TREC 2015 paper submission

UWM-UO @ 2015 Clinical Decision Support Track

: QE by Weighted Keywords using PRF

Xiangming Mu & Sukjin You
University of Wisconsin - Milwaukee

Abstract

In the 2015 CDS track, the queries have been expanded in four different ways which we called four different modes. The results shows statistically significantly improvement in terms of infAP, infNDCG and iP10 for some modes as compared to baseline mode which is generated using original query (summary) only without any expansion terms.

Keywords

Pseudo relevance feedback, weighted keywords, MeSH, Query expansion

Introduction

The Information Retrieval (IR) in health field has some unique characteristics. For example, the literature in medical area can be roughly labeled into three categories – diagnosis, test and treatment. This categorization might be utilized to help to improve health information retrieval. Medical Subject Heading (MeSH) has been widely used in health information seeking and retrieval research and products (e.g. PubMed). Pseudo Relevance Feedback (PRF) model has shown usefulness for general information retrieval. Limited research demonstrated its effectiveness for medical literature.

In this project we would like to test the effectiveness of applying these unique characteristics in retrieving health literature to support clinical decision using TREC CDS 2015 dataset which includes 30 queries and 733,138 health literature documents retrieved from PubMed.

Related Work

Query reformulation including Query Expansion (QE) has been a common method to help improve the IR performance. One way is to use knowledge bases such as Wikipedia, dictionary or domain thesauri. Classification of documents might be useful in re-scoring retrieved results by query types. In health IR MeSH is the most popular thesaurus used for QE. D'hondt et al. (2015) showed that MeSH queries generated high precision. Oh & Jung (2015) have tried to use Wikipedia and Unified Medical Language System (UMLS) to collect concepts in queries for QE. Their result show small improvement by using UMLS. Goodwin & Harabagiu (2014) also

tested QE using knowledge bases using Wikipedia, UMLS, Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT), and Google statistics. Their findings also demonstrated some promising results.

Classification by document type had been tried by some researchers for re-ranking retrieved results. The classification by the document type – diagnosis, test, and treatment - might be helpful. For example, Choi & Choi (2014) applied two task classifiers, namely therapy vs. non-therapy and diagnosis vs non-diagnosis, in TREC CDS 2015 tasks and their result showed some positive findings.

The combination using QE and document type classification (e.g. diagnosis, test, & treatment) might generate more effective results. Soldaini et al. (2015) had combined several methods to ameliorate the IR performance; Pseudo Relevance Feedback (PRF) has been used for QE. Selected Wikipedia pages with medically-related information had been further chosen to identify health-related terms in queries. In addition they used classification based on machine learning to re-order retrieved results. Their conclusion showed that improved results have been achieved than just using QE approach.

Methodology

Data

We used TREC CDS 2014 dataset (http://www.trec-cds.org/2014.html#documents) which is the PubMed Central (PMC) snapshot provided by NIST (2014) including 733,138 articles. For each article the content, abstract, title and keywords have been indexed.

Tasks

There are two tasks required for TREC CDS 2015: Task A – no diagnosis description for the test and treatment topics. Task B – added diagnosis description for the test and treatments topics. In this project we participated in both tasks.

Search Engine

In this study, we used Terrier search engine (http://terrier.org/) as our all runs.

Runs

In this study we provided six runs with different approaches to test effectiveness.

UWMUO1: Base Run. The summary fields from original queries (http://www.trec-cds.org/topics2015A.xml) have been used as the queries. We chose this as the base run because in the 2014 CDS

track, Choi & Choi (2014) showed that using summary as a query generated high score. Bayesian smoothing with Dirichlet Prior had been set up as default for the retrieval in Terrier. Porter stemmer has been adopted.

UWMUO2: Query expansion with MeSH keywords. Top 20 documents were chosen based on base run for each query. All the terms in keyword field of each document have been selected as expansion terms in addition to the original base run query. To avoid over-expansion, we calculated the frequency of these keyword terms that appeared in the top 20 documents, and removed all those keyword terms which only appeared once. In addition, we also removed those keyword terms that do not contain any MeSH terms. In other words, we only keep the terms that appeared in the top 20 documents at least twice and contain at least one MeSH term as expanded terms to the original base run query.

