A Large Scale Study on Health Information Retrieval for Laypersons

COMP8755 – Final Presentation

Supervisor: Assoc/Prof Hanna Suominen & Mr. Chenchen Xu
Zhengyang Liu, u6022937
The Australian National University
02/04/2020
Overview

- Introduction
- Motivation & Goals
- Datasets
- Methods
- Results & Discussion
- Conclusion
Introduction

Fig. Percentage of health related search

- specific disease/problem
- treatment/procedure
- diet/nutrition
- exercise
- prescription drugs
- alternative treatments
- health insurance
- depression/stress
- doctor/hospital

Fig. Percentage of accessible way for searching health related information

- Search Engine (Google, Bing, etc)
- Health Portal (e.g. WebMD)
- General site, e.g. Wikipedia
- Social Network (Facebook)
- Other
Motivation & Goals

• Goals:
  A. Improve availability of health searching system
  B. Help provide reliable medical knowledge for laypersons

• What we do:
  A. A large scale dataset is built as a benchmark for different search engine
  B. A global competition based on the dataset is held by us
Datasets

- clef2018collection_B Dataset
  1653 health domains (websites) from CommonCrawl
  3,813,987 files (HTML, XHTML, XML, etc.)
Datasets

<table>
<thead>
<tr>
<th>query_id</th>
<th>query_text</th>
</tr>
</thead>
<tbody>
<tr>
<td>151001</td>
<td>anemia diet therapy</td>
</tr>
<tr>
<td>152001</td>
<td>emotional and mental disorders</td>
</tr>
<tr>
<td>153001</td>
<td>american diabetes association</td>
</tr>
<tr>
<td>154001</td>
<td>high blood pressure</td>
</tr>
<tr>
<td>155001</td>
<td>infectious disease prevention</td>
</tr>
<tr>
<td>156001</td>
<td>food allergy test</td>
</tr>
<tr>
<td>157001</td>
<td>orlistat drug profile</td>
</tr>
<tr>
<td>158001</td>
<td>radiation health effect</td>
</tr>
<tr>
<td>159001</td>
<td>smoking and heart disease</td>
</tr>
<tr>
<td>160001</td>
<td>smoking cessation products</td>
</tr>
<tr>
<td>161001</td>
<td>omega 3 fatty acids</td>
</tr>
<tr>
<td>162001</td>
<td>common medication errors</td>
</tr>
<tr>
<td>163001</td>
<td>Anxiety coping skills</td>
</tr>
<tr>
<td>164001</td>
<td>health benefits of spirulina</td>
</tr>
<tr>
<td>165001</td>
<td>do allergies cause migraines</td>
</tr>
<tr>
<td>166001</td>
<td>ketamine and its potential for abuse</td>
</tr>
<tr>
<td>167001</td>
<td>head and neck cancer</td>
</tr>
<tr>
<td>168001</td>
<td>hiv vaccine phase</td>
</tr>
<tr>
<td>169001</td>
<td>normal lab values</td>
</tr>
<tr>
<td>170001</td>
<td>rheumatoid arthritis prognosis</td>
</tr>
<tr>
<td>171001</td>
<td>breast reduction and lift</td>
</tr>
<tr>
<td>172001</td>
<td>involuntary trembling or quivering</td>
</tr>
<tr>
<td>173001</td>
<td>drug food interaction</td>
</tr>
<tr>
<td>174001</td>
<td>breakfast better student</td>
</tr>
<tr>
<td>175001</td>
<td>Veitaren Erumbel 1%</td>
</tr>
<tr>
<td>176001</td>
<td>pelvic inflammatory disease</td>
</tr>
<tr>
<td>177001</td>
<td>breakfast school children</td>
</tr>
<tr>
<td>178001</td>
<td>drug addiction organization</td>
</tr>
</tbody>
</table>

- **Query Set**
  Top 50 most commonly used queries for health searching
Method

Process of information retrieval system

Main Contribution

Initial Dataset

Queries

Preprocessing

Clean document collection

ElasticSearch

Index

Query Expansion

Extended Queries

Ranked Docs

Retrieval

Evaluation

Main Contribution
Method

- Data Preprocessing:
  
