Unsupervised Detection and Promotion of Authoritative Domains for Medical Queries in Web Search

Manoj K. Chinnakotla*
Microsoft, Hyderabad, India
manojc@microsoft.com

Rupesh K. Mehta
Microsoft, Hyderabad, India
rupeshme@microsoft.com

Vipul Agrawal
Microsoft, Bellevue, USA
vipulag@microsoft.com

Abstract

Medical or Health related search queries constitute a significant portion of the total number of queries searched everyday on the web. For health queries, the *authenticity* or *authoritativeness* of search results is of utmost importance besides relevance. So far, research in automatic detection of authoritative sources on the web has mainly focused on - a) link structure based approaches and b) supervised approaches for predicting *trustworthiness*. However, the aforementioned approaches have some inherent limitations. For example, several content farm and low quality sites artificially boost their link-based authority rankings by forming a syndicate of highly interlinked domains and content which is algorithmically hard to detect. Moreover, the number of positively labeled training samples available for learning trustworthiness is also limited when compared to the size of the web.

In this paper, we propose a novel unsupervised approach to *detect* and *promote* authoritative domains in health segment using click-through data. We argue that standard IR metrics such as NDCG are relevance-centric and hence are not suitable for evaluating authority. We propose a new authority-centric evaluation metric based on side-by-side judgment of results. Using real world search query sets, we evaluate our approach both quantitatively and qualitatively and show that it succeeds in significantly improving the authoritativeness of results when compared to a standard web ranking baseline.

1 Introduction

The web is growing at an enormous rate with information spread across billions of web pages. Medical and Health related web sites constitute a significant portion of the web and its growth. The Pew Internet Project, one of the largest national surveys undertaken in the U.S, reveals that 59% of U.S adults have looked online for health information in the past year (Pew , 2013). Moreover, the study states that 35% of the U.S adults were *"Online Diagnosers (OD)"* - people who turn to the internet to figure out which medical condition they have. Interestingly, 41% of ODs had their condition confirmed by the clinician. As per the study, 80% of the people with a health information need start off their inquiry with a web search engine. Not just patients, recent studies (AMA , 2002) show that physicians too rely on the web for their research and studies. In view of its impact on decisions related to people's health, it's highly imperative for search engines to provide information which is not just relevant and accurate but also *authoritative*. We define *authoritativeness* of a search result as follows:

Definition 1. A search result is said to be authoritative for a query if:

- It is widely accepted to be an authentic source of information by experts in the domain
- It is the site of an organization or corporation vested with the right to give first-hand information on the entity or topic

The quality of content available on the web poses a serious challenge to search engines while trying to provide accurate and reliable information. The content quality varies a lot, ranging from shallow content written by amateurs, automatic content generated by engines, plagiarized and spammy content to deep and authentic articles written by domain experts. Besides, many low

[*] Corresponding Author

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 388–394,
Goa, India. December 2014.

Domain	Type of Site	PageRank
ehow.com	Content Farm	64,678
ezinearticles.com	Content Farm	61,566
webmd.com	Authoritative	65,504
mayoclinic.com	Authoritative	63,832

Table 1: Comparison of PageRank values of a few popular content farm sites and authoritative sites.

quality sites employ a variety of Search Engine Optimization (SEO) techniques to trick search engines and artificially boost their rankings. Hence, attempts have been made (Spink et al. , 2004) to aid information seekers by identifying and accrediting web sites which provide high quality, well-researched, reliable and trust-worthy information. For example, ”Health on Net (HON)” logo on web sites helps in identifying reliable health information sources. However, since it is a manual effort, it is not scalable for the web.

