

# Efficiently Scoring the Health-relatedness of Web Pages

---

1st International Workshop on Open Web Search 2024  
28th March 2024, Glasgow, Scotland



**Ferdinand  
Schlatt**



[webis.de](http://webis.de)

# Efficiently Scoring the Health-relatedness of Web Pages

## Motivation

About 70% of people with internet access use search engines as their primary source for medical information [Bondarenko et al., CIKM'21]

# Efficiently Scoring the Health-relatedness of Web Pages

## Motivation

About 70% of people with internet access use search engines as their primary source for medical information [Bondarenko et al., CIKM'21]



<https://stablediffusionweb.com/app/image-generator>

# Efficiently Scoring the Health-relatedness of Web Pages

## Motivation

About 70% of people with internet access use search engines as their primary source for medical information [Bondarenko et al., CIKM'21]

*However:* A substantial portion of the provided "answers" are incorrect

# Efficiently Scoring the Health-relatedness of Web Pages

## Motivation

About 70% of people with internet access use search engines as their primary source for medical information [Bondarenko et al., CIKM'21]

*However:* A substantial portion of the provided "answers" are incorrect

The image shows a Google search interface. The search bar contains the text "Does garlic help with thrush?". Below the search bar, there are tabs for "All", "Images", "Videos", "News", "Books", and "More". The search results are as follows:

- Healthline** (https://www.healthline.com > health > garlic-for-yeast-i...):  
Title: [Garlic for Yeast Infection: Treatment & Side Effects](#)  
Date: 26 Jul 2018 — Studies are inconclusive as to whether garlic, garlic tablets, or garlic extract can provide a cure for yeast infections. However, its chemical ...  
Status: Question mark icon (?)
- Scientific American** (https://www.scientificamerican.com > article > fact-or-fi...):  
Title: [A Clove of Garlic Can Stop a Vaginal Yeast Infection](#)  
Date: 3 Oct 2014 — They did a small study looking at what oral garlic does to the growth of yeast in the vagina and they found that there was no impact. That was a ...  
Status: Red X icon (✗)
- The Candida Diet** (https://www.thecandidadiet.com > garlic):  
Title: [How To Use Garlic For Candida and Yeast Infections](#)  
Date: 27 Jan 2022 — Numerous research studies have shown garlic to be an effective treatment for yeast and fungal infections, especially Candida overgrowth. Not ...  
Status: Green checkmark icon (✓)

# Efficiently Scoring the Health-relatedness of Web Pages

## Motivation

About 70% of people with internet access use search engines as their primary source for medical information [Bondarenko et al., CIKM'21]

*However:* A substantial portion of the provided "answers" are incorrect

The screenshot shows a Google search interface with the query "Does garlic help with thrush?". Below the search bar, there are three search results, each with a status icon in a box to its right:

- Healthline**: "Garlic for Yeast Infection: Treatment & Side Effects" (26 Jul 2018). Status icon: Question mark (grey).
- Scientific American**: "A Clove of Garlic Can Stop a Vaginal Yeast Infection" (3 Oct 2014). Status icon: Red X (red).
- The Candida Diet**: "How To Use Garlic For Candida and Yeast Infections" (27 Jan 2022). Status icon: Green checkmark (green).

