the cases, and retrieval from the *Rare* index results in finding the correct diagnosis in 38.46% of the cases. No major differences in MRR or NDCG are observed.

Although the index size increased by 207.8% with the inclusion of genetic articles, efficiency measures results are not deteriorated and score below the efficiency limit of 0.5 seconds for a response, which is perceived as an instantaneous reply [45].

Therefore, the inclusion of genetic disease articles improves the overall effectiveness of the system for both query collections without a major speed impact. In the future, it would be interesting to see if similar improvements can be achieved with an even wider collection of articles.

Using prior probabilities to increase the relevance of rare disease articles

RQ3 *Does increasing the prior probabilities of the relevance of rare disease articles in contrast to the relevance of genetic disease articles improve the effectiveness of the system in diagnosing rare diseases?*

With the inclusion of genetic disease articles into the index, many of the documents returned for the tested queries belonged to the resources on genetic diseases. As a consequence, some articles that were relevant using the *Rare* index were not retrieved in top 20 any more. In order for those articles to be ranked higher, we decided to increase the prior probabilities of the relevance of all rare disease articles.

We measured the effectiveness for retrieval from the *RareGenet* index on the *rare diseases query collection* using boosting factors ϕ of 2 and 4 for the rare disease articles relevance, and compared these values with retrieval without using prior probabilities. MRR, average precision and NDCG scores

showed a slight improvement, although the number of cases for which the correct diagnosis was found in the results remained the same.

When the experiments were repeated for two additional smoothing values (800 and 4000), the best overall effectiveness was observed on the *RareGenet* index with a boosting factor $\phi = 4$ and a smoothing value $\mu = 4000$. However, due to the fact that we manually evaluated the effectiveness, we did not perform the evaluation on a range of μ and ϕ values. As a result, we cannot assess how these values correlate with the effectiveness measurements.

Reducing the amount of time spent searching for diagnostic hypotheses

RQ4 *Does the use of our system, in comparison with other systems, decrease the search time spent by clinicians looking for rare disease diagnostic hypotheses?*

Our experiments show that if a clinician is to read the results of a query from start to finish, and then sequentially go through the linked articles if the correct disease is not found in the results page, using our system would be faster than using Google or PubMed.

These results are to be expected, as the system is optimised for the task of generating diagnostic hypotheses. It should be mentioned that most of the titles for the documents indexed by our system are those of the disease they cover. As a result, almost always, if the correct diagnosis is retrieved, the name of the correct disease or one of its synonyms appear directly in the results page, without further clicks being necessary.

Although the time it takes to arrive at the first mention of the correct diagnosis is better for our system, it is not clear if this would necessarily translate into clinicians generating diagnostic hypotheses faster. The only way to assess this with certainty would be by observing the clinicians using the systems themselves.

5.2 Limitations and Directions of Future Work

The most important deficiency in evaluating the vertical search engine was the fact that the authors themselves performed all the experiments and measurements. Moreover, only previously reported patient cases were used for the query collections. In the future, it is crucial that a tighter collaboration with medical personnel is established in order to gather original rare disease cases, have the clinicians perform relevance assessments, and understand their searching behaviour. Even if the overall performance of the system is better than that of other systems used by clinicians, there are numerous directions for future work that could be explored.

5.2.1 Evaluation strategy

The tests and experimental evaluations were performed by two non-medical experts not blinded to the correct diagnoses. While these experimental laboratory tests show satisfactory results, further tests are needed before the search engine can be used in clinical practice.

It would be interesting to further investigate the performance of the system by designing a randomized controlled laboratory trial, where expert evaluators (medical students, clinicians, general practitioners, or rare disease experts) are blinded to the correct diagnosis. Such an investigation could establish the effectiveness of the system when used by medical specialists. Moreover, a field test could be conducted at a rare disease centre, to provide an insight into what are the specific clinical needs, how to best integrate the system into the clinical workflow, and test if such a system could improve the diagnostic process.

The two query collections used to evaluate the system consist of 30, respectively 26, queries representing rare disease and difficult cases. A larger query collection comprised of novel rare disease cases should be constructed. However, it is difficult to collect original descriptions of rare disease cases. One way to collect such information is through clinicians that have encountered rare disease cases, but care must be taken to remove personal patient information first.

One of the flaws of using previously used query collections or creating collections from previously reported cases is that Google, as well as PubMed, usually index the original articles discussing the cases. This issue has been solved in our evaluations by merely eliminating the original articles, but this approach could introduce some bias.

The query collections used in the evaluations do not have associated relevance judgements. As we compare the developed system with web search engines where it is difficult to identify all relevant documents for a given query, we did not add relevance judgements. However, if the focus is more on testing the performance of the system alone, relevance judgements could be added, and thus the evaluation process could be improved. By automating the evaluation, a larger number of tests and experiments could be performed. For example, it would be possible to find the best value for the μ smoothing parameter for a given index. Moreover, because the number of relevant documents for a query would be known in advance, the system's recall could also be measured.

As part of our efforts to understand the needs of clinicians, we had a meeting with two experts on rare and genetic diseases from Rigshospitalet. Even if we received positive feedback on a demonstration of the system, we strongly believe that a better understanding of the clinical needs is crucial for having such a system accepted by the medical professionals.

5.2.1.1 Ranking diseases instead of documents

The disease ranking experimental version of the vertical search engine presents to the clinician a list of ranked diseases extracted from the first 20 documents returned by the document ranking algorithm. This approach could further decrease the search time for diagnostic hypotheses, since instead of documents, every disease could have associated a short summary containing the disease description, symptoms, disease confirmation tests, genetic causes, differential diagnoses, or other information aggregated from medical resources.

In our tests, disease ranking sometimes manages to extract the correct disease for more cases. For example, for the *difficult cases query collection*, it finds four additional correct diagnoses for the *Rare* index, and three for the *RareGenet* index (Appendix C). However, there are now on average 52 results for retrieval from the *Rare* index, and 341 for the *RareGenet* index, and the MRR values decrease considerably.

Although for now disease ranking does not seem to be better at generating diagnostic hypotheses, starting from the idea of returning diseases as results, it would be probably interesting to further develop a visual graph of the results and their interconnections. Such a visualization could prove useful to see, for example, if the results come from a certain class of diseases, possibly streamlining the diagnostic process.

5.2.1.2 Search time

The most important issue with our measurement of the time it would take a clinician to arrive at the correct diagnosis is that we cannot observe the clinicians using the evaluated systems. For example, our naïve approach of considering that the clinician sequentially reads and clicks through the results, is certainly far from real practice.

If a field study is to be conducted, a logging and timing of the clinicians' actions would be a more appropriate and precise measurement. Additionally, a qualitative test could point into the areas that could be improved to decrease the time it takes clinicians to generate hypothesis ideas using the vertical search engine.

5.2.2 System design

In our system, we have used the textual information from a variety of rare and genetic disease resources. However, as hinted in Figure 2.2, many of these resources provide additional information such as references to entries in medical classifications, databases, or ontologies. Moreover, most of the resources have a well-defined structure which means that disease synonyms, symptoms, evolution, or treatment are usually separated into different sections. Although we have not used the structure or references to other med-

Translations:	
21 year old	"adult"[MeSH Terms] OR "adult"[All Fields] OR "21 year old"[All Fields]
female	"women"[MeSH Terms] OR "women"[All Fields] OR "female"[All Fields] OR "female"[MeSH Terms]
foot malformation	"foot deformities, congenital"[MeSH Terms] OR ("foot"[All Fields] AND "deformities"[All Fields] AND "congenital" [All Fields]) OR "congenital foot deformities"[All Fields] OR ("foot"[All Fields] AND "malformation"[All Fields]) OR "foot malformation"[All Fields]
function	"physiology"[Subheading] OR "physiology"[All Fields] OR "function"[All Fields] OR "physiology"[MeSH Terms] OR "function"[All Fields]
hand	"hand"[MeSH Terms] OR "hand"[All Fields]
menorrhagia	"menorrhagia"[MeSH Terms] OR "menorrhagia"[All Fields]
menses	"menstruation"[MeSH Terms] OR "menstruation"[All Fields] OR "menses"[All Fields]
ovarian cyst	"ovarian cysts"[MeSH Terms] OR ("ovarian"[All Fields] AND "cysts"[All Fields]) OR "ovarian cysts"[All Fields] OR ("ovarian"[All Fields] AND "cyst"[All Fields]) OR "ovarian cyst"[All Fields]

Figure 5.1: **PubMed translation into MeSH terms** for the query "21 year old, female, irregular menses, menorrhagia, hand and foot malformation, ovarian cyst, basic cognitive function".

ical resources present in the documents indexed in our system, it is obvious that this information could be used to improve the system. One possible scenario would be the use of this information when displaying the results for the most relevant documents or diseases. Another possibility would be to try to improve the disease ranking algorithm by using the references to other medical resources to compute similarity measures between diseases. Additionally, the structure of the documents and natural language processing (NLP) techniques could be used to identify a timeline of the disease's evolution or concepts associated to the disease, such as risk population, age of onset, inheritance, or differential diagnosis.

Information extraction could be used to identify in the documents the medical terms that appear in medical vocabularies such as those from the UMLS Metathesaurus. This information could be used to filter the results according to symptoms, identify known disease names, recognize temporal expressions (e.g. previous symptoms or conditions for a patient), detect coreference (e.g. linking names that refer to the same disease) and identify relations between entities (e.g. a disease considered in the differential diagnosis of another disease). For annotating the medical text, the following classifications and ontologies could be used: UMLS Metathesaurus, HPO, OMIM, ICD, Orphanet classification of rare diseases, Gene Ontology (GO), MeSH headings, or the London Dysmorphology Database.

The same information extraction techniques can be applied on the user queries, by matching the query terms with terminology from the medical vocabularies and ontologies. For example, queries submitted to PubMed are automatically translated into MeSH terms, as shown in Figure 5.1. Although most of the queries we use in our evaluation include the age of the patient, this information is not used in the vertical search engine. However, by using the MeSH headings, an expression such as "21 year old" is asso-

ciated with the term "adult", which is added to the list of synonyms to be used in the search. Besides enriching the query with synonyms, the medical vocabularies could be used to provide spell checking functionality, and the Metathesaurus in particular could help translate specialised medical terms given in a different language.

Because the sources of information indexed by the vertical search engine are heterogeneous in quality and style, it might be useful for the clinicians to have the resources classified based on the type of language in which they are written, either using expert medical terms or basic English [55]. In this way, patient-oriented articles could be distinguished from articles addressed to medical specialists.

