York University at CLEF eHealth 2015: Medical Document Retrieval

Andia Ghoddousi Jimmy Xiangji Huang

Information Retrieval and Knowledge Management Research Lab
Department of Computer Science and Engineering
School of Information Technology
York University
Canada
andia@yorku.ca jhuang@yorku.ca

Abstract. This paper presents the results for task 2 of ShAre/CLEF 2015 eHealth Evaluation Lab. We use BM25 as our base normalization method and Pseudo Relevance Feedback to retrieve information regarding patients' health and find the best results to expand the efficiency of information retrieval system. Participants in task 2 are provided with a collection of datasets focused on health web pages. We used queries and submitted 10 runs in TREC style. The runs include the top1000 documents returned for each query. The objective of task 2 is to evaluate the effectiveness of information retrieval system and develop search engines for searching on the web looking for health related documents[1].

Keywords: Information Retrieval, BM25, Pseudo Relevance Feedback

1 Introduction:

The goal of the CLEFeHealth is to develop a way to help and support people for searching and understanding their health [1].

When searching through the database in molecular biology a search term is submitted, then the program will check the query terms and keywords to find information, then information retrieval software will classify the entered data with the existing information in the database and returns the result. When searching in the databases it is usually hard to find the exact term that we are looking for therefore we should modify the query, after the term is found we need to develop our search to find relevant documents, we sometimes need to look in to different databases and link the contents [2]. Databases retrieve and analyze the information in different steps: first they retrieve the sequences by features and annotations or by patterns. Then it compares the sequences [3].

We use Terrier, which is an open source search engine for collecting, indexing and querying the documents, and retrieves the results. This program was developed in Java in the University of Glasgow, Computer Science department. Terrier index the queries from the dataset in order to index, Terrier Parse the collection of documents then develop the tokens and create compacted index structures. Indexing uses Lexical or direct inverting to index. Direct Indexing consists of Pseudo-relevance feedback, document clustering, classification and similarity. [4]

2 Information Retrieval Model

2.1 BM25

In this research we used BM25 normalization model scaling rang from 0 to 1 in order to get the best results possible. In BM25, the weight of each term is assigned by taking in to account the query term frequency in the documents. [6] A document's weight for a query is given by the sum of its weight for each term in the query,

$$BM25(D) = \sum_{i=1}^{|Q|} w(q_i, D)$$ [7]

i=1 where w is the term weight obtained from Equation (1), and |Q| is the length of the query Q. [7]

BM25 developed in okapi system and started to be used in TREC competition. BM25 is used as a baseline and is one of the most established probabilistic term weight model, BM stands for Best Model, [8] which is why we used this model in our research.

Terrier provides two different implementation of BM25, one is the standard BM25 implementation and the second one is BM25-DFR. [8]

DivergenceFromRandomness (DFR) is also one of the first models of Information Retrievals. DFR first selects as basic randomness model, after applying the first normalization tries to normalize term frequency. [8] The term weight is contrarily related to the probability of the term frequency in the document d obtained by model M of randomness [4]

$$\text{weight}(t|d) \propto -\log \text{Prob}_M(t \in d|\text{Collection})$$ [4]

In other words the term weights are measured by calculating the divergence between a term allocation obtained by a random process and the actual term distribution. [8]

sudo bin/trec_setup.sh ../IR/CLEF/data/
sudo bin/trec_terrier.sh -i
sudo bin/trec_terrier.sh -r