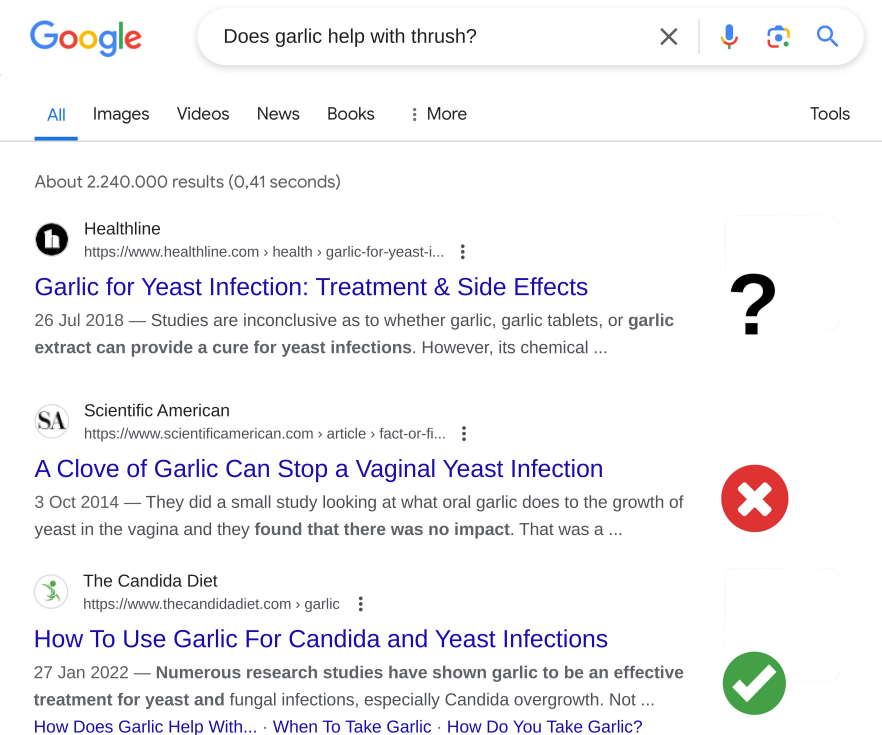
Health-related queries warrant special care

# Efficiently Scoring the Health-relatedness of Web Pages

## Motivation

About 70% of people with internet access use search engines as their primary source for medical information [Bondarenko et al., CIKM'21]

*However:* A substantial portion of the provided "answers" are incorrect



Health-related queries warrant special care

→ Building a search engine for health information requires filtering health-related queries and documents

# Efficiently Scoring the Health-relatedness of Web Pages

## Termhood Scores

In previous work, we developed an approach based on termhood scores to efficiently identify health-related phrases / terms [Schlatt et al., COLING'22]



# Efficiently Scoring the Health-relatedness of Web Pages

## Termhood Scores

In previous work, we developed an approach based on termhood scores to efficiently identify health-related phrases / terms [Schlatt et al., COLING'22]

Termhood scores determine the degree to which a phrase is specific for a domain using the term frequencies of an in-domain and a contrastive corpus



Phrase	Health-related Corpus	Contrastive Corpus
actor	5,590	<b>539,180</b>
carcinoma	<b>987,164</b>	7,410
diagnosis	<b>1,851,514</b>	34,218
study	<b>10,630,098</b>	508,740
the	<b>200,926,211</b>	196,374,618
ward	47,099	<b>186,811</b>

# Efficiently Scoring the Health-relatedness of Web Pages

## Discriminative Weight

We use the discriminative weight (DW) which combines [Wong et al., AusDM'07]

1. a corpus-oriented  $tf \cdot idf$  measure
2. a domain-specificity measure

# Efficiently Scoring the Health-relatedness of Web Pages

## Discriminative Weight

We use the discriminative weight (DW) which combines [Wong et al., AusDM'07]

