Diagnose This If You Can

On the effectiveness of search engines in finding medical self-diagnosis information

Guido Zuccon[1], Bevan Koopman[2], and João Palotti[3]

[1] Queensland University of Technology, Brisbane, Australia, g.zuccon@qut.edu.au
[2] Australian e-Health Research Centre, CSIRO, Brisbane, Australia, bevan.koopman@csiro.au
[3] Vienna University of Technology, Vienna, Austria, palotti@ifs.tuwien.ac.at

Abstract. An increasing amount of people seek health advice on the web using search engines; this poses challenging problems for current search technologies. In this paper we report an initial study of the effectiveness of current search engines in retrieving relevant information for diagnostic medical circumlocutory queries, i.e., queries that are issued by people seeking information about their health condition using a description of the symptoms they observes (e.g. hives all over body) rather than the medical term (e.g. urticaria). This type of queries frequently happens when people are unfamiliar with a domain or language and they are common among health information seekers attempting to self-diagnose or self-treat themselves. Our analysis reveals that current search engines are not equipped to effectively satisfy such information needs; this can have potential harmful outcomes on people's health. Our results advocate for more research in developing information retrieval methods to support such complex information needs.

Keywords: Medical Information Retrieval, Self-Diagnosis, Evaluation, Medical Circumlocution

1 Introduction and Motivations

The use of the Web as source of health-related information is a wide-spread phenomena. Qualitative research carried out by the Pew Research Center has found that 80% of the interviewed U.S.-based population uses the Web to acquire health information [2]. Health-related websites available on the Internet range from those providing information and support for people with diagnosed conditions, to those (developed both from private companies and recognised healthcare providers) suggesting diagnoses for particular symptoms, and those providing self-treatment options and cures [6].

Search engines are commonly used as a means to access health information available online. An analysis of query logs obtained a dozen of years ago from three commercial search engines revealed that health-related queries amounted to about 10% of the total number of queries issued to web search engines [7]. This

trend has grown enormously in recent years [2]. A survey from the Pew Research Center reports that nearly 70% of search engine users in the U.S. have performed health-related searches; many of these searches were for self-diagnosis purposes, and of these about half lead to users seeking professional medical attention [2].

Previous research has, however, shown that exposing people with no or scarce medical knowledge to complex medical language may lead to erroneous self-diagnosis and self-treatment [1]. White and Horvitz have shown that access to medical information on the Web can lead to the escalation of concerns about common symptoms (e.g., cyberchondria) [9].

It is therefore important to develop and evaluate search methodologies that effectively support users in finding topical, high-quality, and accessible health information on the web. The ShARe/CLEF eHealth Evaluation Labs 2013 and 2014 (Task 3) have focused on evaluating information retrieval systems aimed at health consumers to improve how they access medical information on the Web [3,4]. The tasks focused on queries used by health consumers to find information about their diseases or disorders as reported in a discharge summary they were given upon discharge from a hospital admission. The results from the 2014 campaign showed that effective systems can be created using statistical language modelling techniques along with sophisticated query expansion mechanisms based on structured domain knowledge and the exploitation of information from discharge summaries.

The queries investigated by the CLEF evaluation labs so far were seeking information about a medical term (usually the name of a medical condition) users encountered in their discharge summaries. As mentioned above, these are only one part of the health-related queries issued to search engines, with queries aimed at self-diagnosis purposes being another important type of health-related information needs [2,9,10,8]. A recent study by Stanton et al. [8] has suggested that self-diagnosis queries observed from search engines query logs tend to be in a *circumlocutory* form, where the information seeker describes the symptoms they are observing in a colloquial way and using a "talking around" style, instead of the actual medical expression, e.g., [white part of the eye turned green] in place of [jaundice]. Answering such circumlocutory self-diagnosis queries correctly is of critical importance to avoid the risk of harm from incorrect self-diagnosis or self-treatment.

Our Contribution. In this paper, we perform an initial investigation of the effectiveness of current commercial search engines in retrieving information that helps the information seekers to correctly self-diagnose themselves. We investigate 8 main symptoms and for each of these we consider 3 to 4 queries (26 queries in total) obtained from the work of Stanton and colleagues [8], who have proposed a method to generate medical circumlocution diagnostic queries that resemble what users may issue to search for self-diagnosis information. Queries are issued to two commercial search engines (Google and Bing), their search results recorded and assessed to evaluate whether users may find relevant information that helps self-diagnoses their conditions (the 8 main symptoms). The results reveal that only half of the top 10 results retrieved by the considered search

Symptom Group	Crowdsourced Circumlocutory Queries
alopecia	baldness in multiple spots, circular bald spots, loss of hair on scalp in an inch width round
angular cheilitis	broken lips, dry cracked lips, lip sores, sores around mouth
edema	fluid in leg, puffy sore calf, swollen legs
exophthalmos	bulging eye, eye balls coming out, swollen eye, swollen eye balls
hematoma	hand turned dark blue, neck hematoma, large purple bruise on arm
jaundice	yellow eyes, eye illness, white part of the eye turned green
psoriasis	red dry skin, dry irritated skin on scalp, silvery-white scalp + inner ear
urticaria	hives all over body, skin rash on chest, extreme red rash on arm

Table 1. Crowdsourced queries with associated symptoms obtained from [8] and used in this work to evaluate the effectiveness of state-of-the-art search engines.

engines provide information that is somewhat relevant to the self-diagnosis of the medical condition; only about 3 out 10 results on average are highly useful for self-diagnosis purposes.

