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

G. Amati and F. Crestani (Eds.): ICTIR 2011, LNCS 6931, pp. 356–359, 2011.

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[1]. 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[2] 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.

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

[2] 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[3]; (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.

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

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 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[4]; 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.

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

[4] 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. Bouwman, M.G., Teunissen, Q.G.A., Wijburg, F.A., Linthorst, G.E.: 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), 642–644 (2010)
2. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. 22(2), 179–214 (2004)