A Full-Text Learning to Rank Dataset for Medical Information Retrieval

Vera Boteva, Demian Gholipour, Artem Sokolov, Stefan Riezler
{boteva, gholipour, sokolov, riezler}@cl.uni-heidelberg.de

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Motivation and Approach

- **bridge lexical gap between layman's English and scientific medical literature**
- **website NutritionFacts.org (NF) provides information on health and nutrition topics with references to research articles**
- **use NF content and references to assemble a corpus for medical IR**
- **extract queries from NF content and documents from scientific articles linked on NF pages**
- **graded relevance scheme from internal linking structure and citation links**

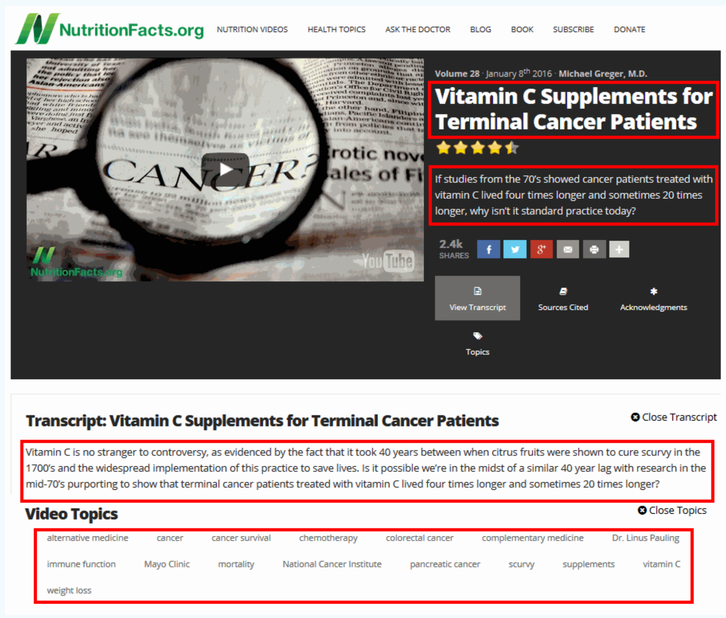
Extracting Queries

Sections in NF pages used for queries:

- **video pages**: title, short description, video transcript
- **blog articles, Q&A pages, topic pages**: title, text

Links between NF pages:

- used as indirect connection between queries and documents through intermediate queries
- videos and blog articles include a list of topics, useful for introducing additional links between queries



Statistics for different query types:

type	# queries	mean/median # tokens per query	mean # docs per query lev. 2	lev. 1	lev. 0
all fields	3244	1890.0/43.5	4.6	41.6	33.8
all titles	3244	3.6/ 1.5	4.6	41.6	33.8
titles of non-topic pages	1429	6.0/ 4.0	4.6	25.4	26.3
video titles	1016	5.5/ 6.0	4.9	23.6	27.1
video descriptions	1016	24.3/21.0	4.9	23.6	27.1

Experiments

Baselines: retrieval with classical **tfidf** and **Okapi BM25** ranking scores

Two Learning-to-Rank methods:

- let $\mathbf{q} \in \{0,1\}^Q$ and $\mathbf{d} \in \{0,1\}^D$ be query and document vectors, dimensions indicating word occurrence for dictionaries of size Q and D
- score function $f(\mathbf{q},\mathbf{d}) = \mathbf{q}^T W \mathbf{d} = \sum_{i=1}^{Q}\sum_{j=1}^{D} \mathbf{q}_i W_{ij} \mathbf{d}_j$, where $W \in \mathbb{R}^{Q\times D}$ is a matrix of word associations between query and document dictionaries
- $\mathcal{R}$ is a set of tuples $(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-)$, document $\mathbf{d}^+$ being more relevant for query $\mathbf{q}$ than $\mathbf{d}^-$
- relevance rank $r_{q,d}$, rank difference $m(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-) = r_{\mathbf{q},\mathbf{d}^+} - r_{\mathbf{q},\mathbf{d}^-}$
- **RankBoost:** $\mathcal{L}_{exp} = \sum_{(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-)\in\mathcal{R}} m(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-) e^{f(\mathbf{q},\mathbf{d}^+)-f(\mathbf{q},\mathbf{d}^-)}$
  - using batch boosting and bagging
- **SGD:** $\mathcal{L}_{hng} = \sum_{(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-)\in\mathcal{R}} (f(\mathbf{q},\mathbf{d}^+) - f(\mathbf{q},\mathbf{d}^-))_+ + \lambda\|W\|_1$
  - where $(x)_+ = \max(0, m(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-) - x)$
  - using Stochastic Gradient Descent