Another possibility would be to correlate the retrieved results with actions that could be taken by the clinician in order to confirm or eliminate hypotheses. For example, disease-associated tests for confirming the presence of a disease in the patient could be presented, as well as disease inheritance information that might guide the clinician in verifying the patient's medical history. Photographic images associated to those rare diseases that present known dysmorphic features could also be integrated as additional information for the retrieved results.

Besides rare diseases, there are also other situations in which elucidating the diagnosis could be difficult, such as patients simultaneously suffering from multiple diseases, or exhibiting atypical or non-specific presentations [27]. It could be interesting to extend the collection of articles indexed by the system and evaluate it on these types of difficult cases as well.

Chapter 6

Conclusions

As a consequence of the low prevalence and unspecific symptoms of rare diseases, patient suffering from a rare disease are faced with high misdiagnosis rates and long diagnostic delays. Therefore, the area of rare disease diagnosis is one where information retrieval could improve practice, by making use of the domain-specific knowledge from medical databases and the relations between these. As part of this study, a functional rare disease vertical search engine was developed with the goal of retrieving relevant documents given textual patient descriptions.

The goal of the thesis was to create a freely available vertical search engine dedicated to rare diseases that could be used by clinicians in their differential diagnostic process. A document collection was extracted from publicly available rare and genetic disease articles, on which state-of-the-art IR techniques were applied. The user interface, inspired by the simplicity of web search engines, allows clinicians to introduce any patient data as free text and have a quick overview of the results, together with the option of viewing more details for selected documents.

An important part of the study was to evaluate the effectiveness of the developed vertical search engine as compared to other systems currently used by clinicians when faced with diagnosing difficulties. Although no field study was performed, our experiments show that the developed vertical search engine has better overall performance than other systems.

Nevertheless, in order to establish if using such a system improves the clinical practice, and in order to further investigate the clinical workflow and understand how clinicians use the system, laboratory control trials and field test will have to be conducted.

Acknowledgements

We are thankful to our supervisors Ole Winther and Christina Lioma for their interest in the project and their availability for questions and discussions. Their help and guidance was invaluable for the completion of the project. We are also grateful to Birger Larsen, associate professor at the Royal School of Library and Information Science, for his input on the project.

We would also like to thank chief physician Henrik L. Jørgensen for his input on the devised query collection, and expert physicians Allan M. Lund from Klinik for Sjældne Handicap, Klinisk Genetisk Afdeling, Rigshospitalet, and Malene B. Rasmussen from the Genetisk Rådgivning department of the same clinic for their feedback on the developed system and their proposals for future development.

We are grateful to our colleagues Henrik G. Jensen and Michael Andersen who, through their bachelor thesis, inspired us to pursue the problems tackled in this work.

We recognize the importance of the work done by medical specialists who worked on the medical resources used in this project. We are especially grateful to Orphanet for allowing us to use their rare disease resources in our project.

Authors contribution

The authors (RD and PP) equally contributed and closely collaborated at the design and development of the vertical search engine, the experimental evaluations, and in writing this report.

Most of the code was written using the pair programming technique. The article links from the rare and genetic diseases web databases were scraped by PP and fetched by RD. PP signed the license for the data acquisition of the Orphanet databases, while RD signed for access to the OMIM entries, as well as for access to other NLM databases. PP was mainly responsible for transforming the scrapped webpages into TREC-formatted files. The web API was developed by RD, while both authors contributed to the web UI.

The queries from the *OJRD query collection* were extracted jointly by the authors after reading the case reports. The effectiveness evaluations were performed jointly by the authors. Pages of results from Google and PubMed where saved simultaneously, and then evaluated. In principle, one of the authors screened the results (RD), while the other validated the findings and wrote the relevant results in a spreadsheet (PP), as well as noted the observations made by both authors. The evaluation methodology was devised by both authors. PP customized the Google custom search engines. RD administered and maintained the servers on which the IR system runs.

The authors jointly wrote the manuscript. PP was mostly responsible for drafting the report and including the tables, figures, and appendices. RD was mostly responsible with reviewing the content and discussing the experimental evaluation results. Both authors discussed, analysed and interpreted the results and commented on the work for this report.

Bibliography

[1] D. F. Sittig, M. A. Krall, R. H. Dykstra, A. Russell, and H. L. Chin, "A survey of factors affecting clinician acceptance of clinical decision support," *BMC Medical Informatics and Decision Making*, vol. 6, p. 6, Jan. 2006.

[2] J. Wyatt, "Computer-based knowledge systems," *The Lancet*, vol. 338, pp. 1431–1436, 1991.

[3] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, and D. W. Bates, "Grand challenges in clinical decision support," *Journal of Biomedical Informatics*, vol. 41, pp. 387–92, Apr. 2008.

[4] B. C. Delaney, D. A. Fitzmaurice, A. Riaz, and F. D. Hobbs, "Can computerised decision support systems deliver improved quality in primary care?," *BMJ Clinical Research Ed.*, vol. 319, p. 1281, Nov. 1999.

[5] D. W. Bates, G. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, C. Spurr, R. Khorasani, M. Tanasijevic, and B. Middleton, "Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality," *Journal of the American Medical Informatics Association*, vol. 10, no. 6, pp. 523–530, 2003.

[6] C. Lombardi, E. Griffiths, B. McLeod, A. Caviglia, and M. Penagos, "Search engine as a diagnostic tool in difficult immunological and allergologic cases: is Google useful?," *Internal Medicine Journal*, vol. 39, no. 7, pp. 459–464, 2009.

[7] T. Kortteisto, M. Kaila, I. Kunnamo, P. Nyberg, A.-M. Aalto, and P. Rissanen, "Self-reported use and clinical usefulness of second-generation decision support - a survey at the pilot sites for evidence-based medicine electronic decision support," *Finnish Journal of eHealth and eWelfare*, vol. 1, no. 3, pp. 161–169, 2009.

[8] M. Sim, E. Khong, and M. Jiwa, "Does general practice Google?," *Australian Family Physician*, vol. 37, no. 6, pp. 471–474, 2008.

[9] P. N. Hider, G. Griffin, M. Walker, and E. Coughlan, "The information-seeking behavior of clinical staff in a large health care organization," *Journal of the Medical Library Association*, vol. 97, pp. 47–50, Jan. 2009.

[10] M. G. Bouwman, Q. G. A. Teunissen, F. A. Wijburg, and G. E. Linthorst, "'Doctor Google' ending the diagnostic odyssey in lysosomal storage disorders: parents using internet search engines as an efficient diagnostic strategy in rare diseases," *Archives of Disease in Childhood*, vol. 95, pp. 642–4, Aug. 2010.

[11] EURORDIS, "Eurordiscare2: survey of diagnostic delays, 8 diseases," 2004.

[12] "Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products," 2000.

[13] EURORDIS, "Rare diseases: understanding this public health priority, founding paper," November 2005.

[14] S. Ayme, "Orphanet, an information site on rare diseases," *Soins*, no. 672, p. 46, 2003.

[15] B. Go, J. J. Cimino, J. Hupp, and E. P. Hoffer, "DXplain - an evolving diagnostic decision-support system," *Journal of the American Medical Association*, vol. 258, pp. 67–74, 1987.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the web. Technical Report," *Computer Science Department Stanford University*, 1999.

[17] R. Dragusin and P. Petcu, "Improving clinical practice using a computerized clinical decision support system for diagnosing rare diseases: literature review, challenges, and possible paths forward. Technical Report," *Institute of Computer Science at Copenhagen University*, 2010.

[18] H. Tang and J. H. K. Ng, "Googling for a diagnosis - use of Google as a diagnostic aid: internet based study," *BMJ Clinical Research Ed.*, vol. 333, pp. 1143–5, Dec. 2006.

[19] R. A. Miller, "Medical diagnostic decision support systems - past, present, and future," *Journal of the American Medical Informatics*, vol. 1, no. 1, p. 8, 1994.

[20] D. Crombie, "Diagnostic process," *Journal of the College of General Practitioners*, vol. 6, no. 4, p. 579, 1963.

[21] R. A. Miller, "Computer-assisted diagnostic decision support: history, challenges, and possible paths forward," *Advances in Health Sciences Education*, vol. 14 Suppl 1, pp. 89–106, Sept. 2009.

[22] J. Wyatt, "Use and sources of medical knowledge," *The Lancet*, vol. 338, no. 8779, pp. 1368–1373, 1991.

[23] E. Campbell, "The diagnosing mind," *The Lancet*, vol. 329, no. 8537, pp. 849–851, 1987.

[24] O. Kostopoulou, J. Oudhoff, R. Nath, B. Delaney, C. Munro, C. Harries, and R. Holder, "Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine," *Medical Decision Making*, vol. 28, no. 5, p. 668, 2008.

[25] E. Antman, J. Lau, B. Kupelnick, F. Mosteller, and T. C. Chalmers, "A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction," *Journal of the Americal Medical Association*, vol. 268, no. 2, p. 240, 1992.

[26] B. C. Delaney, "Potential for improving patient safety by computerized decision support systems," *Family Practice*, vol. 25, no. 3, pp. 137–8, 2008.

[27] O. Kostopoulou, B. Delaney, and C. Munro, "Diagnostic difficulty and error in primary carea systematic review," *Family practice*, vol. 25, no. 6, p. 400, 2008.

[28] P. Ramnarayan and J. Britto, "Paediatric clinical decision support systems," *Archives of Disease in Childhood*, vol. 87, p. 361, Nov. 2002.

[29] C. Sneiderman, D. Demner-Fushman, M. Fiszman, N. Ide, and T. C. Rindflesch, "Knowledge-based methods to help clinicians find answers in MEDLINE," *Journal of the American Medical Informatics Association*, pp. 772–780, 2007.

[30] J. J. Cimino, J. Li, S. Bakken, and V. L. Patel, "Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users," *Proceedings of the AMIA Symposium*, p. 170, Jan. 2002.

[31] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success," *BMJ Clinical Research Ed.*, vol. 330, p. 765, Apr. 2005.

[32] I. Sim, P. Gorman, R. A. Greenes, R. Haynes, B. Kaplan, H. Lehmann, and P. C. Tang, "Clinical decision support systems for the practice of evidence-based medicine," *Journal of the American Medical Informatics Association*, vol. 8, no. 6, pp. 527–34, 2001.

[33] D. C. Classen, "Clinical decision support systems to improve clinical practice and quality of care," *Journal of the American Medical Association*, vol. 280, p. 1360, Oct. 1998.

[34] B. Kaplan, "Evaluating informatics applications – clinical decision support systems literature review," *International Journal of Medical Informatics*, vol. 64, pp. 15–37, Nov. 2001.