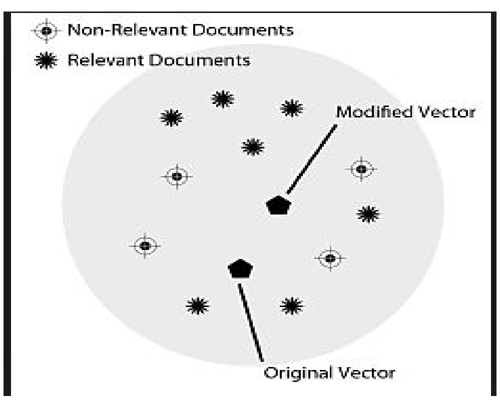
BM25b0.25_7.res — Locked

```
clef2015.test.1 Q0 :skinc4437_12_002296 0 10.641967335027942 BM25b0.25
clef2015.test.1 Q0 :heart3138_12_000039 1 10.597597350287128 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000073 2 10.278342869063291 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000155 3 10.277193969375432 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000123 4 10.277193969375432 BM25b0.25
clef2015.test.1 Q0 :healt3090_12_001617 5 10.17923506777506 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000472 6 10.178496352975618 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000277 7 10.178496352975618 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000175 8 10.131908080424937 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000306 9 10.055348916040149 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000216 10 10.046318481487145 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000164 11 10.045816814262363 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000151 12 10.036627901674809 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000427 13 10.00411518238705 BM25b0.25
clef2015.test.1 Q0 :healt3090_12_001013 14 9.91752274801746 BM25b0.25
clef2015.test.1 Q0 :metho3695_12_000272 15 9.90614063781699 BM25b0.25
clef2015.test.1 Q0 :skinc4437_12_002060 16 9.904471352579513 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000263 17 9.836676061478062 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000053 18 9.83544627947344 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000368 19 9.83544627947344 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000391 20 9.832157371587837 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000322 21 9.80160780082998 BM25b0.25
clef2015.test.1 Q0 :lymph3532_12_001146 22 9.752720724317026 BM25b0.25
clef2015.test.1 Q0 :lymph3532_12_000914 23 9.752720724317026 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000302 24 9.751098868760067 BM25b0.25
clef2015.test.1 Q0 :stret4575_12_000073 25 9.713294871782411 BM25b0.25
clef2015.test.1 Q0 :stret4575_12_000461 26 9.70962494818875 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000333 27 9.692481689681873 BM25b0.25
clef2015.test.1 Q0 :mydr.3757_12_000097 28 9.611868807297343 BM25b0.25
clef2015.test.1 Q0 :plast4085_12_000919 29 9.561235405201224 BM25b0.25
```

2.2 Relevance Feedback and Pseudo Relevance Feedback:

We use relevance feedback to be involved in retrieval process in order to achieve better results and give feedback on the relevance of the documents. In this approach we first dispute a query then using any information retrieval application (Terrier in our research) to index and gain results. When we have the results we can determine the relevant and no relevant documents. [9]

Rocchio algorithm is used for implementing relevance feedback, which was introduced in 1970 and was industrialized using Vector Space Model. Using this algorithm changes search queries in to relevant and non-relevant documents. [10]



The formula for Rocchio Relevance is calculated as follow:

$$\overrightarrow{Q_m} = \left(a \cdot \overrightarrow{Q_o}\right) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}\right) - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}\right)$$

‘a’ is the original query weight, ‘b’ is the weight of the related documents and ‘c’ is the weight of the non-related documents. [12]

2.3 Pseudo Relevance Feedback (PRF):

Which is also known as blind relevance feedback, is used for automatic local analysis. This method is used to do normal retrieval to find initial set of the most relevant documents. This method was used to improve the performance in TREC and ad-hoc retrieval tasks. [12] Pseudo relevance feedback is used to improve retrieval results; this technique is used to obtain results that are originally returned from query to determine if the information is relevant or non-relevant. [13] The relevant documents are clustered together.

2.4 Relevance Feedback:

This theory tried to add the query terms and adjust the weight of each query term of relevant and non-relevant documents and rank the lists, in a way the relevant documents get higher ranks. The best query is the one that has the most similarity to the relevant document. Relevance feedback is used to improve the efficiency of Information Retrieval. [14] Relevance feedback created long revised queries and is sometimes expensive to process.

