Exploring effective information retrieval technique for the medical web documents: SNUMedinfo at CLEFeHealth2014 Task 3

Sungbin Choi, Jinwook Choi

Medical Informatics Laboratory, Seoul National University, Seoul, Republic of Korea

wakeup06@empas.com, jinchoi@snu.ac.kr

Abstract. This paper describes the participation of the SNUMedinfo team at the CLEFeHealth2014 task 3. We submitted 7 runs to Task3a (monolingual information retrieval): 1 baseline run using query likelihood model in Indri search engine; 3 runs applying UMLS based lexical query expansion utilizing discharge summary as an expansion term filter; 3 runs applying learning to rank technique utilizing various document features. We submitted 4 runs to Task3b (multilingual information retrieval): 1 baseline run using Google Translate for the English translation; 3 runs applying learning to rank technique on the translated query.

Keywords: Learning to Rank, Document quality, Query expansion, UMLS, Web document, Medical information retrieval

1 Introduction

In this paper, we describe the methods used for our participation of the CLEFeHealth2014 Task 3 User-centered health information retrieval. For detailed task description, please see the overview paper of task3 [1].

2 Methods

2.1 Baseline run

We submitted 1 baseline run (SNUMEDINFO_EN_Run.1) using unigram language model with Dirichlet smoothing [2] in the Indri search engine [3]. Default parameter setting is used. Documents are indexed with Indri without stopword removal. Only title field is used as a query. The queries are stopped at the query time using the

standard 418 INQUERY stopword list, case-folded, and stemmed using Porter stemmer.

In task3b, queries are expressed in French, German and Czech. We used Google Translate [4] for the translation of queries into English. Then, this translated query is applied to the unigram language model.

2.2 Lexical query expansion using discharge summary as an expansion term filter

We submitted 3 runs applying UMLS based lexical query expansion, utilizing discharge summary as an expansion term filter. In this method, document content of discharge summary is assumed as user context information.
Firstly, MetaMap [5] is applied to the query, and UMLS concepts are recognized. We took all of the concepts in the top-scoring final mapping results from MetaMap. Preferred terms per each UMLS concepts are extracted as a candidate expansion terms. Candidate expansion terms which do not occur in the discharge summary are removed, and remaining expansion terms are used as expansion terms. For example, if UMLS concept C0559769 : 'Pelvic cavity structure' is recognized, but discharge summary text does not contain term 'cavity' and 'structure', only 'pelvic' is used as an expansion term.
Original query part is weighted by 0.9 and expansion query part is weight by 0.1.

e.g., Indri query example
#weight (
  0.9 #combine (convalescence open pelvic fracture right superior rami fracture)
  0.1 #combine (fractures open pelvis pelvic right superior))

We applied different parameter settings per each run. In Run Number 2 (which corresponds to SNUMEDINFO_EN_Run.2), we used title field as a query. In Run Number 3, we used title, description and narrative field as a query. In Run Number 4, we used title field as a query, and Dirichlet prior parameter is changed to 1,000.

2.3 Applying learning to rank technique

We applied learning to rank to incorporate various document features into the ranking model.

Document features.

The following features are extracted per each top 1,000 documents.

1. Relevance score of query likelihood model from fulltext index
2. Rank of query likelihood model from fulltext index
3. Relevance score of query likelihood model from title+url index
4. Rank of query likelihood model from title+url index

5. Document quality features

Feature 1 and 2 refers to the relevance score and rank acquired from baseline retrieval model on fulltext index.
Feature 3 and 4 refers to the retrieval result acquired from the index built from keyword content of document. We assumed that terms occurring in the title (e.g., <title> Community-acquired pneumonia </title>) and the url field (e.g., http://www.merckmanuals.com/home/lung_and_airway_disorders/pneumonia/community-acquired_pneumonia.html) as keywords of document.

With regard to the feature 5, we hypothesized a certain notion of document content quality of medical web documents, which defines ideal characteristics that relevant documents could have in common across different types of queries. For example, content reliability, comprehensiveness, whether writing is well structured and so on.

For general web search task, there are several prior studies trying to assess web page quality [6-10]. Many of them utilized link analysis techniques such as PageRank, textual features, webpage design features and so on. In [11], Bendersky et al. tried quality-biased ranking. They used several features to evaluate web document quality such as number of visible terms on the page, depth of the URL path or fraction of table text on the page. These features are informative to assess readability and layout of web page. They combined document quality assessment with query-document relevance ranking

In the medical domain, several prior studies tried to evaluate methodological quality of medical literatures [12, 13], or reliability of medical web document [14-18]. Those prior works focused on the quality classification task only.

In [19], Choi et al. tried to incorporate document evidence quality into document ranking for the high quality literature search. Document quality is defined by the methodological evidence quality of the literature. Target corpus is research literatures in MEDLINE, not a web document. Target users are professionals such as medical doctors and researchers, not the general public. We think that criteria for evaluating document quality in [19] is different from our task.