UWMUO3: It is same to UWMUO2 except we did not exclude keyword terms selected from the top 20 documents that contain no MeSH term. In other words, any keyword terms in the retrieved top 20 documents that appeared at least twice in the document's keyword field have been used as expansion terms to the corresponding base run query.

UWMUO4: The queries have been formulated by manually using the meaningful keywords existing in the summary.

UWMUO5: Query expansion with MeSH (Keywords & Title). It is same to UWMUO2 except expansion terms were selected from both keyword field and title of the top 20 retrieved documents.

UWMUO6: Query expansion with MeSH Keywords enhanced by diagnosis information. It is same as UWMUO2 except we included additional diagnosis information provided by Task B topics into expansion terms.

Result

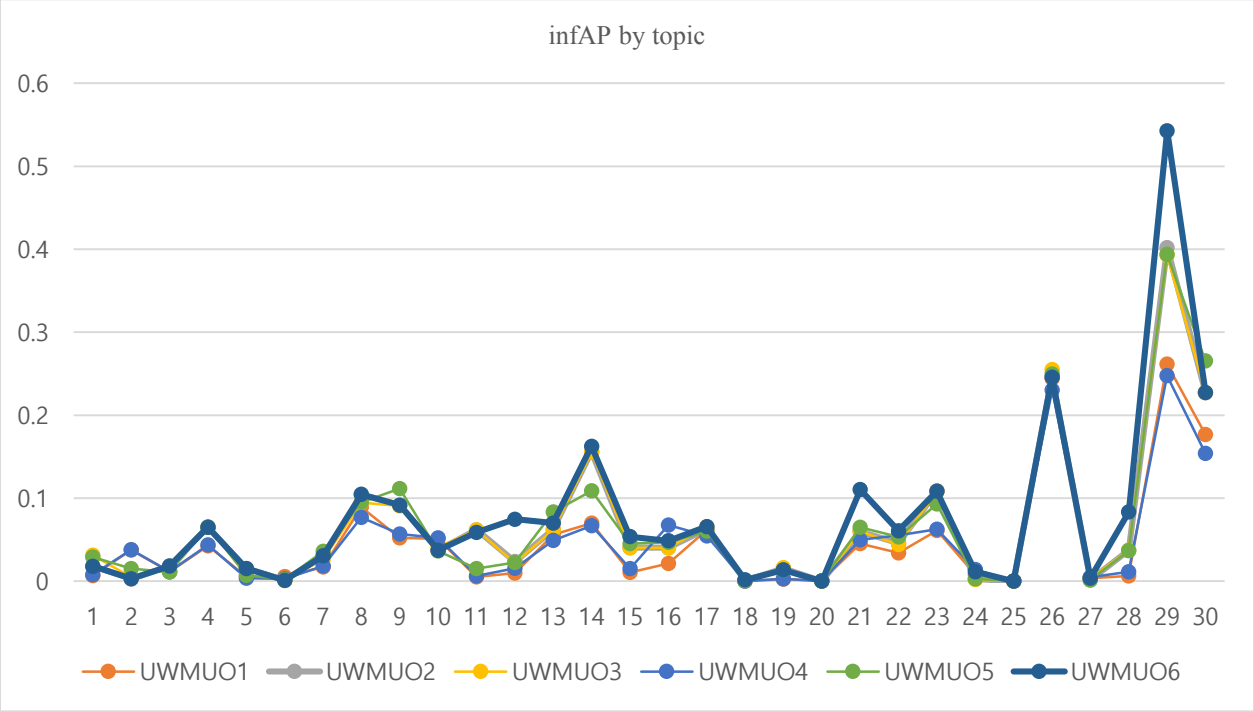
For the runs, infAP (inferred average precision), infNDCG (inferred Normalized Discounted Cumulative Gain) computed at a cut-off of 100 results and iP10 (inferred Precision at 10 results) had been evaluated (Table 1). TREC CDS Track evaluation tool and judgment file have been used for calculation (http://trec.nist.gov/data/clinical2014.html).

