What to Trust When Searching for Health-Related Symptoms on Google

Author: Nina Sachdev Faculty Advisor: Professor Eni Mustafaraj
Department of Computer Science, Wellesley College, Wellesley MA, 02481

1. Introduction

- Previous research explored use of online searching as a diagnostic engine [13], as a notable number of queries are targeted towards self-diagnosing one's symptoms [9]
- Motivating factors of this project are cyberchondria [13] and online health misinformation [7], which can lead to user anxiety and misinformed health decisions
- In this study, Google search results of medical queries are compared to those of problematic health queries
- Findings will prompt users to pay attention to the types of health websites they click on, the key aspects to look for on search results pages, and how to best formulate health search queries

2. Research Questions

RQ1: Which websites are the most popular in medical and problematic health searches?
RQ2: How do search results differ between medical searches and problematic searches? Are any features of search results an indication of trustworthiness?
RQ3: How should symptom searchers formulate their queries in order to maximize the helpful content their receive?

3. Data and Methods

- 48 legitimate health queries with 642 top results generated
- 31 problematic health queries with 398 top results generated

Legitimate Queries	Problematic Queries		
Extract common medical symptoms from Wikipedia	Extract pseudoscience diagnoses from Wikipedia	Extract alternative health terms from Snopes	Extract recommended search queries from Google

Enter both sets of queries into Google and use **web scraping** to retrieve various features from each search results page

4. Results

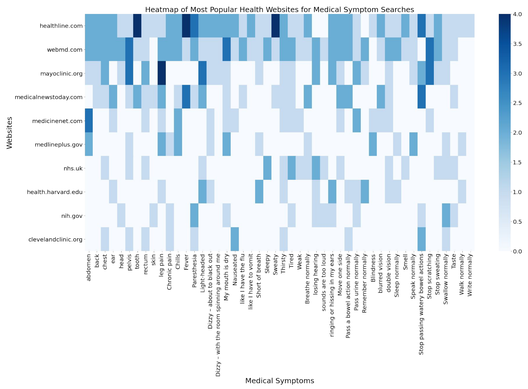

Figure 1: Heatmap shows the density of the most popular health websites across all medical symptom searches. A value of 4.0 (dark blue) corresponds to a website that showed up four times in the top search results for a given symptom. A value of 0.0 (light blue) corresponds to a website that didn't show up in the top search results for a given symptom. There are 48 medical symptoms along the x-axis and 10 top health websites along the y-axis.

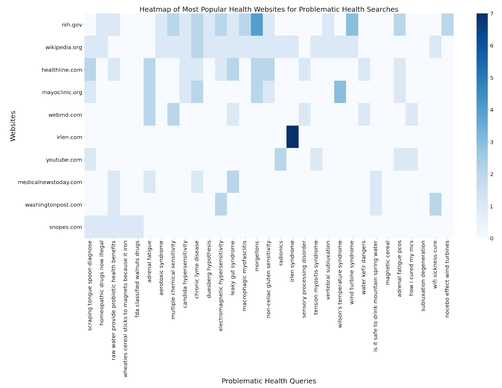

Figure 2: Heatmap shows the density of the most popular websites across all problematic health searches. A value of 7 (dark blue) corresponds to a website that showed up seven times in the top search results for a given health query. A value of 0 (light blue) corresponds to a website that didn't show up in the top search results for a given health query. There are 31 problematic health queries along the x-axis and 10 top websites along the y-axis.

- Most popular websites in medical symptom searches were **Healthline** and **WebMD**, showing up at least once for most queries (Fig. 1)
- **Wikipedia** and **NIH** (and its related subdomains) were the most popular websites in problematic health searches (Fig. 2)