  A. Vital information extraction for each web page (title, heading, content, author, date, link information)
Method

- **Data Preprocessing:**

  B. Document pooling for checking the extraction ratio (proportion of web pages that can extract information) and initial training (1/2000 + drop <2000 domains)

<table>
<thead>
<tr>
<th></th>
<th>title</th>
<th>heading</th>
<th>content</th>
<th>author</th>
<th>date</th>
<th>linkInfo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction ratio(%)</td>
<td>99.32%</td>
<td>94.39%</td>
<td>98.38%</td>
<td>35.41%</td>
<td>28.18%</td>
<td>98.65%</td>
</tr>
</tbody>
</table>

Information extraction ratio
Method

- Query Expansion: build a query reformulator based on REINFORCE algorithm

Query reformulator framework

Illustration of REINFORCE based reformulator

Method

- Information retrieval model: Okapi BM25
  bag-of-words retrieval function, based on frequency

\[
score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i, D)(k_1 + 1)}{f(q_i, D) + k_1(1 - b + b * \frac{|D|}{avgdl})}
\]

\[
IDF(q_i) = \log\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}\right)
\]
Method

● Evaluation

A. Precision@K

\[
\text{Precision@K} = \frac{\text{relevant items in top } k}{K}
\]

B. Normalized discounted cumulative gain (NDCG) @K

\[
\text{DCG@K} = \sum_{i=1}^{K} \frac{r(i)}{\log_2(i+1)}
\]

\[
\text{NDCG}_u@K = \frac{\text{DCG@K}}{\text{IDCG}_u}
\]

\[
\text{NDCG@K} = \frac{\text{NDCG}_u@K}{|u|}
\]
## Results

<table>
<thead>
<tr>
<th></th>
<th>HTML only (baseline)</th>
<th>EI only</th>
<th>HTML + QE</th>
<th>EI + QE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision@10</td>
<td>0.498</td>
<td>0.506</td>
<td>0.504</td>
<td>0.510</td>
</tr>
<tr>
<td>NDCG@10</td>
<td>0.440</td>
<td>0.462</td>
<td>0.464</td>
<td>0.470</td>
</tr>
</tbody>
</table>

Test performance for different methods on 2018 pooling dataset
Results

<table>
<thead>
<tr>
<th>Weighting Ratio</th>
<th>Precision@10</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1:1:1:1:1</td>
<td>0.510</td>
<td>0.470</td>
</tr>
<tr>
<td>1:1:1:1:1:0</td>
<td>0.524</td>
<td>0.490</td>
</tr>
<tr>
<td>1:1:1:1:0:1</td>
<td>0.508</td>
<td>0.454</td>
</tr>
<tr>
<td>1:1:1:0:1:1</td>
<td>0.508</td>
<td>0.466</td>
</tr>
<tr>
<td>5:3:2:1:1:0</td>
<td>0.534</td>
<td>0.492</td>
</tr>
</tbody>
</table>

Test performance for proposed method with different weighting ratio (title : heading : content : author : date : link information) on 2018 pooling dataset
Conclusion

- **Achievement**
  
  A. Build an applicable search engine for laypersons on the clef2018collection_B dataset
  
  B. Set up the evaluation tool and help organize the CLEF eHealth Evaluation Lab(2020)

- **Limitation**

  May not work well on the other dataset due to the various of data preprocessing

- **Future work**

  A. Other commonly used query expansion methods can be applied and compared (like Rocchio-Relevance Feedback)
  
  B. Two CLEF eHealth lab overview papers will be published
Thanks

I would like to show my respect and express my thanks to my supervisor Assoc/Prof Hanna Suominen, and co-supervisor Mr. Chengchen Xu for their guidance, support, and precious time throughout my individual project period.

And I am thankful to Prof. Weifa Liang who is the course convener. He taught me how to deliver the presentation and thesis.