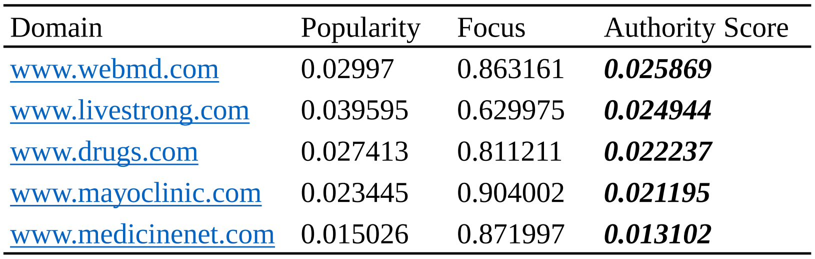
So far, research in automatic detection of authoritative sources on the web has mainly focused on - a) link structure based approaches and b) supervised approaches for predicting trustworthiness. Link based approaches such as PageRank, HITS, SALSA (Sergey and Larry , 1998; Kleinberg , 1999; Lempel and Moran , 2001) analyse the hyperlinking structure of the web graph for identifying the authoritative sources. In PageRank, which is the most popular variant, the notion of authority is similar to the notion of citations in scientific literature. The authority of a page is proportional to the number of incoming hyperlinks and to the authority of the pages which point to it. However, since links are created by content curators, it is easy to manipulate them. Bianchini *et. al.* (Bianchini et al. , 2003) point to one such limitation of link based algorithms - it’s possible to artificially boost the authority scores of pages through the creation of artificial communities which are algorithmically hard to detect. Many low quality information sources, such as content farms, exploit this weakness to gain decent PageRank scores by forming a syndicate of highly interlinked content. Table 1 shows a comparison of PageRank values for a few content farm and authoritative domains. It can be observed that their PageRanks are almost comparable. On the other hand, supervised approaches for learning trustworthiness of domains rely on the availability of gold standard labeled data which is hard to obtain in large quantities.

In this paper, we propose a novel unsupervised approach to automatically detect authoritative domains in the medical query segment using user click logs. We aggregate the click signals at a domain level and assign an authority score which is based on - a) popularity of the domain and b) specificity of content with respect to the query segment. Although click signals are noisy, the advantage of relying on clicks for authority computation is that it is hard to manipulate them, and they provide a user-centric view of authority. In practice, a search engine has to optimize both *relevance* and *authority*. Hence, we fire the initial query to get the top *k* documents and then rerank them based on a combination of relevance and domain authority scores. We also show that the standard NDCG metric is not sensitive to authority and hence can’t be used to measure it. In our current work, we define a measure based on side-by-side authority-centric judging of ranking results and show that our approach improves the authoritativeness of results when compared to a standard web ranking baseline.

2 Related Work

Several researchers have reported the presence of unreliable and low quality information, especially in the context of medical domain, on the web (Matthews et al. , 2003; Tang et al. , 2006; Marriott et al. , 2008).

(A1 et al. , 2007b; A1 et al. , 2007a; Sondhi et al. , 2012; Olteanu et al. , 2013) have employed supervised machine learning techniques to learn the notion of trustworthiness or credibility of web pages. (A1 et al. , 2007b; A1 et al. , 2007a) use the Health On Net (HoN) label data as gold standard and learn a prediction model based on content and URL based features. (Sondhi et al. , 2012) use both content based features and link based features. (Olteanu et al. , 2013) further experimented with social features from popular sources such as Facebook and Twitter and web page design. However, the main problem with supervised methods is the lack of training samples. The number of web sites which apply and refresh their HoN rat-



Domain	Popularity	Focus	Authority Score
www.webmd.com	0.02997	0.863161	***0.025869***
www.livestrong.com	0.039595	0.629975	***0.024944***
www.drugs.com	0.027413	0.811211	***0.022237***
www.mayoclinic.com	0.023445	0.904002	***0.021195***
www.medicinenet.com	0.015026	0.871997	***0.013102***

Figure 1: Top 5 Medical Domains identified through Authority Scores defined in Section 3

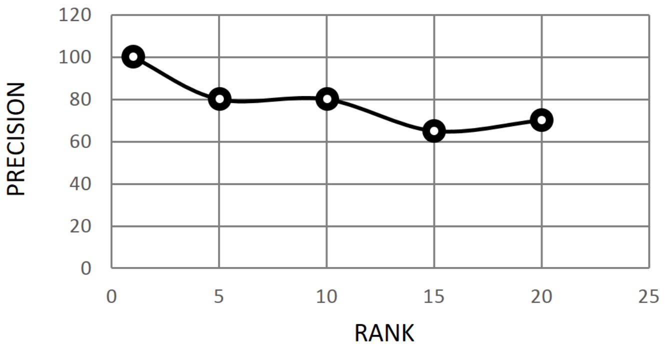


Figure 2: Precision vs. Rank for Authority Scores defined in Section 3

ings regularly is usually very less. Due to this, curating enough number of training samples to learn reasonably accurate models is hard. Our approach differs from the other approaches discussed so far in the following ways:

- We propose an *unsupervised technique* to detect and score segment-specific authoritative domains using click data

- We also propose a re-ranking technique to promote the authoritative domains in the ranking