1. a corpus-oriented  $tf \cdot idf$  measure
2. a domain-specificity measure

with two different in-domain corpora

1. PubMed (medical / scientific language)
2. consumer-oriented online encyclopedias (layperson language)

# Efficiently Scoring the Health-relatedness of Web Pages

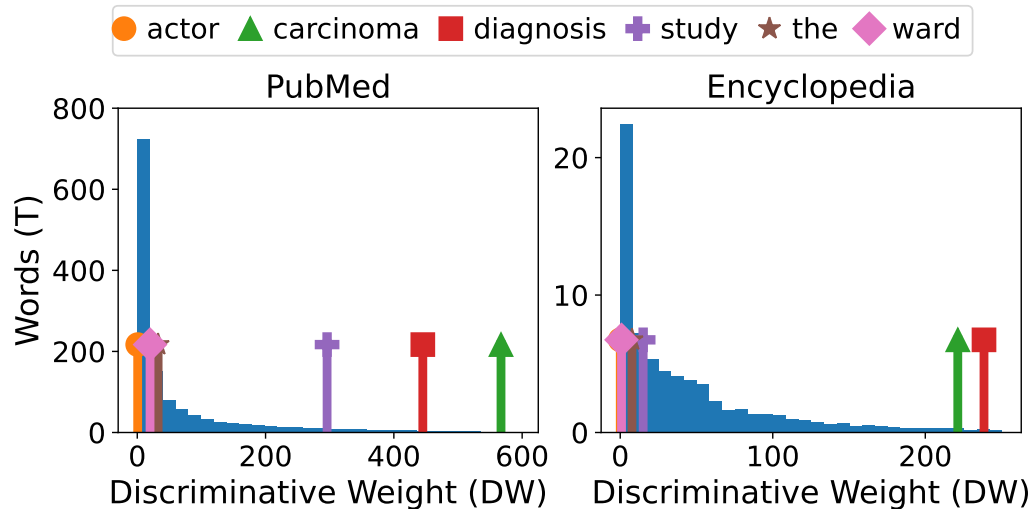
## Discriminative Weight

We use the discriminative weight (DW) which combines [Wong et al., AusDM'07]

1. a corpus-oriented  $tf \cdot idf$  measure
2. a domain-specificity measure

with two different in-domain corpora

1. PubMed (medical / scientific language)
2. consumer-oriented online encyclopedias (layperson language)



# Efficiently Scoring the Health-relatedness of Web Pages

## Filtering Sentences

We applied the discriminative weight to classify if a sentence is health-related

<b>Approach</b>	<b>P</b>	<b>R</b>	<b>F1</b>	<b>M</b>
cTakes	0.57	0.46	0.51	0.42
ScispaCy	0.42	0.60	0.49	0.37
BERT	0.76	0.74	<b>0.75</b>	<b>0.70</b>
PubMedBERT	<b>0.87</b>	0.57	0.69	0.66
DW	0.71	<b>0.77</b>	0.74	0.68

# Efficiently Scoring the Health-relatedness of Web Pages

## Filtering Sentences

We applied the discriminative weight to classify if a sentence is health-related

<b>Approach</b>	<b>P</b>	<b>R</b>	<b>F1</b>	<b>M</b>
cTakes	0.57	0.46	0.51	0.42
ScispaCy	0.42	0.60	0.49	0.37
BERT	0.76	0.74	<b>0.75</b>	<b>0.70</b>
PubMedBERT	<b>0.87</b>	0.57	0.69	0.66
DW	0.71	<b>0.77</b>	0.74	0.68

→ DW more effective than entity linkers and (almost) as effective as BERT

# Efficiently Scoring the Health-relatedness of Web Pages

## Filtering Sentences

We applied the discriminative weight to classify if a sentence is health-related

Approach	P	R	F1	M
cTakes	0.57	0.46	0.51	0.42
ScispaCy	0.42	0.60	0.49	0.37
BERT	0.76	0.74	<b>0.75</b>	<b>0.70</b>
PubMedBERT	<b>0.87</b>	0.57	0.69	0.66
DW	0.71	<b>0.77</b>	0.74	0.68

→ DW more effective than entity linkers and (almost) as effective as BERT

Approach	ms	Speedup
cTakes	212.12	0.2
ScispaCy	15.96	3.0
(PubMed) BERT	47.77	1.0
CW / TDS / DW	<b>1.02</b>	<b>46.8</b>

# Efficiently Scoring the Health-relatedness of Web Pages

## Filtering Sentences

We applied the discriminative weight to classify if a sentence is health-related

Approach	P	R	F1	M
cTakes	0.57	0.46	0.51	0.42
ScispaCy	0.42	0.60	0.49	0.37
BERT	0.76	0.74	<b>0.75</b>	<b>0.70</b>
PubMedBERT	<b>0.87</b>	0.57	0.69	0.66
DW	0.71	<b>0.77</b>	0.74	0.68

→ DW more effective than entity linkers and (almost) as effective as BERT

Approach	ms	Speedup
cTakes	212.12	0.2
ScispaCy	15.96	3.0
(PubMed) BERT	47.77	1.0
CW / TDS / DW	<b>1.02</b>	<b>46.8</b>