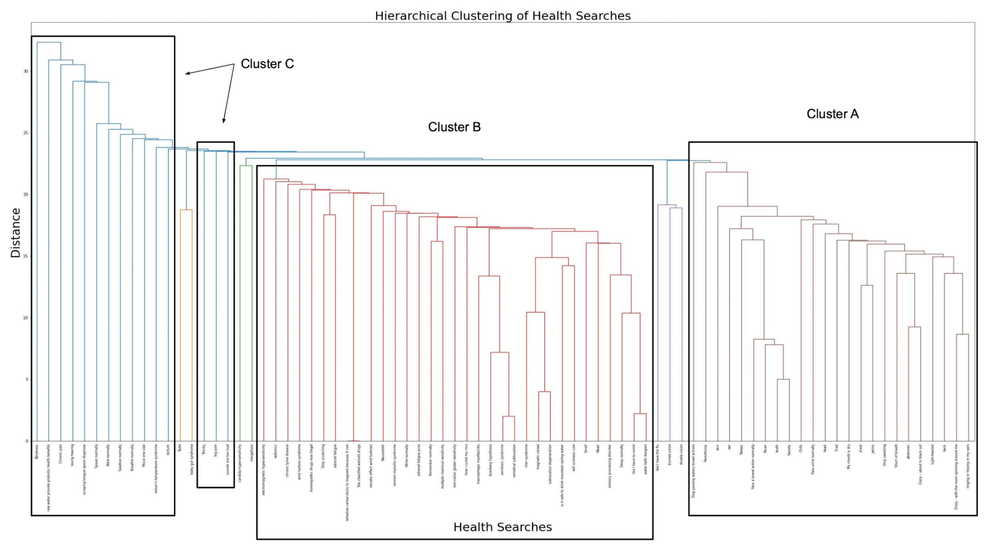

Figure 3: Visualization of agglomerative hierarchical clustering as a dendrogram. Each node at the bottom corresponds to one of the 79 total health searches in Google. The node labels are a shortened version of the query that was used and they are clustered together according to their computed Euclidean distances. Nodes of the same color belong to the same cluster. Clusters from left to right: blue (Cluster C), orange, green, red (Cluster B), purple, and brown (Cluster A).

- Hierarchical clustering of all queries by 16 features: various website rankings, presence of Google knowledge panels, and bolded query text in search results description (Fig. 3)
- **Cluster A:** only medical queries with popular health websites and knowledge panels in search results (Fig. 3)
- **Cluster B:** mostly problematic queries that **lack** many of the same features (i.e. knowledge panel) (Fig. 3)
- **Cluster C:** large distance between nodes reveals limitation in features not being normalized to binary digits (Fig. 3)

5. Conclusion

- Popular health websites are commonplace in legitimate searches and contain easily digestible information (Fig. 1), and are less present in problematic searches (Fig. 2)
- Although NIH and Wikipedia appear more frequently in problematic health searches, they aren't necessarily untrustworthy—their high ranking is in accordance with Google's PageRank algorithm [2]
- Combination of website domains and visual knowledge panels creates starkest contrast between medical searches and problematic health searches (Fig. 3)
- Preliminary findings (see my paper in References) suggest small changes in query formulation can significantly alter search results
- A design suggestion includes flagging potential problematic queries in Google searches
- Limitations include sampling bias against problematic queries and lack of normalized features
- Future work will continue to explore RQ3, understand role of authoritative pages in health searches, and ultimately minimize risk of getting cyberchondria and becoming exposed to health misinformation

6. References

Read my full research paper here

[1] Beeferman, D., and Berger, A. 2000. Agglomerative clustering of a search engine query log. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 407–416.
[2] Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7, 107–117. NLD: Elsevier Science Publishers B. V.
[3] Cartright, M.-A.; White, R. W.; and Horvitz, E. 2011. Intentions and attention in exploratory health search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 65–74.
[4] De Choudhury, M.; Morris, M. R.; and White, R. W. 2014. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1365–1376.
[5] Doherty-Torstrick, E. R.; Walton, K. E.; and Fallon, B. A. 2016. Cyberchondria: parsing health anxiety from online behavior. *Psychosomatics* 57(4):390–400.
[6] Fox, S., and Duggan, M. 2013. Health online 2013. *Health* 2013:1–55.
[7] Ghenai, A., and Mejova, Y. 2018. Fake cures: user-centric modeling of health misinformation in social media. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):1–20.
[8] Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H.; and Gay, G. 2017. Accurately interpreting clickthrough data as implicit feedback. *SIGIR Forum* 51(1):4–11.
[9] Kao, H.-C.; Tang, K.-F.; and Chang, E. Y. 2018. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *AAAI*, 2305–2313.
[10] Laurent, M. R., and Vickers, T. J. 2009. Seeking health information online: does Wikipedia matter? *Journal of the American Medical Informatics Association* 16(4):471–479.
[11] Schoenherr, G. P., and White, R. W. 2014. Interactions between health searchers and search engines. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 143–152.
[12] Sillence, E.; Briggs, P.; Fishwick, L.; and Harris, P. 2004. Trust and mistrust of online health sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 663–670.
[13] White, R. W., and Horvitz, E. 2009. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems (TOIS)* 27(4):1–37.
[14] White, R. W., and Horvitz, E. 2013. Captions and biases in diagnostic search. *ACM Transactions on the Web (TWEB)* 7(4):1–28.
[15] White, R. 2013. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3–12.
[16] Yue, Y.; Patel, R.; and Roehrig, H. 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th International Conference on World Wide Web*, 1011–1018.

7. Acknowledgements

I would like to thank Professor Eni Mustafaraj for her knowledgeable guidance and support throughout this project.