3 Optimal Query feedback:

This formula tries to maximize the likeness to the relevant documents and minimize the likeness to the non-relevant documents and it can be calculated as follow:

$$\vec{q}_{\text{opt}} = \frac{1}{C_r}\sum_{\vec{d}\in C_r}\vec{d} - \frac{1}{N - C_r}\sum_{\vec{d}\notin C_r}\vec{d} \quad [10]$$

N is the total number of documents.

4 Results:

This Year Share/CLEF eHealth 2015 built results pools from the submissions. Run2 and run3 had the highest priority run. The primary measurement used was P@5 and the secondary measurement used was normalized cumulative gain at rant 10. [17]

4.1 Evaluation with standard TREC_eval metric for Run2 and Run3:

./trec_eval -c -M1000 qrels.clef2015.test.bin.txt runName

YorkU_EN_Run.2.dat

Table Results of Run2

runid	all	BM25b0.31
num_q	all	66
num_ret	all	66000
num_rel	all	1972
num_rel_ret	all	1082
map	all	0.1385
gm_map	all	0.0385
Rprec	all	0.1745
bpref	all	0.2086
recip_rank	all	0.5113
iprec_at_recall_0.00	all	0.5490
iprec_at_recall_0.10	all	0.4097
iprec_at_recall_0.20	all	0.3065
iprec_at_recall_0.30	all	0.2080
iprec_at_recall_0.40	all	0.1147
iprec_at_recall_0.50	all	0.0766
iprec_at_recall_0.60	all	0.0397
iprec_at_recall_0.70	all	0.0179
iprec_at_recall_0.80	all	0.0098
iprec_at_recall_0.90	all	0.0046
iprec_at_recall_1.00	all	0.0046
P_5	all	0.3455
P_10	all	0.2924
P_15	all	0.2596
P_20	all	0.2265
P_30	all	0.1985
P_100	all	0.0964
P_200	all	0.0577
P_500	all	0.0290
P_1000	all	0.0164

[18]

4.2 Reliability Biased-Evaluation*:*

java -jar /tools/ubire.0.1.jar --qrels-file=qrels/qrels.clef2015.test.bin.txt --qread-file=qrels/qread.clef2015.test.graded.txt --readability --rbp-p=0.8 --ranking-file=runName

YorkU_EN_Run.2.dat

Table Results of Reliability Biased-Evaluation of Run2

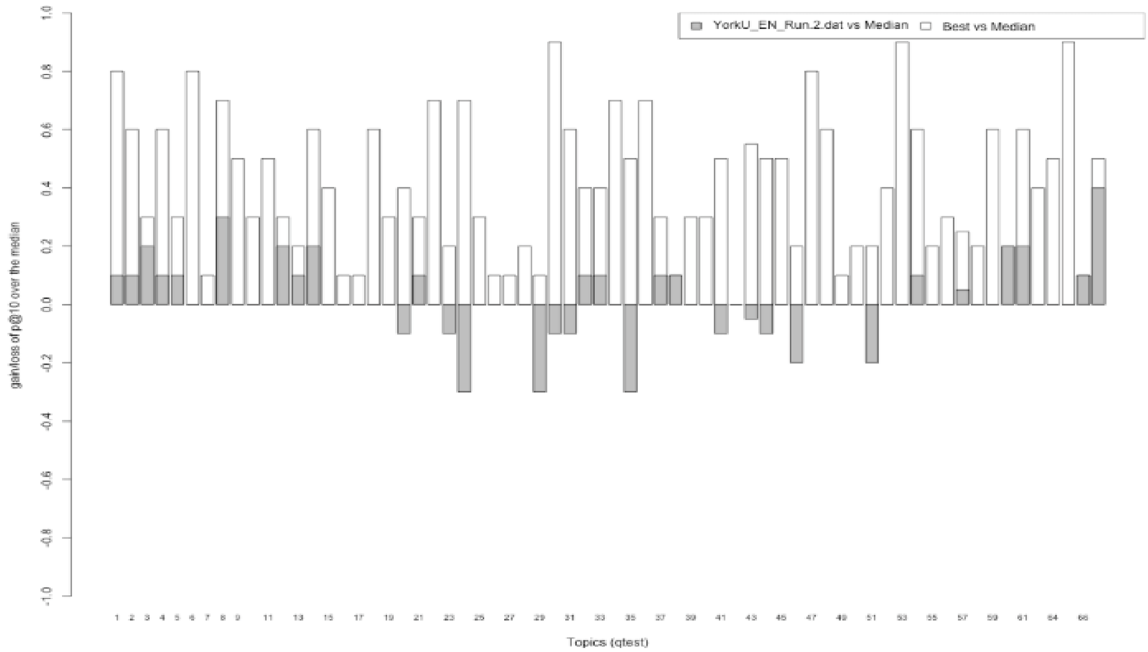
RBP(0.8)	all	0.3151
uRBP(0.8)	all	0.2334
uRBPgr(0.8)	all	0.2404