[35] E. Berner and T. J. L. Lande, "Overview of clinical decision support systems," *Decision Support Systems, Healthcare Information Management Systems, Third Edition, Chapter 36*, vol. 6, pp. 463–477.

[36] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?," *Journal of Biomedical Informatics*, vol. 42, pp. 760–772, Oct. 2009.

[37] C. P. Friedman, "Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems," *Journal of the American Medical Association*, vol. 282, pp. 1851–1856, Nov. 1999.

[38] G. Barnett, E. P. Hoffer, M. J. Feldman, K. T. Famiglietti, and R. J. Kim, "DXplain - a niche resource for just-in-time knowledge access," *Proceedings of the AMIA Symposium*, vol. 2005, no. 1, p. 1170, 2005.

[39] S. Köhler, M. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. Robinson, "Clinical diagnostics in human genetics with semantic similarity searches in ontologies," *The American Journal of Human Genetics*, vol. 85, no. 4, pp. 457–464, 2009.

[40] "Online mendelian inheritance in man," *McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University and National Center for Biotechnology Information, National Library of Medicine*, 2011.

[41] R. M. Winter, M. Baraitser, and J. M. Douglas, "A computerised data base for the diagnosis of rare dysmorphic syndromes," *Journal of Medical Genetics*, vol. 21, pp. 121–3, Apr. 1984.

[42] "Possum-web," *Murdoch Childrens Research Institute*, 2011.

[43] G. De Leo, C. LeRouge, C. Ceriani, and F. Niederman, "Websites most frequently used by physician for gathering medical information," *Proceedings of the AMIA Symposium*, p. 902, Jan. 2006.

[44] M. Falagas, F. Ntziora, G. Makris, G. Malietzis, and P. Rafailidis, "Do PubMed and Google searches help medical students and young doctors reach the correct diagnosis? a pilot study," *European Journal of Internal Medicine*, vol. 20, no. 8, pp. 788–790, 2009.

[45] B. Croft, D. Metzler, and T. Strohman, *Search engines: information retrieval in practice*. Addison-Wesley Publishing Company, USA, 2009.

[46] E. Eckard and J. Chappelier, "Free software for research in information retrieval and textual clustering," 2007.

[47] C. Middleton and R. Baeza-Yates, "A comparison of open source search engines. Technical Report," *Universitat Pompeu Fabra, Department of Technologies*, Oct. 2007.

[48] I. Witten, A. Moffat, and T. Bell, *Managing gigabytes: compressing and indexing documents and images*. Academic Press, Morgan Kaufmann, 1999.

[49] T. Strohman, D. Metzler, H. Turtle, and W. Croft, "Indri: a language model-based search engine for complex queries," in *Proceedings of the International Conference on Intelligence Analysis*, 2004.

[50] R. Krovetz, "Viewing morphology as an inference process," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 191–202, ACM, 1993.

[51] A. Hoogendam, A. Stalenhoef, P. Robbé, and A. 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*, vol. 8, p. 42, 2008.

[52] H. G. Jensen and M. Andersen, "Support decision system for diagnosing rare diseases using vector space model and medical text mining. Technical Report," *Institute of Computer Science at Copenhagen University*, January 2010.

[53] R. Thiele, N. Poiro, D. Scalzo, and E. Nemergut, "Speed, accuracy, and confidence in Google, Ovid, PubMed, and UpToDate: results of a randomised trial," *Postgraduate Medical Journal*, vol. 86, no. 1018, p. 459, 2010.

[54] J. Herskovic, L. Tanaka, W. Hersh, and E. Bernstam, "A day in the life of PubMed: analysis of a typical days query log," *Journal of the American Medical Informatics Association*, vol. 14, no. 2, p. 212, 2007.

[55] N. Grabar and S. Krivine, "Application of cross-language criteria for the automatic distinction of expert and non expert online health documents," *Artificial Intelligence in Medicine*, pp. 252–256, 2007.

Appendix A

Query Collections

Query ID	Correct Diagnosis	Query	Source	Synonyms (from Orphanet)
H1-1	Fibrodysplasia ossificans progressiva	Boy, normal birth, deformity of both big toes (missing joint), quick development of bone tumor near spine and osteogenesis at biopsy	Henrik L. Jørgensen	'FOP', 'Man of stone', 'Myositis ossificans progressiva', 'Fibrodysplasia ossificans progressiva'
H2-1	Adrenoleukodystrophy autosomal neonatal form	Normally developed boy age 5, progressive development of talking difficulties, seizures, ataxia, adrenal insufficiency and degeneration of visual and auditory functions	Henrik L. Jørgensen	'ADL', 'Adrenoleukodystrophy, X-linked', 'Adrenoleukodystrophy, neonatal', 'Pseudoadrenoleukodystrophy', 'Acyl-CoA oxidase deficiency', 'Adrenoleukodystrophy, X-linked, cerebral form'
H3-1	Papillon Lefevre syndrome	Boy age 14, yellow, keratotic plaques on the skin of palms and soles going up onto the dorsal side. Both hands and feet are affected	Henrik L. Jørgensen	'Keratosis palmoplantar - periodontopathy'
H4-1	Kleine Levin Syndrome	Jewish boy age 16, monthly seizures, sleep deficiency, aggressive and irritable when woken, highly increased sexual appetite and hunger	Henrik L. Jørgensen	
H5-1	Schinzel-Giedion Syndrome	Male child, malformations at birth, midfacial retraction with a deep groove under the eyes, and hypertelorism, short nose with a low nasal bridge and large lowset ears, wide mouth and retrognathia. Hypertrichosis with bright reddish hair and a median frontal cutaneous angioma, short neck with redundant skin, Bilateral inguinal hernias, hypospadias with a megameatus, and cryptorchidism	Henrik L. Jørgensen	'Midface retraction syndrome, Schinzel-Giedion type'
1-1-1	Rothmund-Thomson syndrome	6 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	http://www.ojrd.com/content/5/1/37	'RTS', 'Poikiloderma of Rothmund-Thomson', 'RTS1', 'Rothmund-Thomson syndrome type 1 or 2', 'Poikiloderma of Rothmund-Thomson type 1 or 2', 'RTS2'
2-1-1	Autosomal recessive centronuclear myopathy (ARCNM)	13 year old, teenage girl, skeletal muscle defects (muscle weakness), mild mental retardation, ophthalmoparesis	http://www.ojrd.com/content/5/1/35	'Autosomal dominant centronuclear myopathy'
2-2-1	Autosomal recessive centronuclear myopathy (ARCNM)	14 year old, teenage boy, mild mental retardation, proximal muscle weakness, unable to walk (wheelchair-bound), premature ventricular complexes, ophthalmoparesis	http://www.ojrd.com/content/5/1/35	'Autosomal dominant centronuclear myopathy'
3-1-1	Cerebrotendinous xanthomatosis (CTX) Synonym: Sterol 27-hydroxylase deficiency	35 year old, female, progressive disturbance of gait (difficulties in walking), recurrent diarrhea, bronchitis, growth retardation, mild retardation of psychomotor development in infancy, bilateral juvenile cataracts, swelling of the Achilles tendons, high arched feet, exaggerated tendon reflexes	http://www.ojrd.com/content/5/1/27	'Xanthomatosis cerebrotendinous', 'Sterol 27-hydroxylase deficiency'
4-1-1	Cogan's syndrome	25 year old, woman, conjunctival hyperaemia, interstitial keratitis, moderate bilateral sensorineural hearing loss, tinnitus, dizziness, nausea and vertigo	http://www.ojrd.com/content/5/1/18	
5-1-1	CDG (Congenital Disorders of Glycosylation) syndrome type Ic Synonyms: Carbohydrate deficient glycoprotein syndrome type Ic, Congenital disorder of glycosylation type 1c (or Ic)	11 year old, boy, severe psychomotor retardation, seizures, strabismus, inverted nipples, dilated cardiomyopathy, hypotonia, wheelchair-bound	http://www.ojrd.com/content/5/1/7	'CDG syndrome', 'Congenital Disorders of Glycosylation', 'Carbohydrate deficient glycoprotein syndrome'
6-1-1	Mayer-Rokitansky-Kster-Hauser syndrome	17 year old, woman, congenital right pulmonary hypoplasia, right hip dysplasia, absence of uterus, rudimentary uterine horn	http://www.ojrd.com/content/5/1/6	'MURCS association', 'Klippel-Feil deformity - conductive deafness - absent vagina', 'Mullerian aplasia - renal aplasia - cervicothoracic somite dysplasia', 'MRKH syndrome', 'Rokitansky sequence', 'Rokitansky syndrome'
7-1-1	Congenital hepatic fibrosis (CHF)	10 year old, girl, thrombocytopenia, splenomegaly, headache, itching rubeoliform rash	http://www.ojrd.com/content/5/1/4	
8-1-1	Hypophosphatemic rickets with hypercalciuria	11 year old, girl, intermittent abdominal pain, mild dorsal scoliosis, low serum phosphate / hypophosphatemia, hypercalcuria, elevated serum 1.25 dihydroxyvitamin D	http://www.ojrd.com/content/5/1/1	
9-1-1	Type I tyrosinemia Synonyms: Fumarylacetoacetase deficiency, Hepatorenal tryosinosis / tyrosinemia	4 month old, boy, epistaxis, haematemesis, haematochezia, subconjunctival bleeding, petechiae, haematomas, haemangioma, slightly enlarged liver, elevated serum transaminases	http://www.ojrd.com/content/4/1/28	'Tyrosinemia type 1', 'Hepatorenal tyrosinemia', 'Fumarylacetoacetase deficiency'
10-1-1	Loeys-Dietz syndrome (LDS) type I	7 year old, boy, dysmorphic signs, blue sclerae, high-arched palate, bifid uvula, joint hypermobility, muscular hypotrophy, translucent skin, aortic root dilatation, camptodactyly and ulnar deviation	http://www.ojrd.com/content/4/1/24	'Aortic aneurysm syndrome, Loeys-Dietz type', 'Aortic aneurysm syndrome, due to TGFbta receptors anomalies'