2 Methodology

We use the 26 crowdsourced queries from the work of Stanton and colleagues [8]. Along with the queries, we extracted the name of the symptoms each queries referred to: queries can be divided in 8 groups which correspond to the 8 different symptoms. We used this symptom information for relevance assessment. The considered queries and symptoms are reported in Table 1.

Two large, commercial search engines (Google and Bing) were used as representative of current state-of-the-art search engines; these search engines were used to retrieve the top-10 results in answer to each of the 26 queries. Queries were issued against the (deprecated) Google Ajax API and the Microsoft Azure Marketplace API from Australia on the same day. The URL of the returned top 10 results were recorded.

A purposely customised version of the Relevation! assessment tool [5] was used to carry out the relevance assessment exercise. Eight higher degree students and researchers from Queensland University of Technology were employed to assess the relevance of the retrieved results. The assessors were not medical experts: this was deliberate to realistically simulate the situation of people with little or no medical knowledge searching for health information on the Web, similar to the actual task we investigate. Web pages returned for queries belonging to the same symptom were shown to a single assessor. Assessors were instructed to evaluate whether each webpage provided relevant information that would allow the information seeker to self-diagnose, i.e., individuate the correct medical term of the symptom they are experiencing. Assessors could assign

	ndcg@1		ndcg@5		ndcg@10		P@5		P@10	
System	Rel	Hrel	Rel	Hrel	Rel	Hrel	Rel	Hrel	Rel	Hrel
Bing	.3846	.2308	.3812	.2654	.3802	.2764	.4385	.2769	.4308	.2769
Google	.3846	.3077	.4242	.3142	.4252	.3138	.5000	.3154	.4923	.3115

Table 2. Retrieval effectiveness achieved by two widely used commercial search engines when prompted with circumlocutory medical queries aimed at self-diagnosis purposes. Results are averaged over 26 queries

one of the following relevance label to each result: Not relevant (assigned to 226 documents), On topic but unreliable (assigned to 54 documents), Somewhat relevant (assigned to 87 documents) and Highly relevant (assigned to 153 documents). Queries, webpage URLs and relevance assessments are made available at https://github.com/ielab/ecir2015-DignoseThisIfYouCan.

To evaluate the effectiveness of two search engines we consider precision at ranks 5 and 10 (P@5, P@10), which indicates the proportion of relevant documents among the top 5 (10) search results, and nDCG at 1, 5 and 10 (ndcg@1, ndcg@5, ndcg@10), which indicates the usefulness, or gain, of the document ranking based on the position of relevant documents in the result list.

3 Results and Analysis

Table 2 reports the effectiveness of the two commercial search engines. We distinguish between Somewhat relevant (Rel) and Highly relevant only documents (Hrel only) (see below for an analysis of these two relevance categories). The results reveal differences in effectiveness between the two search engines (in particular beyond rank 1). Similarly, Figure 1 reports the effectiveness of the systems at a query level, showing that differences are not due to the contribution of outliers, e.g., a single query where one system was particular good or bad. More importantly though, the results highlight that, on average, only about 4 to 5 out of the first 10 results provide information that can help people self-diagnose themselves. This reduces to 3 out of the first 10 documents if highly relevant information is sought.

An analysis of documents assessed as "Somewhat relevant" reveals that a prototypic somewhat relevant document contained information that was not focused on only the relevant symptom, e.g., it provided a list of symptoms with corresponding definition that included the relevant symptom. A similar analysis revealed that documents assessed as highly relevant instead contained information that was mostly solely focused on the relevant symptom, providing descriptions and causes of the symptoms, often aided by photographic material showing visual examples of symptoms occurrences. Pages that were deemed as on topic but unreliable were considered irrelevant for the purpose of this evaluation. These pages contained information that was somewhat related to the sought symptoms, but it was of suspicious origin and often involved the pur-

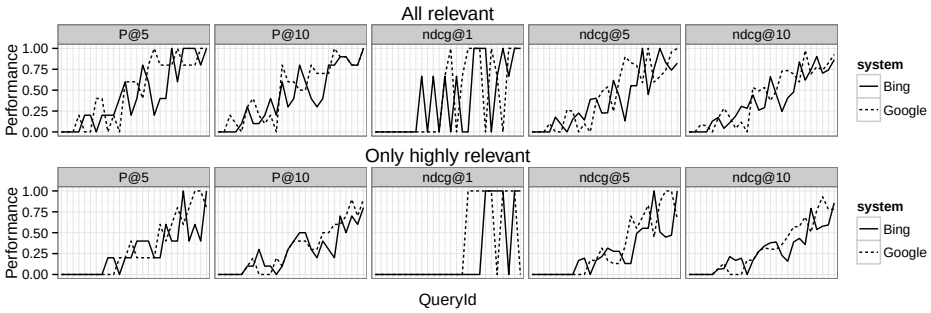

Fig. 1. Retrieval effectiveness of the two studied search engines for each individual query; results are reported for different level of relevance.

chase of a service or a product (for example, selling anti hair loss shampoos for alopecia or glasses for jaundice).