In this study, we tried to combine medical web page quality assessment with topical relevance ranking. We used only textual features for medical webpage quality assessment. We tried to identify terms possibly relevant to the medical document quality evaluation. Each term's document term frequency is used as a feature value. First author arbitrarily collected about 80 words which considered to be relevant to the document quality (Table 5). These terms are punctuation removed, case-folded, tokenized and stemmed, so finally 82 unique terms are used. We included terms such as 'etiology', 'prognosis', and 'treatment'. When we write any text, we often asked to write in terms of 6W such as what, why, where. We thought that these 'etiology', 'prognosis' terms could be necessary condition to be well-organized medical document, analogous to the 6W. We also included terms like 'md'. When medical doctors wrote document content, in many cases they write their name and qualification at the end of text. We thought that these information could give some assurance about the quality of document content.

Learning.

We used CLEFeHealth 2013' Task 3 [20] test collection as a training set. Relevance assessments are conducted on a 4 point scale (0-4), and a scale 1 correspond to a document that was topical to the query, but unreliable. There are 50 queries, and 6,217 query-document paired relevance assessment in CLEFeHealth 2013'. Using baseline method described in Section 2.1, we retrieved 1,000 documents per each query and prepared dataset. Documents whose relevance assessment is not conducted on CLEFeHealth 2013' gold standard is presumed as non-relevant document (scale 0).

We used random forest algorithm [21] [22] as a learning to rank algorithm. Evaluation metric to optimize on the training data was NDCG@10.

We applied different parameter settings per each run. In its default setting, minimum leaf support parameter is 1, and number of bags parameter is 300. In Run Number 5, we set minimum leaf support parameter to 10. In Run Number 6, we set the number of bags parameter to 50. In Run Number 7, we set the minimum leaf support parameter to 50 (For French, German and Czech language queries in task 3b, parameter setting is same as English query).

3 Results and Discussion

3.1 Evaluation results

Experimental results are described in Table 1, Table 2, Table 3 and Table 4.

Table 1. Task3a result

Runid	P@10	NDCG@10	MAP
SNUMEDINFO_EN_Run.1	0.7380	0.7238	0.3703
SNUMEDINFO_EN_Run.2	0.7540	0.7406	0.3753
SNUMEDINFO_EN_Run.3	0.6940	0.6896	0.3671
SNUMEDINFO_EN_Run.4	0.6920	0.6679	0.3514
SNUMEDINFO_EN_Run.5	0.7520	0.7426	0.3814
SNUMEDINFO_EN_Run.6	0.7420	0.7223	0.3655
SNUMEDINFO_EN_Run.7	0.7420	0.7264	0.3716

Table 2. Task3b French query result

Runid	P@10	NDCG@10	MAP
SNUMEDINFO_FR_Run.1	0.7280	0.7077	0.3344
SNUMEDINFO_FR_Run.5	0.7320	0.7090	0.3371
SNUMEDINFO_FR_Run.6	0.7160	0.6940	0.3254
SNUMEDINFO_FR_Run.7	0.7180	0.6956	0.3295

Table 3. Task3b German query result

Runid	P@10	NDCG@10	MAP
SNUMEDINFO_DE_Run.1	0.7240	0.6874	0.3121
SNUMEDINFO_DE_Run.5	0.7200	0.6790	0.3158
SNUMEDINFO_DE_Run.6	0.7140	0.6716	0.3081
SNUMEDINFO_DE_Run.7	0.6980	0.6645	0.3120

Table 4. Task3b Czech query result

Runid	P@10	NDCG@10	MAP
SNUMEDINFO_CZ_Run.1	0.7220	0.6940	0.3404
SNUMEDINFO_CZ_Run.5	0.7400	0.7011	0.3424
SNUMEDINFO_CZ_Run.6	0.7320	0.6871	0.3327
SNUMEDINFO_CZ_Run.7	0.7220	0.6891	0.3378

In Task3a (Table1), primary evaluation metric was P@10 and secondary evaluation metric was NDCG@10. Our baseline P@10 was 0.738.

SNUMEDINFO_EN_Run.2 (UMLS query expansion method) and SNUMEDINFO_EN_Run.5 (Learning to rank method)'s performance is slightly improved than SNUMEDINFO_EN_Run.1 (Baseline). The amount of performance gain was not large enough: on average, 1~2% improved in terms of P@10 compared to the baseline. If we compare SNUMEDINFO_EN_Run.2 to the baseline performance in terms of P@10, 5 queries improved, 41 queries unaffected, 4 queries harmed. If we compare SNUMEDINFO_EN_Run.5 to the baseline performance in terms of P@10, 14 queries improved, 32 queries unaffected, 4 queries harmed. But both of them failed to attain significant improvement against baseline method when we performed two-tailed paired t-test. Nevertheless, we think that both of our methods; (1) lexical query expansion using discharge summary as context filter; (2) learning to rank approach utilizing medical web page quality features; showed good potentials of improvement against tough baseline. We will try to improve our methods in our future research.