3 Segment Authority

In this section, we describe our approach to score domains based on their authority. The main intuition behind the technique is that typically content farm and other low-quality sites such as eHow, ezinearticles, create shallow and generic content across a wide variety of query segments. This is mainly done to maximize their revenue through ad monetization. On the other hand, authoritative sites curate high-quality and deep content for a particular segment. There are very few sites which do both of the above with notable exceptions such as wikipedia (i.e. deep content spread across a wide variety of segments). Hence, for each domain, we model this notion of ”generic” vs ”specific” and popularity and combine them as a single score for authority. We define *Focus* and *Popularity* of a domain *d* with respect to a query segment *seg* as follows:

Definition 2. Focus of a domain d with respect to query segment *seg*, is defined as the probability of a user choosing a query from segment *seg* when there is a click on the domain *d*.

$$Focus(d, seg) = Pr(seg|d) \tag{1}$$

Definition 3. Popularity of a domain d within the query segment *seg*, is defined as the probability of a user clicking on the domain *d* when the user query belongs to the segment *seg*.

$$Popularity(d, seg) = Pr(d|seg) \tag{2}$$

Given the click logs, we first run query classifiers (Cao et al. , 2009) to classify each query into segments such as health, movies, sports, technology, finance, *etc.* Later, the probability scores mentioned above are computed as follows:

$$Score(seg|d) = \frac{\text{No. of queries on d where seg classifier is ON}}{\text{Total no. of queries on d}} \tag{3}$$

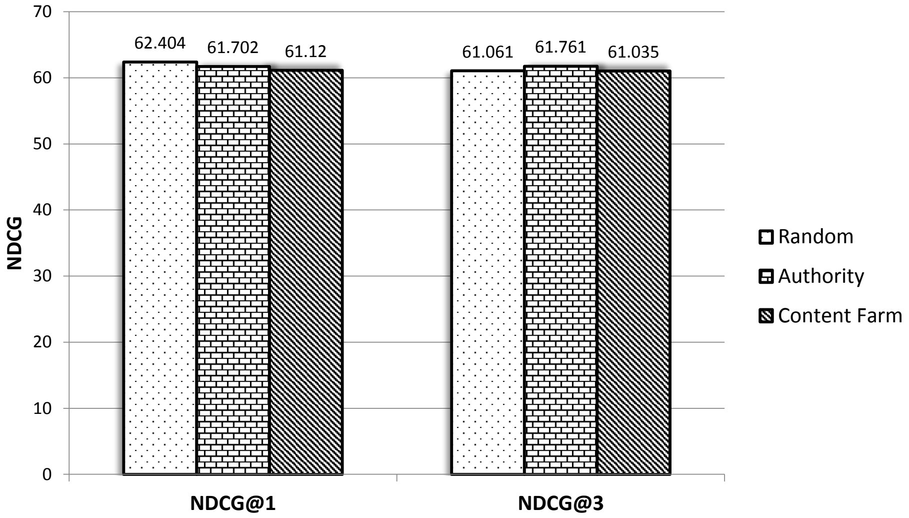


Figure 3: Sensitivity of NDCG Metric to Authority

The above score is unnormalized since the query classifiers were independently trained. The normalized probability is given below:

$$Pr(seg|d) = \frac{Score(seg|d)}{\Sigma_{seg} Score(seg|d)} \quad (4)$$

$$Pr(d|seg) = \frac{Pr(seg|d) \times Pr(d)}{\Sigma_{seg} Pr(seg|d) \times Pr(d)} \quad (5)$$

$$Pr(d) = \frac{\text{No. of queries on d}}{\text{Total no. of queries}} \quad (6)$$

$$Authority(d) = Pr(seg|d) \times Pr(d|seg) \quad (7)$$

Figure 1 shows the top five authoritative domains, when sorted using the authority score, for the medical segment. These scores were generated using three months of click logs from the commercial search engine Microsoft Bing. We perform a simple validation of our authority scores by checking how many of the top ranked domains have valid HON certification. The results are shown in Figure 2. We can observe that the automatically mined top domains correlate well with the HON labels given by medical experts.

4 Authority Based Reranking of Results

Although, both relevance and authority are both important, Jones *et. al.* (Jones et al. , 2011) show that users prefer relevant information from a spammy domain than irrelevant results. We combine our domain authority score with relevance score of web ranker to promoting both relevant and authoritative pages.