Table A.1: **Rare diseases query collection (part 1)**

Query ID	Correct Diagnosis	Query	Source	Synonyms (from Orphanet)
10-2-1	Loeys-Dietz syndrome (LDS) type II	48 year old, woman, aortic aneurysm, haematoma, translucent skin, bilateral venous varicosities, recurrent wrist dislocations	http://www.ojrd.com/content/4/1/24	'Aortic aneurysm syndrome, Loeys-Dietz type', 'Aortic aneurysm syndrome, due to TGFbta receptors anomalies'
11-1-1	Arterial tortuosity syndrome (ATS)	8 months old, male, progressive signs of respiratory distress, tachypnea, pulmonary hypertension, tortuosity of aortic arch, facial dysmorphisms	http://www.ojrd.com/content/4/1/20	
11-2-1	Arterial tortuosity syndrome (ATS)	5 year old, male, dyspnoea, asthenia, pulmonary hypertension, severe stenoses elongation and tortuosity of pulmonary arteries branches aortic arch sovraortic trunks and iliac arteries, dysmorphic features, joints hypermobility	http://www.ojrd.com/content/4/1/20	
12-1-1	Whipple's disease Synonyms: Intestinal lipodystrophy, Intestinal lipophagic granulomatosis, Secondary non-tropical sprue	64 year old, male, inflammatory back pain, flares of arthritis, multisegmental spondylitis	http://www.ojrd.com/content/4/1/13	'Whipple disease', 'Intestinal lipodystrophy', 'Secondary non-tropical sprue', 'Intestinal lipophagic granulomatosis'
13-1-1	Pulmonary hemorrhage syndrome associated with dengue fever	70 year old, male, massive hemoptysis, respiratory distress, anemia, hemodynamic instability, renal failure, intense headache, arthralgia, myalgias, ecchymoses over arms and abdomen, acidosis, pleural effusions, blood tinged secretion from lungs	http://www.ojrd.com/content/4/1/8	'Dengue hemorrhagic fever'
14-1-1	Abetalipoproteinemia (ABL) Synonyms: Bassen-Kornzweig disease, Homozygous familial hypobetalipoproteinemia (HoFHBL)	46 year old, female, ptosis, acanthocytosis, history of diarrhea, ataxia, paresthesia	http://www.ojrd.com/content/3/1/19	'Acanthocytosis', 'Abetalipoproteinemia', 'Bassen-Kornzweig disease', 'Homozygous familial hypobetalipoproteinemia'
14-2-1	Abetalipoproteinemia (ABL) Synonyms: Bassen-Kornzweig disease, Homozygous familial hypobetalipoproteinemia (HoFHBL)	16 year old, girl, persistent diarrhea, acanthocytosis, mild dysarthria, reduced muscle bulk, bilateral proximal muscle weakness, absent deep-tendon reflexes, upgoing plantar reflexes, reduced sensitivity to light, dysdiadochokinesia	http://www.ojrd.com/content/3/1/19	'Acanthocytosis', 'Abetalipoproteinemia', 'Bassen-Kornzweig disease', 'Homozygous familial hypobetalipoproteinemia'
15-1-1	Methylmalonic acidemia (MMA) Synonyms: Methylmalonic aciduria	teenager, girl, hypotonia, dehydration, acidosis, massive ketonuria, hyperammonemia	http://www.ojrd.com/content/3/1/2	'Adenosylcobalamin deficiency', 'Methylmalonic acidemia, vitamin B12 responsive', 'Methylmalonic aciduria, vitamin B12 responsive'
15-2-1	Propionic acidemia (PA) Synonyms: Propionic aciduria, Ketotic glycinemia, Propionyl-CoA carboxylase deficiency	girl, hypotonia, seizures, dehydration, polypnea, acidosis, massive ketonuria, hyperammonemia	http://www.ojrd.com/content/3/1/2	'Propionic acidemia', 'Glycinemia, ketotic', 'Propionyl-CoA carboxylase deficiency'
16-1-1	Alstrom syndrome (Alstrm syndrome)	27 year old, woman, blindness, obesity, type 2 diabetes, renal dysfunction, chronic pyelonephritis, hypertension, hirsutism, retinitis pigmentosa, cataract	http://www.ojrd.com/content/2/1/49	
17-1-1	Pulmonary alveolar proteinosis (PAP)	17 year old, boy, lysinuric protein intolerance, mild restrictive functional impairment, digital clubbing, atypical abdominal and thoracic pain, ground glass attenuation, interlobular septa thickening, moderate restrictive ventilatory defect, mild anemia, thrombocytopenia, increase in lactate dehydrogenase	http://www.ojrd.com/content/2/1/14	
18-1-1	Ligase IV deficiency syndrome (LIG4 syndrome) Synonyms: Ligase 4 syndrome	girl, pronounced microcephaly, short stature, psychomotoric delay, distinctive facial appearance, thrombocytopenia, anemia, leukocytopenia, pancytopenia, growth retardation, telecanthus, epicanthal folds, ptosis, infections of the inner ear and respiratory tract, hypoplastic marrow with cellular dysplasia	http://www.ojrd.com/content/2/1/5	'Ligase 4 syndrome'
19-1-1	Oromandibular-limb hypogenesis-Mbius syndrome	5 year old, boy, congenital malformations, malformations of the hands and feet, bilateral strabismus, small tongue, impaired coordination, expressionless face, prominent forehead, depressed nasal bridge, hypoplastic thumbs, bilateral adactyly of the feet, short stature, severe myopia	http://www.ojrd.com/content/2/1/2	'Moebius syndrome', 'Congenital facial diplegia'
20-1-1	Terminal deletion of chromosome 4q	21 year old, female, irregular menses, menorrhagia, hand and foot malformation, ovarian cyst, basic cognitive function	http://www.ojrd.com/content/2/1/9	'Deletion 4q', 'Monosomy 4q', 'Distal monosomy 4q', 'Distal deletion 4q', 'Telomeric deletion 4q', 'Non-distal monosomy 4q', 'Non-distal deletion 4q', 'Non-telomeric monosomy 4q'

Table A.2: **Rare diseases query collection (part 2)**

Query ID	Initial Query	Truncated Query
H5-1	Male child, malformations at birth, midfacial retraction with a deep groove under the eyes, and hypertelorism, short nose with a low nasal bridge and large lowset ears, wide mouth and retrognathia. Hypertrichosis with bright reddish hair and a median frontal cutaneous angioma, short neck with redundant skin, Bilateral inguinal hernias, hypospadias with a megameatus, and cryptorchidism	Male child, malformations at birth, midfacial retraction with a deep groove under the eyes, and hypertelorism, short nose with a low nasal bridge and large lowset ears, wide mouth and retrognathia, Hypertrichosis with bright reddish hair and a median frontal cutaneous angioma, short neck
17-1-1	17 year old, boy, lysinuric protein intolerance, mild restrictive functional impairment, digital clubbing, atypical abdominal and thoracic pain, ground glass attenuation, interlobular septa thickening, moderate restrictive ventilatory defect, mild anemia, thrombocytopenia, increase in lactate dehydrogenase	17 year old, boy, lysinuric protein intolerance, mild restrictive functional impairment, digital clubbing, atypical abdominal and thoracic pain, ground glass attenuation, interlobular septa thickening, moderate restrictive ventilatory defect, mild anemia, thrombocytopenia, increase in lactate

Table A.3: **The two truncated queries from the *rare diseases query collection***

BMJ Case	BMJ Synopsys	BMJ Google Search Terms	BMJ Final Diagnosis	In Orphanet?	In Rare?	In RareGenet?	Synonyms (from Orphanet)
5	53 yo man with depression, Aortic regurg, heart block and acute puloedema.	Acute Aortic regurgitation, depression, abscess	Infective endocarditis	Yes - Eosinophilic endocarditis (ORPHA75566)	Yes	Yes	Loeffler's endocarditis
6	58 yo newly diagnosed oesophageal cancer, refractory hic cups and vomiting	oesophageal cancer, refractory hic cups, nausea, vomiting	Linitis plastica with bowel obstruction	Yes - Linitis plastica of the stomach (ORPHA36273)	Yes - Gastric linitis plastica	Yes	Gastric LP, Gastric linitis plastica, Borrmann gastric cancer type 4, Linitis plastica of the stomach
7	59 yo with difficult to control hypertension, exsmoker with adrenal mass	hypertension, adrenal mass	Cushings secondary to adrenal adenoma	Yes - Familial adrenal adenoma (ORPHA404), Cushing syndrome (ORPHA553)	Yes - Cushing, Adrenal adenoma, Hyperaldosteronism	Yes - Cushing, Adrenal adenoma, Hyperaldosteronism	"Cushing syndrome, Hyperadrenocorticism, Hypercortisolism, Familial adrenal adenoma, FH2, Familial hyperaldosteronism type 2"
8	10 yo boy with right thigh pain and CT showed lytic R hip lesion	hip lesion, older child	Osteoid osteoma	No	Yes	Yes	-
9	67 yo man with acute respiratory failure, exposure to bird dropping	HRCT centrilobular nodules, acute respiratory failure	Hot tub lung secondary to M avium	Yes - type of Hypersensitivity Pneumonitis (ORPHA31740)	Yes - Hypersensitivity Pneumonitis, Mycobacterium Avium	Yes - Hypersensitivity Pneumonitis, Mycobacterium Avium	Hypersensitivity pneumonitis, HP, Extrinsic allergic alveolitis, EAA, Mycobacterium avium, Mycobacterium avium-intracellulare, MAI
10	73 yo fever, thigh pain, urinary frequency, previous statin use	fever, bilateral thigh pain, weakness	Ehrlichiosis	Yes (ORPHA1902)	Yes	Yes	-
11	30 yo female with fever and anterior mediastinal mass	fever, anterior mediastinal mass and central necrosis	Lymphoma	Yes (ORPHA223735)	Yes - types of lymphoma	Yes - types of lymphoma	-
12	48 yo man with multiple spinal tumours and skin tumours	multiple spinal tumours, skin tumours	Neurofibromatosis type 1	Yes (ORPHA636)	Yes	Yes	"NF1, Von Recklinghausen disease"
14	38 yo man with ulcerative colitis, fever, blurred vision and dyspnoea	ulcerative colitis, blurred vision, fever	Vasculitis	Yes (ORPHA52759)	Yes	Yes	"Vasculitides, Systemic vasculitis"
15	80 yo man with dyspnoea and proteinuria	nephrotic syndrome, Bence Jones, ventricular failure	Amyloid light chain	Yes - Amyloidosis (ORPHA69)	Yes - Amyloidosis	Yes - Amyloidosis	AL amyloidosis
16	9 yo female with headache, hypertension, visual disturbance	hypertension, papilledema, headache, renal mass, cafe au lait	Pheochromocytoma	Yes - Pheochromocytoma and secreting paraganglioma (ORPHA717)	Yes	Yes	Phaeochromocytoma, PCC
17	22 yo female with back pain, pulmonary infiltrates, rapidly progressing to death	sickle cell, pulmonary infiltrates, back pain	Acute chest syndrome	No (*it is a complication of Sickle Cell Disease)	No	No	-
18	45 yo female with painful abdo mass	fibroma, astrocytoma, tumor, leiomyoma, scoliosis	Endometriosis	Yes - Rare endometriosis (ORPHA137820)	Yes	Yes	-
19	17 yo female Tsunami survivor with respiratory distress and R hemiplegia	pulmonary infiltrates, cns lesion	Aspiration pneumonia and brain abscess (polymicrobial)	No	No - Aspiration Pneumonia only as symptom in other conditions	No - Aspiration Pneumonia only as symptom in other conditions	-
22	81 yo with cough, fever, weakness and confusion.	CLL, encephalitis	West Nile fever	Yes (ORPHA83476)	Yes - West Nile encephalitis	Yes	West-Nile encephalitis
25	40 yo with wt loss, sweats and persistent fever after food poisoning.	portal vein thrombosis, cancer	Pylephlebitis	No	No	No	Infective suppurative thrombosis of the portal vein