Regarding Task3b (Table2, 3 and 4), we just used Google Translate to translate French, German, and Czech query. These translated queries showed comparable performance compared to the original English query.

4 Conclusion

In CLEFeHealth 2014' Task 3, we tried to test various retrieval technique. We tried lexical query expansion methods and learning to rank approach utilizing various features for medical web document. Evaluation results shows potentially promising results. We will try to improve our methods with more experiments in the future study.

Acknowledgements

This study was supported by a grant of the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea. (No. HI11C1947).

5 References

1. Lorraine Goeuriot, et al. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centered health information retrieval. 2014.
2. Zhai, C. and J. Lafferty, A study of smoothing methods for language models applied to Ad Hoc information retrieval, in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001, ACM: New Orleans, Louisiana, USA. p. 334-342.
3. Strohman, T., et al. Indri: A language model-based search engine for complex queries. in Proceedings of the International Conference on Intelligent Analysis. 2005. McLean, VA.
4. Google Translate. 2014 [cited 2014 June 12]; Available from: https://translate.google.com/.
5. Aronson, A.R. and F.-M. Lang, An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 2010. **17**(3): p. 229-236.
6. Page, L., et al., The PageRank Citation Ranking: Bringing Order to the Web. 1999, Stanford InfoLab.
7. Olteanu, A., et al., Web Credibility: Features Exploration and Credibility Prediction, in Advances in Information Retrieval, P. Serdyukov, et al., Editors. 2013, Springer Berlin Heidelberg. p. 557-568.
8. Rubin, V.L. and E.D. Liddy. Assessing Credibility of Weblogs. in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.
9. Agichtein, E., et al., Finding high-quality content in social media, in Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008, ACM: Palo Alto, California, USA. p. 183-194.
10. Kleinberg, J.M., Authoritative sources in a hyperlinked environment. J. ACM, 1999. **46**(5): p. 604-632.
11. Bendersky, M., W.B. Croft, and Y. Diao, Quality-biased ranking of web documents, in Proceedings of the fourth ACM international conference on Web search and data mining. 2011, ACM: Hong Kong, China. p. 95-104.
12. Aphinyanaphongs, Y., et al., Text Categorization Models for High-Quality Article Retrieval in Internal Medicine. Journal of the American Medical Informatics Association, 2005. **12**(2): p. 207-216.
13. Kilicoglu, H., et al., Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence. Journal of the American Medical Informatics Association, 2009. **16**(1): p. 25-31.
14. Abbasi, A., et al., Detecting Fake Medical Web Sites Using Recursive Trust Labeling. ACM Trans. Inf. Syst., 2012. **30**(4): p. 1-36.
15. Sondhi, P., V.G.V. Vydiswaran, and C. Zhai, Reliability Prediction of Webpages in the Medical Domain, in Advances in Information Retrieval, R. Baeza-Yates, et al., Editors. 2012, Springer Berlin Heidelberg. p. 219-231.

16. Aphinyanaphongs, Y. and C. Aliferis. Text categorization models for identifying unproven cancer treatments on the web. in Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems. 2007. IOS Press.
17. Wang, Y. and R. Richard, Rule-based automatic criteria detection for assessing quality of online health information. 2007.
18. Price, S.L. and W.R. Hersh. Filtering Web pages for quality indicators: an empirical approach to finding high quality consumer health information on the World Wide Web. in Proceedings of the AMIA Symposium. 1999. American Medical Informatics Association.
19. Choi, S., et al., Combining relevancy and methodological quality into a single ranking for evidence-based medicine. Information Sciences, 2012. **214**(0): p. 76-90.
20. Goeuriot, L., et al., ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. Online Working Notes of CLEF, CLEF, 2013.
21. Breiman, L., Random Forests. Machine Learning, 2001. **45**(1): p. 5-32.
22. Dang, V. RankLib. [cited 2014 June 12]; Available from: http://sourceforge.net/p/lemur/wiki/RankLib/.

Table 5. Textual words used as webpage content quality feature

admission
admit
age
background
cause
chief complaint
clinical trial
clinician
condition
ddx
diagnosis
disease
disorder
distinguising diagnosis
dr
drug
dx
epidemiology

ethnicity
etiology
evidence
follow up
followup
frequency
frequent
gender
guide
guideline
history
hospital
illness
introduction
journal
literature
management
md
medical doctor
medication
medicine
morbidity
mortality
ms
occurrence
op
operation
overview
pathology
pathophysiology
phd
physiology
prevent

prevention
problem
prognosis
proof
publication
publish
race
radiology
recommend
recommendation
research article
risk factor
sex
sign
statistic
study
sx
symptom
test
therapy
treat
treatment
tx
update
w/u
work up
workup