→ DW is substantially faster than entity linkers and BERT



# Efficiently Scoring the Health-relatedness of Web Pages

## Pilot Study: Filtering Queries

We can also apply DW to classify if a query is health-related

# Efficiently Scoring the Health-relatedness of Web Pages

## Pilot Study: Filtering Queries

We can also apply DW to classify if a query is health-related

We applied our software submission to the TREC Web Tracks from 2009–2012

[Clarke et al., TREC'09–'12]

<b>Text</b>	<b>Rank</b>	<b>DW</b>
sore throat	1	411.7
forearm pain	2	354.0
joints	3	331.5
...		
lower heart rate	15	168.4
ct jobs	16	167.2
...		
angular cheilitis	33	76.90
getting organized	34	74.77
...		

# Efficiently Scoring the Health-relatedness of Web Pages

## Pilot Study: Filtering Queries

We can also apply DW to classify if a query is health-related

We applied our software submission to the TREC Web Tracks from 2009–2012

[Clarke et al., TREC'09–'12]

<b>Text</b>	<b>Rank</b>	<b>DW</b>
sore throat	1	411.7
forearm pain	2	354.0
joints	3	331.5
...		
lower heart rate	15	168.4
ct jobs	16	167.2
...		
angular cheilitis	33	76.90
getting organized	34	74.77
...		

nDCG@10 for health-related vs non-health-related queries

<b>Model</b>	<b>HQ</b>	<b>OQ</b>
Dirichlet	0.27	0.25
BM25	0.24	0.19
MonoT5 3b	0.18	0.20
Splade	0.17	0.19

# Efficiently Scoring the Health-relatedness of Web Pages

## Pilot Study: Filtering Documents

We can also apply DW to classify if a document is health-related

# Efficiently Scoring the Health-relatedness of Web Pages

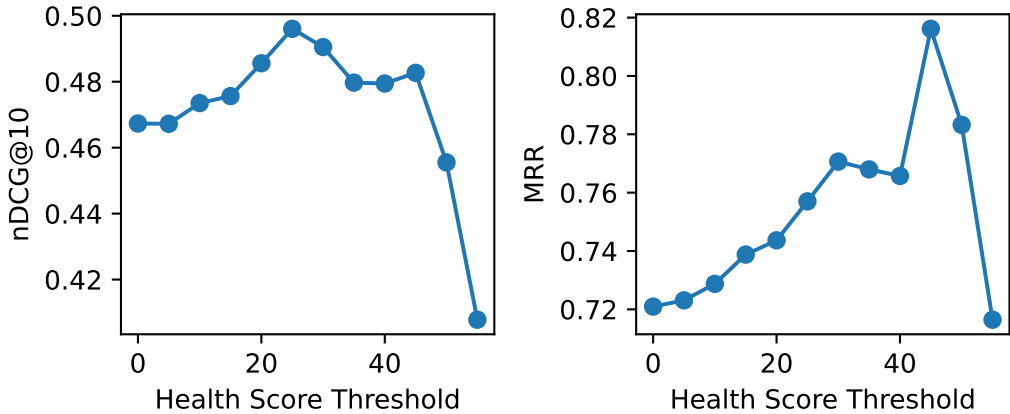
## Pilot Study: Filtering Documents

We can also apply DW to classify if a document is health-related

We applied our software submission to the 2019 TREC Health Misinformation Track

[Abualsaud et al., TREC'19]

TREC Health Misinformation



# Efficiently Scoring the Health-relatedness of Web Pages

## Conclusion

We presented a simple, efficient, effective approach for determining the health-relatedness of queries and documents

- We exemplify that filtering and handling health-related queries may be necessary to accurately retrieve health information
- We show that the effectiveness of a simple lexical approach can be improved by filtering for health-related documents

# Efficiently Scoring the Health-relatedness of Web Pages

## Conclusion

We presented a simple, efficient, effective approach for determining the health-relatedness of queries and documents

- We exemplify that filtering and handling health-related queries may be necessary to accurately retrieve health information
- We show that the effectiveness of a simple lexical approach can be improved by filtering for health-related documents

*Thank you!*