Table A.4: **Difficult cases query collection (part 1).** Source article for the synopsis, search terms and final diagnosis from `http://www.bmj.com/content/333/7579/1143`

BMJ Case	BMJ Synopsys	BMJ Google Search Terms	BMJ Final Diagnosis	In Orphanet?	In Rare?	In RareGenet?	Synonyms (from Orphanet)
26	48 yo man with loss of consciousness while jogging	cardiac arrest, exercise, young	Hypertrophic Obstructive Cardiomyopathy (HOCM)	Yes (ORPHA217569)	Yes - Hypertrophic cardiomyopathy, Familial hypertrophic cardiomyopathy	Yes - Hypertrophic cardiomyopathy, Familial hypertrophic cardiomyopathy	"Hypertrophic cardiomyopathy, Hypertrophic subaortic stenosis, Obstructive hypertrophic cardiomyopathy"
27	80 yo man with fatigue, unsteady gait, confusion, insomnia leading to death	ataxia, confusion, insomnia, death	Creutzfeldt-Jakob disease (CJD)	Yes (ORPHA204)	Yes	Yes	CJD
28	42 yo with 20kg wt loss, weakness, rash, haemapturia and mild hameoptysis	wheeze wt loss, ANCA, haemoptysis, haematuria	Churg Strauss	Yes (ORPHA183)	Yes	Yes	CSS, Churg-Strauss syndrome, Granulomatous allergic angiitis
29	68 yo man with periorbital swelling, rash and weakness	myopathy, neoplasia, dysphagia, rash, periorbital swelling	Dermatomyositis secondary to NHL	Yes - Dermatomyositis (ORPHA221), Non Hodgkin lymphoma NHL (ORPHA547)	Yes - Dermatomyositis, NHL	Yes - Dermatomyositis, NHL	DM, NHL, Non-Hodgkin lymphoma, Non-Hodgkin's malignant lymphomas
30	56 yo renal transplant recipient with fever, lymphadenopathy and cat scratches	renal transplant, fever, cat, lymphadenopathy	Cat scratch disease	Yes (ORPHA50839)	Yes	Yes	CSD, Bartonellosis due to Bartonella henselae infection
31	60 yo man with buttock purpuric rash, chronic renal failure.	buttock rash, renal failure, edema	Cryoglobulinaemia	Yes (ORPHA91139)	Yes	Yes	Cryoglobulinemia
33	43 yo man with lower GI bleed, epistaxis, pulmonary AVM and polyposis.	polyps, telangectasia, epistaxis, anemia	MADH4 mutation (HTT + juvenile polyposis)	Yes - Juvenile gastrointestinal polyposis (ORPHA2929)	Yes - Juvenile polyposis syndrome, Juvenile polyposis of infancy, Juvenile gastrointestinal polyposis	Yes - same and Juvenile Polyposis/Hereditary Hemorrhagic Telangiectasia Syndrome	JIP, JPS, Juvenile intestinal polyposis, Juvenile polyposis syndrome
34	10 yo girl with bullous skin lesions and acute respiratory failure	bullous skin conditions, respiratory failure, carbamazepine	Toxic Epidermal Necrolysis Syndrome (TENS)	Yes - Toxic epidermal necrolysis (ORPHA95455)	Yes - Toxic Epidermal Necrolysis	Yes	Toxic epidermal necrolysis, TEN, SJS-TEN, Toxic epidermolysis
36	61 yo female with seizures, gait disturbance, confusion and dysphasia	seizure, confusion, dysphasia, T2 lesions	MELAS	Yes (ORPHA550)	Yes	Yes	Mitochondrial myopathy-encephalopathy-lactic acidosis and stroke-like episodes
37	35 yo man who had a cardiac arrest while sleeping	cardiac arrest sleep	Brugada	Yes (ORPHA130)	Yes	Yes	"BrS, Idiopathic ventricular fibrillation Brugada type, SUNDS, Sudden unexplained nocturnal death syndrome"
			22 of 26 (84.62%)	23 of 26 (88.46%)	23 of 26 (88.46%)		

Table A.5: **Difficult cases query collection (part 2).** Source article for the synopsis, search terms and final diagnosis from `http://www.bmj.com/content/333/7579/1143`

Appendix B

Rare Disease Search Engines

(I) **Vertical Search Engine for Rare Disease Information Retrieval**

`http://paula.grid.info.uvt.ro/`

APIs:
http://paula.grid.info.uvt.ro/index.[outputformat]?q=[querytext]
where *outputformat* can take the values xml, json, html, and pdf, and the *querytext* must be encoded using the percent-encoding as described in section 2.1 of RFC3986[1].

Code and technical documentation:
`http://code.google.com/p/raredisss/`
The vertical search engine source code is released under the GPL v2 license. On the project website there is a wiki detailing various aspects of the development, implementation, and deployment.

Experimental disease ranking:
`https://costanza.dragusin.ro/rdcdss/default/index`
An experimental version of the vertical search engine, that returns a ranked list of diseases instead of documents.

(II) **Rare Diseases Google CSE Web**

`http://www.google.com/cse/home?cx=017334630119578613104:3s1zsbg1vec`

(III) **Rare Diseases Google CSE Restricted**

`http://www.google.com/cse/home?cx=017334630119578613104:ogownfaoj28`

[1]Uniform Resource Identifier (URI): Generic Syntax, `http://tools.ietf.org/html/rfc3986#section-2.1`

Appendix C

Evaluation Results

Query ID	Rare Index			Reciprocal Rank	Precision at Rank 10	Precision at Rank 20
	Rank	No. relevant @10	No. relevant @20			
1-1	2	1	1	0.500	0.10	0.05
2-1	2	3	5	0.500	0.30	0.25
3-1	2	1	1	0.500	0.10	0.05
4-1	1	2	2	1.000	0.20	0.10
5-1	1	1	1	1.000	0.10	0.05
1-1-1	-	-	-	0.000	0.00	0.00
2-1-1	1	1	1	1.000	0.10	0.05
2-2-1	2	1	1	0.500	0.10	0.05
3-1-1	4	1	1	0.250	0.10	0.05
4-1-1	4	2	2	0.250	0.20	0.10
5-1-1	1	1	1	1.000	0.10	0.05
6-1-1	-	-	-	0.000	0.00	0.00
7-1-1	-	-	-	0.000	0.00	0.00
8-1-1	2	4	4	0.500	0.40	0.20
9-1-1	-	-		0.000	0.00	0.00
10-1-1	1	2	2	1.000	0.20	0.10
10-2-1	1	2	3	1.000	0.20	0.15
11-1-1	2	1	1	0.500	0.10	0.05
11-2-1	1	2	2	1.000	0.20	0.10
12-1-1	-	-	-	0.000	0.00	0.00
13-1-1	-	-	-	0.000	0.00	0.00
14-1-1	1	6	9	1.000	0.60	0.45
14-2-1	-	-	-	0.000	0.00	0.00
15-1-1	7	1	1	0.143	0.10	0.05
15-2-1	5	2	2	0.200	0.20	0.10
16-1-1	2	2	3	0.500	0.20	0.15
17-1-1	1	1	1	1.000	0.10	0.05
18-1-1	-	-	-	0.000	0.00	0.00
19-1-1	-	-	-	0.000	0.00	0.00
20-1-1	-	-	-	0.000	0.00	0.00
				MRR	Average P@10	Average P@20
				0.445	0.123	0.073

Table C.1: Summary per query for retrieval from the *Rare* index on the *rare diseases query collection.*

Query ID	RareGenet Index			RareGenet - with boost = 2			RareGenet - with boost = 4		
	Rank	No. relevant @10	No. relevant @20	Rank	No. relevant @10	No. relevant e@20	Rank	No. relevant @10	No. relevant @20
1-1	1	2	2	1	2	2	1	2	2
2-1	3	1	3	3	2	3	3	2	3
3-1	9	1	1	8	1	1	8	1	1
4-1	1	3	3	1	3	3	1	3	3
5-1	15	0	1	15	0	1	15	0	1
1-1-1	-	-	-	-	-	-	-	-	-
2-1-1	1	2	2	1	2	2	1	2	2
2-2-1	2	2	2	2	2	2	2	2	2
3-1-1	3	1	1	3	1	1	3	1	1
4-1-1	7	1	3	6	1	3	5	1	3
5-1-1	1	3	3	1	3	3	1	3	3
6-1-1	3	2	2	3	2	2	3	2	2
7-1-1	-	-	-	-	-	-	-	-	-
8-1-1	1	4	6	1	4	6	1	4	6
9-1-1	-	-	-	-	-	-	-	-	-
10-1-1	1	5	6	1	5	6	1	5	6
10-2-1	5	2	5	5	3	5	5	4	5
11-1-1	6	1	1	6	1	2	6	1	2
11-2-1	1	2	4	1	2	3	1	2	3
12-1-1	-	-	-	-	-	-	-	-	-
13-1-1	-	-	-	-	-	-	-	-	-
14-1-1	1	6	6	1	6	8	1	6	9
14-2-1	17	0	1	17	0	1	18	0	1
15-1-1	1	4	4	1	4	5	1	4	5
15-2-1	4	1	2	4	2	2	4	2	3
16-1-1	2	1	2	2	1	2	2	2	3
17-1-1	1	2	2	1	2	2	1	2	2
18-1-1	-	-	-	-	-	-	-	-	-
19-1-1	1	1	1	1	1	1	1	1	1
20-1-1	-	-	-	-	-	-	-	-	-