We take the top *'k'* results from the initial retrieval and re-rank them using a function designed to factor in both relevance and authority. Let $U = \{u_1, u_2, \ldots, u_k\}$ be the set of top *'k'* URLs retrieved by the initial ranker. Let $S = \{s_1, s_2, \ldots, s_k\}$ be their corresponding scores assigned by the ranker. Let $D = \{d_1, d_2, \ldots, d_k\}$ be the corresponding domains of these URLs. The new scoring function for re-ranking is:

$$Score(u_i, s_i, q) = s_i \times (1 + \alpha \times \text{Authority}(d_i)) \quad (8)$$

The above function ensures that sites with good relevance and authority scores are boosted higher in the ranking. The above function also ensures that irrelevant results from authoritative domains do not trump a relevant result from a less authoritative source by adjustiing the value of α.

5 Evaluation Metrics for Authority

Standard web ranking metrics such as NDCG (Järvelin and Kekäläinen , 2000) are measures of *graded relevance*. The guidelines which are typically used to grade the web result sets is based purely on relevance. Hence, it can't be directly used to measure authority. To prove this point, from a standard evaluation set for web (size around 13K, where judgments are available), we sampled three equal sized query sets of size 100 with the following variations - a) random queries b) queries with authoritative URLs appearing in top 3 c) queries with content farm sites appearing

Query Set	No. of Queries /Judgment Queries	Surplus$_{\text{strong}}$ (W/L/T)	Surplus$_{\text{weak}}$ (W/L/T)
HealthQuery Set	462/181	+5.52 (24/14/143)	***+14.36***** ***(88/62/31)***
HealthTestSet-1K	1K/1K	+1.2 (41/29/930)	***+6.9***** ***(264/195/541)***

Table 2: Results comparing the performance of Authority Based Reranking with Baseline web ranker on two query sets. Results marked as ** indicate that the surplus was found to be statistically significant over the baseline at 95% confidence level ($\alpha = 0.05$). W/L/T denote the number of Wins, Losses and Ties observed.

in top 3 (b and c are identified manually). We computed NDCG@1 and NDCG@3 on top of these three sets and the results are shown in Figure 3. One can notice that NDCG@1 and NDCG@3 remain almost the same for all three query sets, and hence they are not sensitive to the notion of authority. In view of the above, we define a new metric called "Surplus" which is based on relative comparison of baseline and treatment rankers with respect to their relevance and authoritativeness. The procedure is as follows:

- We show the first page (top 10) of baseline and treatment results to a judge in two separate tabs in a single window.
- For each query, the baseline and treatment results are randomly placed in the left (L) and right (R) tabs of window to avoid any identification and biased judgments.
- Each judge is given detailed guidelines on how to identify authoritative sites[1].
- The ratings are given on a seven-point scale - a) Left Much Better b) Left Better c) Left Slightly Better d) Neutral e) Right Slightly Better f) Right Better g) Right Much Better.
- A technique scores a *Strong Win* for a query if it gets a Better or Much Better rating with respect to the baseline. It scores a *Weak Win* if it gets Slightly Better, Better or Much Better ratings with respect to baseline and a *Tie* if it gets a Neutral rating. *Strong Loss* and *Weak Loss* are similarly defined.

A pool of 25 judges were hired and trained (using the guidelines) specifically for providing these judgments.

Given a query set with n queries, if a technique scores n_W wins, n_L losses and n_T ties, the surplus of a technique is defined as follows:

$$Surplus = \frac{n_W - n_L}{n_W + n_L + n_T} \times 100 \qquad (9)$$

The final metric used for measurement is $Surplus_{strong}$, where strong win/losses are used, and $Surplus_{weak}$ where weak win/losses are used. We also check for the statistical significance of the surplus to ensure more robustness of the metric. A good surplus on a large query set implies that the technique is performing well with respect to the baseline. As discussed in Section 4, a technique has to also ensure that it does not hurt relevance - it should not decrease the NDCG metric significantly.

6 Experimental Setup

We evaluate the performance of our system on a real world dataset from Microsoft Bing search engine. The dataset consists of around 13,711 English queries sampled from one year query logs. From these, we apply query classifiers and filter out only *"health"* segment queries which are around 462. We call this query set as *HealthQuerySet*. For these queries, we also have human generated relevance ratings available on a five-point scale (0-4) where 4 means most relevant and 0 means irrelevant. This judgement data is useful for computing the NDCG metric. We train a LambdaMART (Burges , 2010) baseline

