On Clinical Decision Support

Arman Cohan Luca Soldaini Andrew Yates Nazli Goharian Ophir Frieder
Information Retrieval Lab - Department of Computer Science - Georgetown University
{arman, luca, andrew, nazli, ophir}@ir.cs.georgetown.edu

ABSTRACT

Recent interest in search tools for Clinical Decision Support (CDS) has dramatically increased. These tools help clinicians assess a medical situation by providing actionable information in the form of a select few highly relevant recent medical papers. Unlike traditional search, which is designed to deal with short queries, queries in CDS are long and narrative. We investigate the utility of applying pseudo-relevance feedback (PRF), a query expansion method that performs well in keyword-based medical literature search to CDS search. Using the optimum combination of PRF parameters we obtained statistically significant retrieval efficiency improvement in terms of nDCG, over the baseline.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Query formulation

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Recently focus on Clinical Decision Support systems has intensified dramatically. An important task of Clinical Decision Support (CDS) is retrieving medical literature relevant to diagnosis and treatment of patients. Queries in CDS are long and narrative and thus different from traditional search. CDS search also differs from traditional biomedical search applications, such as genomics search, in which queries are short and great importance is placed on recognizing variants of domain-specific terms (i.e., gene and protein designations). Additionally, previous work has found that techniques which do not perform well in keyword-based medical literature search, such as expansion using thesauri [1], do perform well when searching biomedical genomics literature [2], illustrating the fact that the best methods for genomics search are not necessarily the best methods for CDS search. We investigate the utility of applying pseudo-relevance feedback (PRF), a query expansion method that performs well in keyword-based medical literature search [3] [1] to CDS search.

Previous research in generic search for biomedical literature has been focused on finding documents using short queries or ad-hoc keywords sequences. The default search engine on PubMed, for example, expands each term in the query using MeSH[1] terms and combines them using Boolean operators. Such approach is unsuited for long, discursive case reports. While studies deal with long queries in other biomedical domains, no study has been published yet on searching medical literature based on medical case reports. Many have evaluated the use of relevance feedback techniques to improve basic search in the medical domain. [4] proposed a relevance feedback retrieval system for PubMed by using RankSVM to re-order documents after explicit user feedback which needs user interaction. Related work have also studied search in medical records by query reformulation techniques [5]. While thesauri has been successfully exploited in the genomics domain [2], [6] they do not seem to improve retrieval performance in generic medical literature, as studied by [1]. Other related studies, tried to answer user medical questions by retrieving MEDLINE abstracts [7],[8].

In summary our contributions are: An exploration of the utility of pseudo-relevance feedback in the clinical decision support search domain and analysis of different aspects of PRF that could affect the efficiency of retrieval in CDS domain.

BCB '14, September 20 - 23 2014, Newport Beach, CA, USA
ACM 978-1-4503-2894-4/14/09.
http://dx.doi.org/10.1145/2649387.2660820

2. METHODOLOGY

We propose a combination of a modified version of "IDF Query Expansion" (IDFQE) [3] – a variant of pseudo relevance feedback (PRF) – with a Health Terms Filter (HTF) that exploits Wikipedia to determine the likelihood of each query term to be health-related.

The PRF component of our system works as follows: for each case report Q, the top N documents are retrieved. The system tokenizes the top k retrieved documents and then computes the boosting coefficient b_j for each term in them: $b_j = \log_{10}\left(10 + w_j\right)$

w_j is computed as suggested in [3]:

$$w_j = \alpha \cdot I_Q(t_j) \cdot tf_j + \frac{\beta}{k}\sum_{i=1}^{k} I_{D_i}(t_j) \cdot idf_j$$

Where t_j is the j-th in the top k documents, $I_Q(t_j)$ is an indicator of the presence of term t_j in the query Q (i.e., $I_Q(t_j) = 1 \iff t_j \in Q$), $I_{D_i}(t_j)$ is an indicator of the presence of term t_j in the document D_i, idf_j is the inverse document frequency of the j-th term in the top k documents. Finally, α and β are smoothing factors.

Successively, the top m terms $c_1, \dots, c_m$ from the documents $D_1, \dots, D_k$ that are not in Q are filtered using the HTF component. For candidate c_l we calculate its likelihood of being associated with a health-related Wikipedia entry:

$$OR(c_l) = \frac{\Pr\{\mathrm{P_w} \text{ is health related} \mid c_l \in P_w\}}{\Pr\{\mathrm{P_w} \text{ is not health related} \mid c_l \in P_w\}}$$

Where $OR(c_l)$ is the odds ratio of candidate c_l. A term $c_l \in \{c_1, \dots, c_m\}$ is added to the original query if $OR(c_l) > \delta$, where δ is a tuning parameter. Finally, each term in the modified query Q' is boosted by its boosting factor. To compute the aforementioned probabilities, we used a Wikipedia dump from November 4, 2013. Those pages containing an information box with one or more of the following medically-related codes were designated as health-related: OMIM, eMedicine, MedlinePlus, DiseasesDB and MeSH

[1] http://www.nlm.nih.gov/mesh/

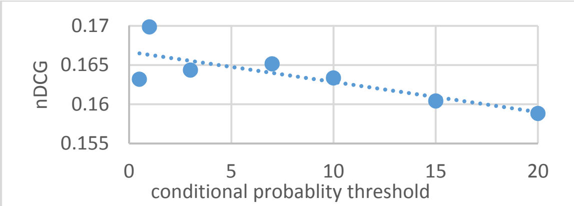

Figure 1 - nDCG in proportion to cond. prob. threshhold

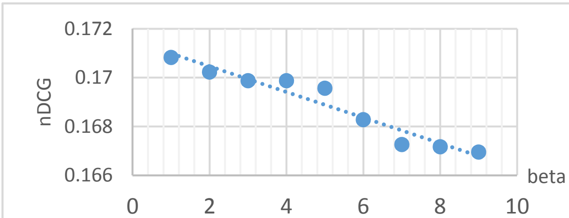

Figure 2. nDCG in proportion to beta

(24,654 pages). The five major parameters affecting our system are α, β, δ, maximum number of added terms and the number of top documents.