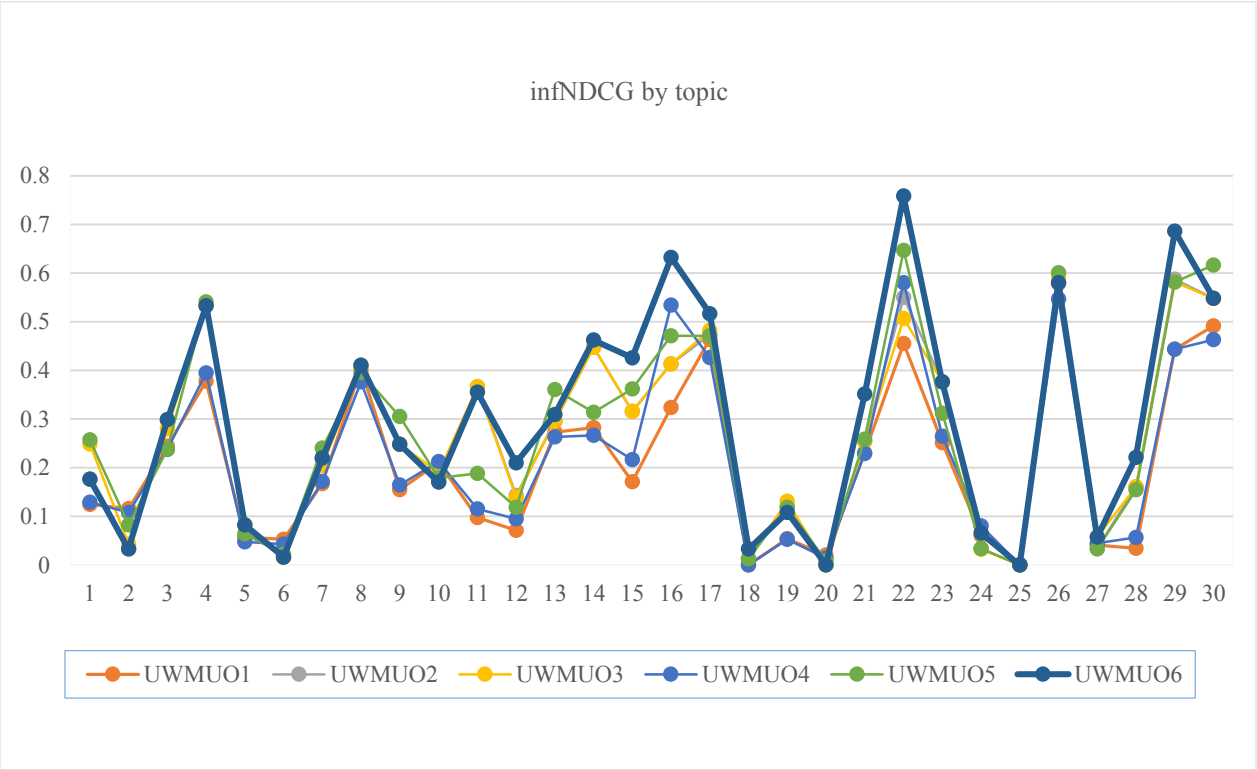
Table 1

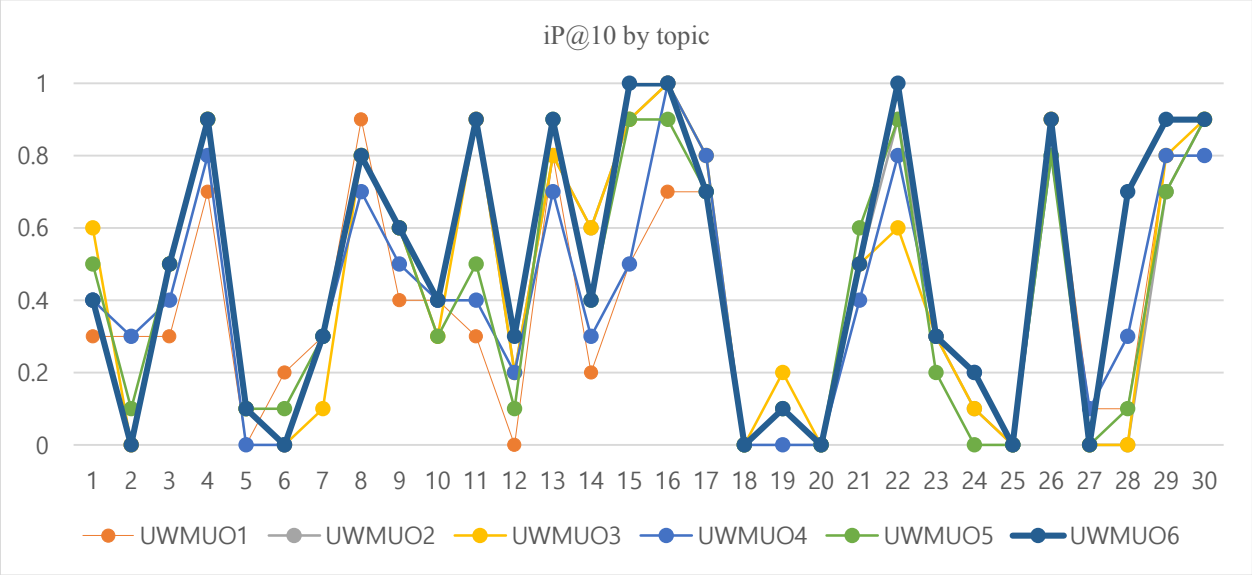
infAP, infNDCG & iP@10

2015	UWMUO1	UWMUO2	UWMUO3	UWMUO4	UWMUO5	UWMUO6
infAP	0.0465	0.0659	0.0661	0.0473	0.0655	**0.0776**
infNDCG	0.2085	0.2634	0.2656	0.2193	0.2663	**0.2962**
iP10	0.3867	0.45	0.4467	0.4067	0.43	**0.49**

For the 30 topics, UWMUO6 generated results with the highest scores in terms of infAP, infNDCG and iP@10. Considering infAP and infNDCG, statistic results using paired T-test demonstrated that there were statistically significant differences between UWMUO1 and UWMUO2, UWMUO3, UWMUO5, and UWMUO6 ($df = 29$, $alpha < 0.05$). We also found that statistically significant differences between UWMUO4 and UWMUO2, UWMUO3, UWMUO5, and UWMUO6 ($df = 29$, $alpha < 0.05$). There was no statistically significant difference between UWMUO1 and UWMUO4 ($df = 29$, $alpha < 0.05$). For iP10, UWMUO6 showed statistically significant difference from UWMUO1 and UWMUO4 ($df = 29$, $alpha < 0.05$).







Conclusion

In general base-run (UWMUO1) and manually selected query approach (UWMUO4) achieved similar performance. Query expansion based approaches showed significantly better performance than base-run (UWMUO1). In other words, the performance of UWMUO2, UWMUO3, UWMUO5 and UWMUO6 has been improved statistically significantly in term of infAP (41.72%, 42.15%, 40.86%, & 66.88% respectively) and