Query ID	Reciprocal Rank			Precision at Rank 10			Precision at Rank 20		
	RareGenet	RareGenet (boost=2)	RareGenet (boost=4)	RareGenet	RareGenet (boost=2)	RareGenet (boost=4)	RareGenet	RareGenet (boost=2)	RareGenet (boost=4)
1-1	1.000	1.000	1.000	0.20	0.20	0.20	0.10	0.10	0.10
2-1	0.333	0.333	0.333	0.10	0.20	0.20	0.15	0.15	0.15
3-1	0.111	0.125	0.125	0.10	0.10	0.10	0.05	0.05	0.05
4-1	1.000	1.000	1.000	0.30	0.30	0.30	0.15	0.15	0.15
5-1	0.067	0.067	0.067	0.00	0.00	0.00	0.05	0.05	0.05
1-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
2-1-1	1.000	1.000	1.000	0.20	0.20	0.20	0.10	0.10	0.10
2-2-1	0.500	0.500	0.500	0.20	0.20	0.20	0.10	0.10	0.10
3-1-1	0.333	0.333	0.333	0.10	0.10	0.10	0.05	0.05	0.05
4-1-1	0.143	0.167	0.200	0.10	0.10	0.10	0.15	0.15	0.15
5-1-1	1.000	1.000	1.000	0.30	0.30	0.30	0.15	0.15	0.15
6-1-1	0.333	0.333	0.333	0.20	0.20	0.20	0.10	0.10	0.10
7-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
8-1-1	1.000	1.000	1.000	0.40	0.40	0.40	0.30	0.30	0.30
9-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
10-1-1	1.000	1.000	1.000	0.50	0.50	0.50	0.30	0.30	0.30
10-2-1	0.200	0.200	0.200	0.20	0.30	0.40	0.25	0.25	0.25
11-1-1	0.167	0.167	0.167	0.10	0.10	0.10	0.05	0.10	0.10
11-2-1	1.000	1.000	1.000	0.20	0.20	0.20	0.20	0.15	0.15
12-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
13-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
14-1-1	1.000	1.000	1.000	0.60	0.60	0.60	0.30	0.40	0.45
14-2-1	0.059	0.059	0.056	0.00	0.00	0.00	0.05	0.05	0.05
15-1-1	1.000	1.000	1.000	0.40	0.40	0.40	0.20	0.25	0.25
15-2-1	0.250	0.250	0.250	0.10	0.20	0.20	0.10	0.10	0.15
16-1-1	0.500	0.500	0.500	0.10	0.10	0.20	0.10	0.10	0.15
17-1-1	1.000	1.000	1.000	0.20	0.20	0.20	0.10	0.10	0.10
18-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
19-1-1	1.000	1.000	1.000	0.10	0.10	0.10	0.05	0.05	0.05
20-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
	Mean reciprocal rank (MRR)			Average Precision at rank 10			Average Precision at rank 20		
	0.467	0.468	0.469	0.157	0.167	0.173	0.105	0.110	0.115

Table C.2: Summary per query for retrieval from the *RareGenet* index (with and without boost) on the *rare diseases query collection* for $\mu = 2500$

Query ID	Google CSE Restricted			Google CSE Web			Google		
	Rank	No. relevant @10	No. relevant @20	Rank	No. relevant @10	No. relevant @20	Rank	No. relevant @10	No. relevant @20
1-1	1	1	1	1	1	1	10	1	1
2-1	-	-	-	19	0	1	-	-	-
3-1	-	-	-	-	-	-	-	-	-
4-1	-	-	-	-	-	-	-	-	-
5-1	-	-	-	1	2	2	1	2	2
1-1-1	-	-	-	-	-	-	-	-	-
2-1-1	-	-	-	-	-	-	-	-	-
2-2-1	-	-	-	-	-	-	-	-	-
3-1-1	-	-	-	-	-	-	-	-	-
4-1-1	-	-	-	1	1	1	7	1	2
5-1-1	-	-	-	1	3	3	3	2	2
6-1-1	-	-	-	7	1	1	10	1	1
7-1-1	-	-	-	-	-	-	-	-	-
8-1-1	-	-	-	-	-	-	-	-	-
9-1-1	-	-	-	-	-	-	-	-	-
10-1-1	-	-	-	1	1	1	-	-	-
10-2-1	-	-	-	-	-	-	-	-	-
11-1-1	-	-	-	-	-	-	-	-	-
11-2-1	-	-	-	-	-	-	-	-	-
12-1-1	-	-	-	-	-	-	-	-	-
13-1-1	-	-	-	-	-	-	-	-	-
14-1-1	-	-	-	-	-	-	-	-	-
14-2-1	-	-	-	-	-	-	-	-	-
15-1-1	-	-	-	-	-	-	-	-	-
15-2-1	-	-	-	-	-	-	-	-	-
16-1-1	-	-	-	-	-	-	-	-	-
17-1-1	-	-	-	-	-	-	-	-	-
18-1-1	-	-	-	-	-	-	-	-	-
19-1-1	-	-	-	-	-	-	-	-	-
20-1-1	-	-	-	-	-	-	-	-	-

Query ID	Reciprocal Rank			Precision at Rank 10			Precision at Rank 20		
	Google CSE Restr.	Google CSE Web	Google	Google CSE Restr.	Google CSE Web	Google	Google CSE Restr.	Google CSE Web	Google
1-1	1.000	1.000	0.100	0.10	0.10	0.10	0.05	0.05	0.05
2-1	0.000	0.053	0.000	0.00	0.00	0.00	0.00	0.05	0.00
3-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
4-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
5-1	0.000	1.000	1.000	0.00	0.20	0.20	0.00	0.10	0.10
1-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
2-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
2-2-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
3-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
4-1-1	0.000	1.000	0.143	0.00	0.10	0.10	0.00	0.05	0.10
5-1-1	0.000	1.000	0.333	0.00	0.30	0.20	0.00	0.15	0.10
6-1-1	0.000	0.143	0.100	0.00	0.10	0.10	0.00	0.05	0.05
7-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
8-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
9-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
10-1-1	0.000	1.000	0.000	0.00	0.10	0.00	0.00	0.05	0.00
10-2-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
11-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
11-2-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
12-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
13-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
14-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
14-2-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
15-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
15-2-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
16-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
17-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
18-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
19-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
20-1-1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00
	Mean reciprocal rank (MRR)			Average Precision at rank 10			Average Precision at rank 20		
	0.033	0.173	0.056	0.003	0.030	0.023	0.002	0.017	0.013

Table C.3: Summary per query for Google Search, Google CSE Web, and Google CSE Restricted on the *rare diseases query collection*.

BMJ Case	Google on Search Terms			Rare on Search Terms			RareGenet on Search Terms			PubMed on Search Terms		
	Rank	No. relevant @10	No. relevant @20	Rank	No. relevant @10	No. relevant @20	Rank	No. relevant @10	No. relevant @20	Rank	No. relevant @10	No. relevant @20
5	-	-	-	-	-	-	-	-	-	1	1	1
6	-	-	-	-	-	-	-	-	-	-	-	-
7	1	1	1	8	1	1	5	1	2	1	3	3
8	-	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-	-	-	-
11	1	2	2	13	0	2	9	1	1	-	-	-
12	4	2	4	-	-	-	7	2	2	11	0	3
14	-	-	-	-	-	-	-	-	-	-	-	-
15	1	5	10	-	-	-	-	-	-	-	-	-
16	13	0	1	-	-	-	-	-	-	-	-	-
17	2	5	8	-	-	-	-	-	-	1	1	1
18	-	-	-	-	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	-	-	-	-	-
22	18	0	1	9	1	2	10	1	2	-	-	-
25	-	-	-	-	-	-	-	-	-	-	-	-
26	1	1	1	-	-	-	2	1	1	12	0	3
27	-	-	-	3	2	5	9	2	2	-	-	-
28	1	2	2	8	1	3	7	1	1	-	-	-
29	1	6	12	-	-	-	9	1	1	1	1	1
30	1	5	7	1	4	4	1	1	3	1	3	3
31	-	-	-	-	-	-	-	-	-	-	-	-
33	-	-	-	3	2	2	1	2	2	1	2	2
34	1	2	5	1	1	1	1	2	2	1	1	1
36	-	-	-	2	1	1	6	1	1	-	-	-
37	1	1	1	2	1	1	4	3	3	-	-	-

BMJ Case	Reciprocal Rank				Precision at Rank 10				Precision at Rank 20			
	Google	Rare	Rare Genet	PubMed	Google	Rare	Rare Genet	PubMed	Google	Rare	Rare Genet	PubMed
5	0	0	0	1	0	0	0	0.1	0	0	0	0.05
6	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0.125	0.2	1	0.1	0.1	0.1	0.3	0.05	0.05	0.1	0.15
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	1	0.077	0.111	0	0.2	0	0.1	0	0.1	0.1	0.05	0
12	0.25	0	0.143	0.091	0.2	0	0.2	0	0.2	0	0.1	0.15
14	0	0	0	0	0	0	0	0	0	0	0	0
15	1	0	0	0	0.5	0	0	0	0.5	0	0	0
16	0.077	0	0	0	0	0	0	0	0.05	0	0	0
17	0.5	0	0	1	0.5	0	0	0.1	0.4	0	0	0.05
18	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0
22	0.056	0.111	0.1	0	0	0.1	0.1	0	0.05	0.1	0.1	0
25	0	0	0	0	0	0	0	0	0	0	0	0
26	1	0	0.5	0.083	0.1	0	0.1	0	0.05	0	0.05	0.15
27	0	0.333	0.111	0	0	0.2	0.2	0	0	0.25	0.1	0
28	1	0.125	0.143	0	0.2	0.1	0.1	0	0.1	0.15	0.05	0
29	1	0	0.111	1	0.6	0	0.1	0.1	0.6	0	0.05	0.05
30	1	1	1	1	0.5	0.4	0.1	0.3	0.35	0.2	0.15	0.15
31	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0.333	1	1	0	0.2	0.2	0.2	0	0.1	0.1	0.1
34	1	1	1	1	0.2	0.1	0.2	0.1	0.25	0.05	0.1	0.05
36	0	0.5	0.167	0	0	0.1	0.1	0	0	0.05	0.05	0
37	1	0.5	0.25	0	0.1	0.1	0.3	0	0.05	0.05	0.15	0
	Mean reciprocal rank (MRR)				Average Precision at rank 10				Average Precision at rank 20			
	0.38	0.158	0.186	0.276	0.123	0.054	0.073	0.046	0.106	0.042	0.044	0.035

Table C.4: Summary per query for Google Search, *Rare*, *RareGenet*, and PubMed on the *difficult cases query collection*.