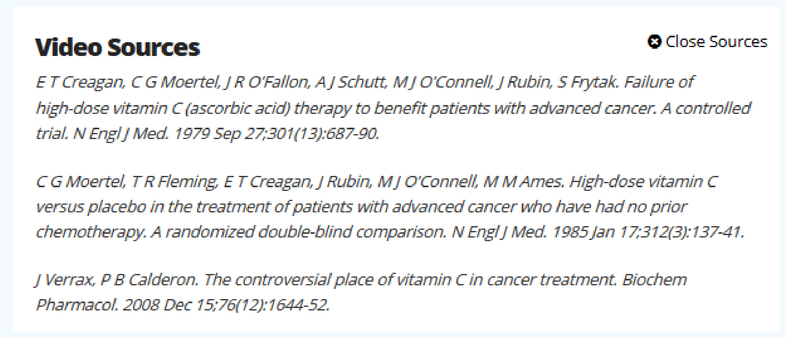
Random division of queries into 80% train, 10% dev, 10% test sets

MAP/NDCG results evaluated for different query types:

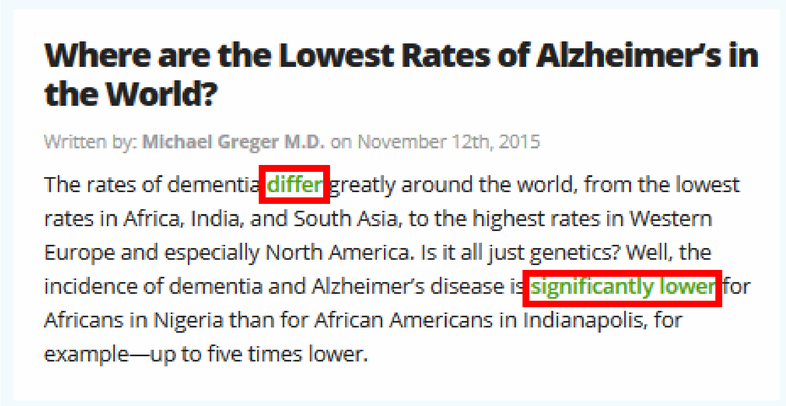
queries	RankBoost	SGD	tfidf	bm25
all fields	0.2632/0.5073	0.3831/**0.6064**	0.1360/0.3932	0.1627/**0.4169**
all titles	0.1549/**0.3475**	0.1360/0.3454	0.1233/0.2578	0.1251/**0.2582**
titles of non-topic pages	0.1615/**0.4039**	0.1775/0.3790	0.0972/0.2851	0.1124/**0.3032**
video descriptions	0.1312/**0.3826**	0.1060/0.3112	0.1110/0.3509	0.1262/**0.3765**
video titles	0.1350/**0.3804**	0.1079/0.3109	0.1010/0.2873	0.1127/**0.3042**

Collecting Documents

Video pages list references, with links to scientific articles:



Other pages link external references in the text:



- 89% of links to external pages lead to articles in **PubMed** or **PubMed Central** archives
- titles and abstracts of these articles were crawled
- 3,633 documents in total, mean/median number of tokens: 147.1/76.

Relevance Levels

- **level 2:** direct link between query **q** and document **d** (e.g. source in a video page)
- **level 1:** indirect link: **q** links **q′** and **q′** links **d** (e.g. article links video, video links source)
- **level 0:** **q′** links **d** and **q** shares at least 70% of topics with **q′**

Conclusion

Key features of the dataset:

- full-text queries of various length, enabling development of complete learning models
- relevance links at 3 levels
- queries in layman's English linked to abstracts of medical research articles
- public availability of the dataset: **www.cl.uni-heidelberg.de/statnlpgroup/nfcorpus/**

Findings of experiments:

- dataset size sufficient for learning ranking models that outperform standard bag-of-words IR methods by far

Acknowledgements

NutritionFacts.org
DFG