[1] http://bit.ly/1uOJeVw

Query	Top 3 Results in Baseline	Top 3 Results in Authority Ranker	Description
flaxseed oil health benefits ??? ***(Strong Win)***	1. www.ehow.com/about_4587396_health-benefits-flaxseed-oil.htm 2. www.webmd.com/diet/features/benefits-of-flaxseed 3. www.livestrong.com/article/112063-benefits-flaxseed-oil-capsules/	1. www.webmd.com/diet/features/benefits-of-flaxseed 2. www.livestrong.com/article/112063-benefits-flaxseed-oil-capsules/ 3. www.ehow.com/about_4587396_health-benefits-flaxseed-oil.htm	*"webmd.com"* and *"livestrong.com"* are much more authoritative sites than "*ehow.com*" which is a content farm site.
high calcium symptoms ***(Weak Win)***	1. www.ehow.com/list_6197078_signs-high-calcium-levels-blood.html 2. www.mayoclinic.org/diseases-conditions/hypercalcemia/basics/symptoms/CON-20031513 3. www.emedicinehealth.com/hypercalcemia_elevated_calcium_levels/page3_em.htm	1. www.mayoclinic.org/diseases-conditions/hypercalcemia/basics/symptoms/CON-20031513 2. www.ehow.com/list_6197078_signs-high-calcium-levels-blood.html 3. www.emedicinehealth.com/hypercalcemia_elevated_calcium_levels/page3_em.htm	*"mayoclinic.com"* is much more authoritative than *"ehow.com"* which is a content farm site.
infected sebaceous cyst ***(Strong Loss)***	1. mackinven.com/sebaceous-cyst-treatment-how-to-treat-an-infected-sebaceous-cyst-at-home/ 2. en.wikipedia.org/wiki/Sebaceous_cyst 3. www.webmd.com/skin-problems-and-treatments/guide/epidermoid-sebaceous-cysts	1. www.webmd.com/skin-problems-and-treatments/guide/epidermoid-sebaceous-cysts 2. en.wikipedia.org/wiki/Sebaceous_cyst 3. mackinven.com/sebaceous-cyst-treatment-how-to-treat-an-infected-sebaceous-cyst-at-home/	Result #1 from *"mackinven.com"* gives the most relevant page. The pages from *"webmd"* and *"wikipedia"* are not relevant.

Table 3: Qualitative comparison of Authority Based Reranking and Baseline through a few representative queries from the query sets.

with standard text based features and document level features defined in the LETOR dataset (Qin et al. , 2010). We use this as the baseline, which does not have any authority oriented features, for comparing with our approach. The baseline was trained on a separate set of 12,124 queries for which judgments (for URLs per query) were available on a five-point scale. We tune the value of α in Equation 8 such that the number of wins is maximised. In order to show that our technique performs well on completely unseen queries as well, we sample a new test set of 1000 queries, called *HealthTestSet-1K*, from only health segment and show our performance on this. While submitting queries for judgments, we only consider queries where both the technique and the baseline rankings differ in some way in the top five positions. This is mainly to save judgement cost and does not have any impact on the evaluation. However, in the HealthTestSet-1K query set, we remove this constraint as well and submit the entire 1K subset for judgments to enable much clearer and straightforward interpretation.

7 Results and Discussion

As mentioned in Section 6, we tune the value of α in Equation 8 using the *HealthQuerySet* such that the number of wins is maximised. The optimal value of α was found to be 0.6.

The results of our technique with respect to the baseline is shown in Table 2. The results show that Authority Based Reranking technique shows significant gains in weak surplus over the baseline web ranker. Since our technique does not lead to any relevance improvements, similar surplus gains were not observed for the strong surplus metric. We also noticed that the NDCG@3 for the baseline ranker was 51.57 whereas it was 52.78 for the treatment authority ranker. This shows that the technique leads to improvements in authority while minimally affecting NDCG@3 which is indicative of relevance. Moreover, the technique also scores significant improvements on the unseen *HealthTestSet-1K* query set.

Table 3 illustrates the qualitative improvement achieved by our technique through some actual examples from the dataset. The first two query examples show the cases where the technique scored wins by promoting authoritative sites which are equally relevant in the top 3. The last case shows a loss where due to the promotion of authoritative site, the relevance was hurt. Since, our technique just does a interpolation of relevance score and domain authority score, this is sometimes bound to happen.

8 Conclusion and Future Work

We proposed a novel unsupervised approach to automatically detect authoritative medical domains using user click logs. We also proposed a technique, which making use of the domain authority scores, reranks the top *'k'* search results from the initial retrieval and promotes the results from top authoritative domains. We argued and experimentally showed that standard web IR metrics such as NDCG are not suitable for measuring authority. Hence, we propose a new authority-centric metric which is based on side-by-side judging of results. Through experiments on different query sets sampled from real web query logs, we showed that our proposed technique significantly improves the *authoritativeness* of results over a standard web ranker baseline. As part of future work, we plan to - a) further refine the concept of authority at a topic-level within each segment and b) come up with a notion of query dependent authority.