Both search engines retrieved documents that were judged irrelevant by the assessors. A large number of irrelevant documents did contain the query terms but were suggesting a different medical symptom than that underlying the issued query. Other irrelevant documents instead did not related to the medical intent of the query (for example the Amazon page selling copies of "Yellow Eyes" by R. G. Montgomery for the query [yellow eyes] but referring to the jaundice symptom) or related to health problems not in human beings (for example a page about cat bald spot diagnosis for the query [baldness in multiple spots]).

The results obtained in this initial investigation suggest that people searching the Web for information for self-diagnosis is likely to encounter misleading advice that could confuse them or, ultimately, cause harm.

4 Conclusion

Previous research has considered the development and evaluation of techniques to support health information seeking; recent efforts have mostly focused on the problem of searching for information that describes or explains a specialistic medical term and effective information retrieval methods have been developed for this task [3,4].

In this paper we have investigated the effectiveness of current state-of-the-art commercial web search engines for retrieving diagnostic information in answer to a different type of health queries: those that describe symptoms in a circumlocutory, colloquial manner, similar to those observed in query logs and likely be issued by people seeking to self-diagnose themselves. The empirical results suggest that current retrieval techniques may be poorly suited to such queries. We advocate for more research be directed towards improving search systems to support such type of queries, as previous research has highlighted that the access to not relevant information can lead to erroneous self-diagnosis and self-treatment and ultimately to possible harm [6,9].

The evaluation reported in this study presents a number of limitations. Firstly, only a small amount of queries were considered in the empirical experiments; nevertheless, the queries refer to common symptoms and are thus likely to appear in search activities. Secondly, the evaluation considered an ad hoc scenario, where only one query was considered while it is likely that health-related queries are part of more complex search sessions [7] and thus the effectiveness of the sessions, rather than the single queries, should also be accounted for. Finally, we did not *fully* consider the factors that come into play when information seekers consider the relevance of the documents: for health information seeking in particular, it has been shown how the reliability and understandability of the retrieved information is critical to determine its utility and these should be accounted for in the evaluation [11].

Acknowledgements: Guido Zuccon is supported by a QUT ECARD grant, and João Palotti is supported by the EU Project FP7/2007-2013 under grant agreement n°257528 (KHRESMOI) and by the FWF project I1094-N23 (MUCKE). The experiments were ran on hardware funded through QUT SEF Large Equipment Grant 94. The authors would like to thank the assessors that took part to this study for their time.

References

1. Mike Benigeri and Pierre Pluye. Shortcomings of health information on the internet. *Health promotion international*, 18(4):381–386, 2003.
2. Susannah Fox. *Health topics: 80% of internet users look for health information online*. Pew Internet & American Life Project, 2011.
3. Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salanterä, Hanna Suominen, and Guido Zuccon. Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *Proc. of CLEF 2013*, 2013.
4. Lorraine Goeuriot, Liadh Kelly, Wei Li, Joao Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth JF Jones, and Henning Mueller. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In *Proc. of CLEF 2014*, 2014.
5. Bevan Koopman and Guido Zuccon. Relevation!: An open source system for information retrieval relevance assessment. In *Proc. of SIGIR 2014*, 2014.
6. Angela Ryan and Sue Wilson. Internet healthcare: do self-diagnosis sites do more harm than good? *Expert Opinion on Drug Safety*, 7(3):227–229, 2008.
7. Amanda Spink, Yin Yang, Jim Jansen, Pirrko Nykanen, Daniel P Lorence, Seda Ozmutlu, and H Cenk Ozmutlu. A study of medical and health queries to web search engines. *Health Information & Libraries Journal*, 21(1):44–51, 2004.
8. Isabelle Stanton, Samuel Ieong, and Nina Mishra. Circumlocution in diagnostic medical queries. In *Proc. of SIGIR '14*, pages 133–142, 2014.
9. Ryen W White and Eric Horvitz. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM TOIS*, 27(4):23, 2009.
10. Ryen W White and Eric Horvitz. Experiences with web search on medical concerns and self diagnosis. In *Proc. of AMIA*, volume 2009, page 696, 2009.
11. Guido Zuccon and Bevan Koopman. Integrating understandability in the evaluation of consumer health search engines. *Proc. of MedIR 2014*, page 29, 2014.