3. EXPERIMENTAL SETUP

Our dataset contained a collection of medical case reports extracted from two USMLE preparation books and a snapshot of the Open Access Subset of PubMed central (728,455 full-text papers as of January 1, 2014). ElasticSearch v0.90.7, a search server built on top of Lucene v4 was used to index PubMed documents and to retrieve results. From the USMLE preparation books we extracted those questions that matched the a free-form patient description, a multiple choices question asking for a diagnosis, possible treatment or test to perform and a short paragraph explaining why the correct answer was correct. To determine relevant documents for each case report we separately issued as query the explanation paragraph (q_E) and each answer choice individually ($q_{a_0}, \dots, q_{a_3}$). Documents retrieved by both sets of queries received a relevance score of two, while documents retrieved by q_E only received score of one. This process identified 129 valid queries (case reports), which were further refined by three human assessor instructed to discard primarily quantitative questions. The three assessors' inter-rater agreement was 0.55 (moderate agreement) as measured by Fleiss' kappa. 90 case reports were labeled as valid by at least two assessors and thus kept.

4. RESULTS & DISCUSSION

We evaluated the performance of our system against that of a popular commercial search engine, which serves as our baseline. We fixed the parameter values to be the suggested values in [3]. We also evaluated the effect of parameters on the retrieval performance of the system. Figure 1 shows the effect of conditional probability threshold on nDCG. Higher values of this threshold means that the probability of term being health related, should be higher to be picked as a query term. When the threshold increases we observe a decrease in performance. That is mainly because when the threshold goes up, we end up picking more focused and specific medical terms whereas lower values result in more general terms to be selected. Figure 2 shows the effect of beta on the performance for a fixed value of alpha. As it can be seen, using more weight for relevance feedback for expansion, causes query drift, resulting in performance decline.

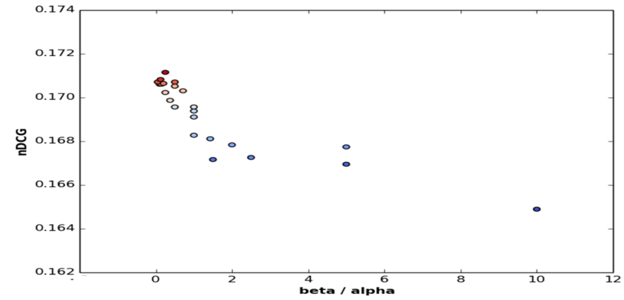

Figure 3.nDCG in proportion to beta/alpha

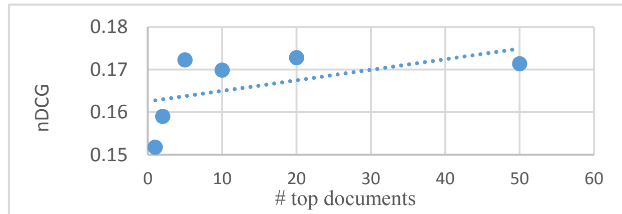

Figure 4 - nDCG in proportion to the # top documents

Figure 3 better shows the combined effect of alpha and beta on the performance. We observe that increased proportion of alpha to beta increases the performance. That means the weight of original query is more important in CDS retrieval efficiency. We observe that when we increase the number of documents to 20, the performance reaches its best (see Figure 4). A very small number of top documents will give us limited resources for expanding the query, resulting in a lower performance. We didn't observe any significant change in the performance by varying number of terms. Using the optimum combination of parameters we obtained an nDCG of 0.172, which is a statistically significant retrieval efficiency improvement over the popular commercial search engine (nDCG of 0.131). Our findings show that the retrieval efficiency, however, still has much room for improvement. This stresses that CDS is a novel search task worthy of further study.

5. ACKNOWLEDGMENTS

This work was partially supported by the US National Science Foundation through grant CNS-1204347.

6. REFERENCES

[1] W. Hersh, S. Price, and L. Donohoe, "Assessing thesaurus-based query expansion using the UMLS Metathesaurus.," *Proc. AMIA Symp* , 2000.

[2] W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli, "TREC 2007 genomics track overview," *TREC*, 2007.

[3] S. Abdou and J. Savoy, "Searching in Medline: Query expansion and manual indexing evaluation," *Inf. Process. Manag.*, vol. 44, no. 2, , Mar. 2008.

[4] H. Yu, T. Kim, J. Oh, I. Ko, and S. Kim, "RefMed: Relevance Feedback Retrieval System for PubMed," *CIKM '09*,

[5] N. Limsopatham, C. Macdonald, and I. Ounis, "A Task-Specific Query and Document Representation for Medical Records Search," in *Advances in Information Retrieval SE -2013*.

[6] N. Stokes, Y. Li, L. Cavedon, and J. Zobel, "Exploring criteria for successful query expansion in the genomic domain," *Inf. Retr. Boston.*, 2009.

[7] T. Bekhuis, E. Tseytlin, K. J. Mitchell, and D. Demner-Fushman, "Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence.," *PLoS One*, vol. 9, no. 1, p. e86277, Jan. 2014.

[8] A. Ben Abacha and P. Zweigenbaum, "Medical question answering," in *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12*, 2012, p. 41.