[18]

This plot compares each of the runs against medium across each has been submitted to CLEF for each query topic where: [18]

grey bars: height(q) = your_p@10(q) - median_p@10(q)
white bars: height(q) = best_p@10(q) - median_p@10(q)

YorkU_EN_Run.2.dat


YorkU_EN_Run.2.dat vs Median
Best vs Median
gain/loss of p@10 over the median
Topics (qtest)

[18]

References:

1. Danielle Mowery, Sumithra Velupillai, Brett R South, Lee Christensen,David Martinez, Liadh Kelly, Lorraine Goeuriot, Noemie Elhadad, Sameer Pradhan, Guergana Savova, Wendy Chapman

2. Hui-Huang Hsu. (2006). Data Mining : Advanced Data Mining Technologies in Bioinformatics.

3. Lesk, A.M. (2001) 'Quality control in databanks for molecular biology'.

4. http://trec.nist.gov

5. D. Kasperowicz and J. Huang. Semantic matching models for medical information retrieval: A case study. In in the Proceedings of the 2012 Advances in Health Informatics Conference (AHIC 2012), 2012.

6. M. Bendersky, D. Metzler, and W. B. Croft. Learning concept importance using a weighted dependence model. WSDM '10, pages 31–40, New York, NY, USA, 2010. ACM.

7. Hsiao, D., and Harary, F. A formal system for information retrieval from files. Communications of the ACM 13, 2 (1970), 67–73.

8. http://ir.dcs.gla.ac.uk/wiki/BM25

9. G. Amati. Probabilistic models for information retrieval based on divergence from randomness. PhD thesis, Department of Computing Science, University of Glasgow, 2003.

10. Buckley, C., Salton, G., and Allan, J., Automatic Retrieval with Locality Information Using Smart, In the First Text REtrieval Conference (TREC-1), National Institute of Standards and Technology, Gaithersburg, MD.

11. http://nlp.stanford.edu/IR-book/html/htmledition/relevance-feedback-and-pseudo-relevance-feedback-1.html

12. http://nlp.stanford.edu/IR-book/html/htmledition/the-rocchio-algorithm-for-relevance-feedback-1.html

13. Jinxi Xu and W. Bruce Croft, Query expansion using local and global document analysis, in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR), 1996.

14. http://www.dcs.bbk.ac.uk/~sven/ir/irchap9.pdf

15. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013

16. http://www.nlm.nih.gov/research/umls

17. http://ceur-ws.org/

18. /CLEF/YorkU_EN_Run/YorkU_EN_Run.html

19. Lorraine Goeuriot and Liadh Kelly and Hanna Suominen and Leif Hanlen and Aurélie Névéol and Cyril Grouin and Joao Palotti and Guido Zuccon . Overview of the CLEF eHealth Evaluation Lab 2015,

20. Palotti, Joao and Zuccon, Guido and Goeuriot, Lorraine and Kelly, Liadh and Hanbury, Allan and Jones, Gareth JF, and Lupu, Mihai and Pecina, Pavel, CLEF eHealth Evaluation Lab 2015, task 2: Retrieving Information about Medical Symptoms, 2015