Figure C.1: Per query performance on the *rare diseases query collection* for the vertical search engine using the *RareGenet* index, and for Google CSE Web

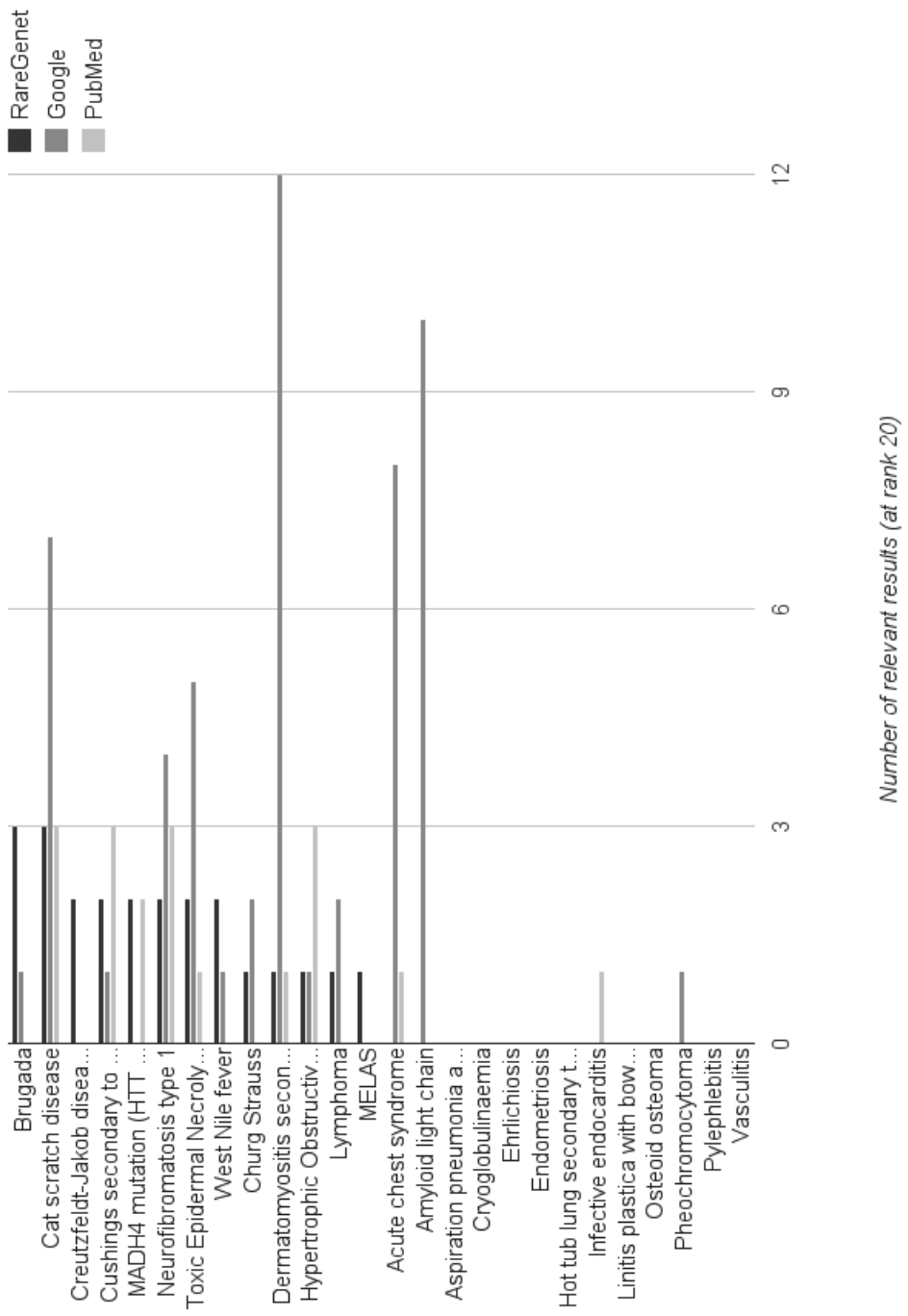


Figure C.2: Per query performance on the *difficult case query collection* for the vertical search engine using the *RareGenet* index, Google Search, and PubMed.

	Rare	RareGenet
Total number of cases	30	30
Correct diagnosis found	18	24
Correct diagnosis in top 10	13	1
Correct diagnosis in top 11-20	0	3
Correct diagnosis in top 21-30	1	5
Correct diagnosis in top 31-40	4	4
Correct diagnosis in top 41-50	0	6
Correct diagnosis in top 51-100	0	3
Correct diagnosis after 100	0	2
Correct diagnosis not found	12	6
Average no. of results	54	326
Mean reciprocal rank (MRR)	0.156	0.028
Average precision rank 10 (P@10)	0.057	0.003
Average precision rank 20 (P@20)	0.033	0.010
Average precision rank 50 (P@50)	0.025	0.018
Average precision rank 100 (P@100)	0.012	0.015

Table C.5: Disease ranking results for the *rare diseases query collection*.

	Rare	RareGenet
Total number of cases	26	26
Correct diagnosis found	14	16
Correct diagnosis in top 10	7	3
Correct diagnosis in top 11-20	2	2
Correct diagnosis in top 21-30	4	0
Correct diagnosis in top 31-40	0	0
Correct diagnosis in top 41-50	1	1
Correct diagnosis in top 51-100	0	6
Correct diagnosis after 100	0	4
Correct diagnosis not found	12	10
Average no. of results	52	342
Mean reciprocal rank (MRR)	0.108	0.039
Average precision rank 10 (P@10)	0.042	0.012
Average precision rank 20 (P@20)	0.029	0.013
Average precision rank 50 (P@50)	0.022	0.009
Average precision rank 100 (P@100)	0.011	0.010

Table C.6: Disease ranking results for the *difficult cases query collection*.

Case	Rare Index					RareGenet Index				
	Rank	No. relevant @10	No. relevant @20	No, relevant @50	No. relevant @100	Rank	No. relevant @10	No. relevant @20	No. relevant @50	No. relevant @100
1-1	6	1	2	2	2	13	0	1	1	2
2-1	2	1	1	2	2	110	0	0	0	0
3-1	26	0	0	1	1	100	0	0	0	1
4-1	3	1	1	1	1	18	0	2	2	2
5-1	-	-	-	-	-	59	0	0	0	1
1-1-1	-	-	-	-	-	-	-	-	-	-
2-1-1	39	0	0	4	4	48	0	0	1	1
2-2-1	35	0	0	4	4	44	0	0	1	3
3-1-1	35	0	0	2	2	39	0	0	1	2
4-1-1	3	1	1	1	1	51	0	0	0	1
5-1-1	5	1	1	1	1	29	0	0	2	2
6-1-1	-	-	-	-	-	45	0	0	1	2
7-1-1	-	-	-	-	-	329	0	0	0	0
8-1-1	-	-	-	-	-	33	0	0	3	3
9-1-1	-	-	-	-	-	-	-	-	-	-
10-1-1	2	1	1	1	1	8	1	1	1	3
10-2-1	2	1	1	1	1	48	0	0	1	3
11-1-1	40	0	0	1	1	36	0	0	1	1
11-2-1	5	1	1	1	1	41	0	0	1	1
12-1-1	-	-	-	-	-	-	-	-	-	-
13-1-1	-	-	-	-	-	-	-	-	-	-
14-1-1	2	4	5	8	8	13	0	2	3	5
14-2-1	-	-	-	-	-	24	0	0	3	3
15-1-1	8	1	1	2	2	49	0	0	1	1
15-2-1	2	1	2	2	2	27	0	0	1	3
16-1-1	3	2	2	2	2	25	0	0	1	2
17-1-1	3	1	1	1	1	23	0	0	1	1
18-1-1	-	-	-	-	-	-	-	-	-	-
19-1-1	-	-	-	-	-	39	0	0	1	1
20-1-1	-	-	-	-	-	-	-	-	-	-
	MRR	P@10	P@20	P@50	P@100	MRR	P@10	P@20	P@50	P@100
	0.156	0.057	0.033	0.025	0.012	0.028	0.003	0.01	0.018	0.015

BMJ Case Rec	Rare Index					RareGenet Index				
	Rank	No. relevant @10	No. relevant @20	No. relevant @50	No. relevant @100	Rank	No. relevant @10	No. relevant @20	No. relevant @50	No. relevant @100
5	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-
7	6	1	1	1	1	45	0	0	1	1
8	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-	-
11	14	0	1	6	6	6	1	1	1	3
12	28	0	0	1	1	5	1	2	4	4
14	-	-	-	-	-	-	-	-	-	-
15	42	0	0	2	2	207	0	0	0	0
16	17	0	1	1	1	104	0	0	0	0
17	-	-	-	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	-	-	-
22	4	1	1	1	1	75	0	0	0	1
25	-	-	-	-	-	-	-	-	-	-
26	-	-	-	-	-	78	0	0	0	2
27	1	3	4	5	5	11	0	1	2	4
28	4	1	1	1	1	120	0	0	0	0
29	-	-	-	-	-	60	0	0	0	2
30	2	2	2	2	2	70	0	0	0	1
31	30	0	0	1	1	204	0	0	0	0
33	5	1	2	3	3	3	1	1	2	4
34	23	0	0	1	1	51	0	0	0	1
36	7	2	2	2	2	65	0	0	0	2
37	25	0	0	1	1	14	0	2	2	2
	MRR	P@10	P@20	P@50	P@100	MRR	P@10	P@20	P@50	P@100
	0.108	0.042	0.029	0.022	0.011	0.039	0.012	0.013	0.009	0.01

Table C.7: Summary per query for *Rare* and *RareGenet* for results of the disease ranking algorithm, on the two query collections: *rare diseases query collection* (above) and *difficult cases query collection* (below).

Query ID	Final Diagnosis	Rank of first relevant result		No. of words until the first mention of the correct disease		Clicks
		RareGenet	Google	RareGenet	Google	
H1-1	Fibrodysplasia ossificans progressiva	1	10	0	487	0
HS5-1	Schinzel-Giedion Syndrome	7	1	28	0	0
4-1-1	Cogan's syndrome	5	7	10	359	0
5-1-1	CDG (Congenital Disorders of Glycosylation) syndrome type Ic	1	3	0	83	0
6-1-1	Mayer-Rokitansky-Kster-Hauser syndrome	3	10	10	484	0
				Average number of words		
				9.6	282.6	

Table C.8: Search time for a subset of queries from the *rare diseases query collection*

BMJ Case Rec	Final Diagnosis	Rank of first relevant result			No. of words until the first mention of the correct disease			Clicks	
		RareGenet	Google	PubMed	RareGenet	Google	PubMed	Google	PubMed
7	Cushings secondary to adrenal adenoma	5	1	1	25	0	0	1 click (76 words)	1 click (131 words)
12	Neurofibromatosis type 1	7	4	11	23	136	135	0	0
26	Hypertrophic Obstructive Cardiomyopathy (HOCM)	2	1	12	4	0	160	1 click (250 words)	0
29	Dermatomyositis secondary to NHL	9	1	1	27	0	0	0	0
30	Cat scratch disease	1	1	1	0	0	0	0	0
34	Toxic Epidermal Necrolysis Syndrome (TENS)	1	1	1	0	0	0	0	0
					Average number of words				
					13.2	22.7	49.2		

Table C.9: Search time for a subset of queries from the *difficult cases query collection*.