infNDCG (26.33%, 27.39%, 27.72%, & 42.06%). For iP10 improvements are 16.37%, 15.52%, 11.20%, &26.71% respectively and UWMUO6 was statistically significantly better than base-run.

The UWMUO2, UWMUO3, and UWMUO5 showed similar performance in terms of infAP, infNGCG, & iP10. The UWMUO6 outperformed UWMUO2, UWMUO3, and UWMUO5 but the improvements were not statistically significant.

The performance of UWMUO2, UWMUO3, UWMUO5 and UWMUO6 was better consistently over almost all 30 topics than that of the UWMUO1 and UWMUO4 in terms of infAP, infNDCG, & iP10.

References

Choi, S., & Choi, J. SNUMedinfo at TREC CDS track 2014: Medical case-based retrieval task In *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*.

D'hondt, E., Grau, B., Darmoni, S., Névéol, A., Schuers, M., & Zweigenbaum, P. (2014). LIMSI@ 2014 Clinical Decision Support Track. In *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*.

Goodwin, T., & Harabagiu, S. M. UTD at TREC 2014: Query Expansion for Clinical Decision Support. In *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*.

Oh, H. S., & Jung, Y. (2015). KISTI at TREC 2014 Clinical Decision Support Track: Concept-based Document Re-ranking to Biomedical Information Retrieval. In *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*.

Soldaini, L., Cohan, A., Yates, A., Goharian, N., & Frieder, O. (2015). Query reformulation for clinical decision support search. In *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*.