References

[A1 et al. 2007a] Gaudinat A1, Grabar N, and Boyer C. 2007a. Automatic Retrieval of Web Pages with Standards of Ethics and Trustworthiness within a Medical Portal: What a Page Name Tell Us. In *Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) AIME 2007. LNCS (LNAI)*, pages 185–189.

[A1 et al. 2007b] Gaudinat A1, Grabar N, and Boyer C. 2007b. Machine Learning Approach for Automatic Quality Criteria Detection of Health Web Pages. In *Proc. of the World Congress on Health (Medical) Informatics Building Sustainable Health Systems*, pages 705–709.

[AMA 2002] AMA. 2002. Study of physicians' use of the world wide web. *American Medical Association*.

[Bianchini et al. 2003] M. Bianchini, M. Gori, and F. Scarselli. 2003. Pagerank and Web Communities. In *IEEE International Conference on Web Intelligence*, pages 365–371.

[Burges 2010] Christopher J. C. Burges. 2010. From RankNet to LambdaRank to LambdaMART: An Overview. Technical report, Microsoft Research.

[Cao et al. 2009] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang. 2009. Context-Aware Query Classification. In *SIGIR '09*, pages 3–10, New York, NY, USA. ACM.

[Järvelin and Kekäläinen 2000] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *SIGIR'00*, pages 41–48, New York, NY, USA. ACM.

[Jones et al. 2011] Timothy Jones, David Hawking, Paul Thomas, and Ramesh Sankaranarayana. 2011. Relative Effect of Spam and Irrelevant Documents on User Interaction with Search Engines. In *CIKM '11*, pages 2113–2116, New York, NY, USA. ACM.

[Kleinberg 1999] Jon M. Kleinberg. 1999. Authoritative Sources in a Hyperlinked Environment. *J. ACM*, 46(5):604–632, September.

[Lempel and Moran 2001] R. Lempel and S. Moran. 2001. Salsa: The Stochastic Approach for Link-Structure Analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, April.

[Pew 2013] The Pew Internet and American Life. Health Online 2013. `http://www.pewinternet.org/2013/01/15/health-online-2013/`.

[Marriott et al. 2008] JV1 Marriott, Stec P, El-Toukhy T, Khalaf Y, Khalaf P, and Coomarasamy A. 2008. Infertility Information on the World Wide Web: A Cross-Sectional Survey of Quality of Infertility Information on the Internet in the UK. *Human Reproduction*, pages 1520–1525.

[Matthews et al. 2003] "Scott C. Matthews, Alvaro Camacho, Paul J. Mills, and Joel E. Dimsdale". "2003". "The Internet for Medical Information about Cancer: Help or Hindrance? ". *"Psychosomatics "*, "44"("2"):"100 – 103".

[Olteanu et al. 2013] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. 2013. Web Credibility: Features Exploration and Credibility Prediction. In *ECIR '13*, pages 557–568, Berlin, Heidelberg. Springer-Verlag.

[Qin et al. 2010] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. 2010. Letor: A Benchmark Collection for Research on Learning to Rank for Information Retrieval. *Inf. Retr.*, 13(4):346–374.

[Sergey and Larry 1998] Brin Sergey and Page Larry. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Computer Networks and ISDN Systems*, pages 107–117.

[Sondhi et al. 2012] Parikshit Sondhi, V. G. Vinod Vydiswaran, and Cheng Xiang Zhai. 2012. Reliability Prediction of Webpages in the Medical Domain. In *ECIR '12*, pages 219–231, Berlin, Heidelberg. Springer-Verlag.

[Spink et al. 2004] Amanda Spink, Yin Yang, Jim Jansen, Pirrko Nykanen, Daniel P. Lorence, Seda Ozmutlu, and H. Cenk Ozmutlu. 2004. A Study of Medical and Health Queries to Web Search Engines. *Health Information and Libraries Journal*, 21(1):44–51.

[Tang et al. 2006] ThanhTin Tang, Nick Craswell, David Hawking, Kathy Griffiths, and Helen Christensen. 2006. Quality and Relevance of Domain-Specific Search: A Case Study in Mental Health. *Information Retrieval*, 9(2):207–225.