Appendix D

ICTIR Poster Paper

Rare Disease Diagnosis as an Information Retrieval Task

Radu Dragusin[1], Paula Petcu[1], Christina Lioma[2], Birger Larsen[3], Henrik Jørgensen[4], and Ole Winther[5]

[1] Computer Science, University of Copenhagen, Copenhagen, Denmark
[2] Informatics, Stuttgart University, Stuttgart, Germany
[3] Royal School of Library and Information Science, Copenhagen, Denmark
[4] Department of Clinical Biochemistry, Bispebjerg Hospital, Copenhagen, Denmark
[5] Informatics, Technical University of Denmark, Lyngby, Denmark
{dragusin,petcu}@diku.dk, liomaca@ims.uni-stuttgart.de, blar@iva.dk, hlj@dadlnet.dk, owi@imm.dtu.dk

Abstract. Increasingly more clinicians use web Information Retrieval (IR) systems to assist them in diagnosing difficult medical cases, for instance rare diseases that they may not be familiar with. However, web IR systems are not necessarily optimised for this task. For instance, clinicians' queries tend to be long lists of symptoms, often containing phrases, whereas web IR systems typically expect very short keyword-based queries. Motivated by such differences, this work uses a preliminary study of 30 clinical cases to reflect on rare disease retrieval as an IR task. Initial experiments using both Google web search and offline retrieval from a rare disease collection indicate that the retrieval of rare diseases is an open problem with room for improvement.

Keywords: rare diseases, clinical information retrieval, web diagnosis

1 Introduction

Recently web Information Retrieval (IR) systems have gained popularity among clinicians to assist them in difficult medical cases, for instance rare diseases that they may not be familiar with [1]. However, such systems are not necessarily designed or optimised for diagnosing rare diseases. For example, clinicians' queries tend to be long lists of symptoms, whereas web IR systems typically expect very short queries. Similarly, the hyperlink popularity and recommendation principles typically applied in web IR tend to favour popular webpages; however, information on rare diseases is generally very sparse and less hyperlinked than other medical content. Motivated by such differences, this work considers rare disease diagnosis as an IR task, and asks what design considerations are needed to build an IR system that clinicians can use to diagnose rare diseases?

To address this question, a small preliminary study with 30 real clinical cases is conducted, involving both Google web search and offline retrieval from a

specialised rare disease collection (Section 2). The resulting findings offer useful insights on the special characteristics, possibilities and challenges of rare disease diagnosis as an IR task (Section 3). Section 4 concludes this work.

2 Retrieving rare diseases: preliminary study

The queries used in this work were created from 30 clinical cases of rare diseases, where the query text was extracted directly from the patient symptoms listed in the clinical cases. This was done by one medical doctor and two non-experts. The correct disease diagnosed for these symptoms was not included in the query text. This is an important difference from standard web search queries, where the topic sought is usually explicitly mentioned in the query. The average query length was 22.17 terms. E.g., query for the rare Kleine-Levine syndrome: `Jewish boy age 16, monthly seizures, sleep deficiency, aggressive and irritable when woken, highly increased sexual appetite and hunger`.

The 30 queries were used to retrieve documents using Google web search, and separately using the Indri IR system on a small rare disease collection specifically created for this task. This dataset contains 31,746 documents, crawled from web sites specialising on rare and genetic diseases[6]. Specifically, we collected 10,280 documents on rare diseases and 21,466 documents on genetic diseases (many of which are rare), to be referred to as RARE and GENET henceforth.

Three runs were realised with Google: (1) using standard Google web search; (2) customing Google[7] on the RARE dataset but retrieving documents from the whole web; (3) restricting Google to retrieve from the RARE & GENET websites, plus 5 websites containing only url links to rare disease information (these 5 websites were excluded from our collection because they included url links only). Three more runs were realised with Indri: (4) retrieval from RARE only; (5) retrieval from RARE & GENET; (6) retrieval from RARE & GENET, with a rank boost of RARE documents by a factor of 4.

Runs with Indri used the query likelihood language model with Dirichlet smoothing at default settings ($\mu = 2500$ [2], Krovetz stemming). For run 6, boosting RARE documents was implemented as the prior probability of a document being relevant ($P(D)$). Unless specified otherwise, the baseline query likelihood model assumes that all documents are a priori equally likely to be relevant, and ignores $P(D)$. Motivated by the intuition that RARE documents should have a higher likelihood to include relevant documents when searching for rare diseases, we computed $P(D)$ directly from the collection statistics as follows. Let C denote the complete retrieval collection containing both RARE and GENET. Then, $P(R|C)x + P(G|C)y = 1$, where $x = \phi y$, and where $P(R|C)$ (resp. $P(G|C)$) denotes the probability of all RARE (resp. GENET) documents in the whole collection. ϕ is the boosting factor, set to $\phi = 4$ in this work; this value of ϕ is ad-hoc and untuned, used only for illustration purposes.

[6] The list of urls is available here: http://code.google.com/p/raredisss/wiki/RareGenetResources.
[7] http://www.google.com/cse/

The relevance of the retrieved documents in these 6 runs was assessed by the two non-experts in the top 20 ranks using graded relevance on 3 points (relevant, marginally relevant, non-relevant): (i) relevant documents should address mainly the correct disease in the title or within the first 400 words, and name it using any of its synonyms listed in Orphanet[8]; (ii) in cases of inherited diseases, e.g `autosomal neonatal form of Adrenoleukodystrophy`, documents about the main disease, e.g. `X-linked Adrenoleukodystrophy`, are relevant; (iii) documents about different types of the correct disease, e.g. `Loeys-Dietz syndrome type 1A` instead of `Loeys-Dietz syndrome type II`, are relevant; (iv) documents about other diseases and mentioning the correct disease as an alternative diagnostic or pointing to it are marginally relevant; (v) documents listing many diseases are not relevant if the correct disease is listed after the first 10.

Collection	Retrieval approach	P@10	P@20	MRR	NDCG@10	NDCG@20
WEB	Standard Google	.023	.013	.056	.168	.189
WEB	Google Custom on RARE	.030	.017	.173	.275	.283
RARE&GENET	Google Restricted	.003	.002	.033	.033	.033
RARE	LM-Dir	.123	.073	.445	**.516**	**.536**
RARE&GENET	LM-Dir	.157	.105	.467	.423	.493
RARE&GENET	LM-Dir prior on RARE	**.173**	**.115**	**.469**	.433	.492

Table 1. Retrieval from the web and our rare disease & genetic disease datasets.

Table 1 shows the retrieval precision at rank k (P@k), the mean reciprocal rank (MRR) and the normalised discounted cumulative gain at rank k (NDCG@k) of our 6 runs averaged for all 30 queries. NDCG uses graded relevance assessments[9]; all other measures use binary relevance assessments which consider marginally relevant documents as non-relevant. Retrieval from the web refers to the part of the web indexed by Google. Two findings emerge: (i) Google overall underperforms for this task, especially when restricted to the sites of our collection; (ii) the MRR scores show that on average the correct diagnosis appears at ranks 2-3 with Indri (.445 - .469) and at best at rank 5-6 with Google (.173). Even though the Google retrieval algorithm is not known, a possible reason for this performance may be the fact that it is not optimised for this task. E.g., if Google uses popularity-based metrics like Page-http://code.google.com/p/raredisss/wiki/RareGenetResourcesRank, the desired relevant documents are not likely to be helped by this, because they are not necessarily as heavily hyperlinked as other medical documents; if Google considers logged user & query features like clickthrough data, rare disease queries are not likely to benefit from this, because they are probably not sufficiently frequent among users; the fact that Google does not accept queries longer than 32 terms indicates that it is optimised for queries shorter than our 22.17 word-long queries.

[8] http://www.orpha.net/

[9] with the following gain values: relevant = 3, marginally relevant = 1.

3 The characteristics of rare disease retrieval

The above observations indicate that rare diseases retrieval may be seen as a distinct IR task with the following user-based and system-based characteristics.

On the user side, the clinicians' information needs are ideally fullfilled by a single document about the correct rare disease, similarly to early-precision tasks such as named-page finding. However, the clinicians' queries are expressed in very different ways than named-page or other web search queries: (a) they are very long; (b) they consist of lists of patient symptoms, where term independence assumptions could lead to topic drift (e.g. `sleep deficiency, increased sexual appetite` is topically different to `sexual deficiency, increased sleep`); (c) some symptoms listed in the query may not apply to the correct disease, and conversely, some pertinent symptoms for the correct disease may be missing from the query because they are masked under different conditions. In short, the clinicians' queries on rare diseases are likely to be more feature-rich but also more noisy than in web IR, and should be treated as such.

On the system side, popularity-based metrics derived from hyperlinking, user visit rates, or other forms of recommendation may not benefit the retrieval of rare diseases. Instead, features that may aid this task could be domain-specific enhancements (such as the prior on the RARE dataset), or information about the rarity, geographic distribution and statistics of a disease. Finally, often efficiency concerns lead to brute-force index pruning for web search, e.g. by removing from the index terms of low frequency or that are unusually long. Such practices may be particularly damaging for rare disease retrieval, as the medical terminology involved may be exceptionally rare or formed by heavy term compounding.

4 Conclusion

This work reflected on rare disease diagnosis as an IR task, where clinicians use symptoms as queries in order to retrieve a correct diagnosis. A small preliminary study involving real clinical cases of rare diseases was conducted in collaboration with a medical doctor. Findings revealed that rare disease retrieval has several distinct features that differentiate it from standard web IR, and that applying standard web IR for this task may not be optimal. Future work includes developing IR approaches for the domain of rare diseases.

References

1. M. G. Bouwman, Q. G. A. Teunissen, F. A. Wijburg, and G. E. Linthorst. 'Doctor' Google ending the diagnostic odyssey in lysosomal storage disorders: parents using internet search engines as an efficient diagnostic strategy in rare diseases. *Arch Dis Child*, 95(8):642–4, 2010.
2. C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.