Author: Axel Ancona Esselmann
Institution: SFSU
Class: CSC 849 Information Retrieval
Due Date: 12/11/15
Date: 12/11/15
Topic: Clinical Decision Support Track
Final Document

Document Pool Reduction through Stop Category and Predictive Classification Modeling

by Axel Ancona Esselmann

Abstract:

Innate structure present in communication could potentially be utilized to reduce the document pool that is used to obtain a ranked list of relevant documents. I hypothesize that this approach may reduce the amount of false positives by narrowing the document pool to fewer, but more relevant categories.

1) Introduction

a) Problem domain and prompt:

Clinical decision support is the delivery of medically relevant documents for a patient's history in free text form. This paper will explore previous research in the area. Subsequently I will attempt to build upon related work and improve the relevancy of the retrieved documents in response to free text patient histories.

b) Relevancy to my studies:

As a side project I am in the early stages of creating a website with the purpose of facilitating civil discourse surrounding potentially controversial topics, with the hope of providing individuals who have a knowledge gap in a particular area with the ability to quickly assess the most relevant arguments for all sides of a discussion and form their own educated opinion. The website, instead of using a "post-then-comment-model" that is common for news sites or forums, constrains all user-interaction to an exchange of evidence supported claims. Claims will have an automated visual indication of how well supported each is. One metric will be how many reliable sources are backing a claim. I would like to implement an automated suggestion mechanism that provides potentially relevant papers or articles that authors of a claims can view, curate and use to support their claims. I would have proposed this as my term project, but since no relevance data exists for my project, I chose a topic that is most similar and might lend itself to modification to be used by my discussion portal. In my assessment the task of finding medically relevant papers to help with diagnosis or treatment is almost the same problem as finding relevant publications to support a claim, since both involve queries based on text passages where it is highly likely that very relevant keywords will not be present.

2) Related Work

The paper by Liu and Chu explores query expansion to bridge a vocabulary gap between the query text and the document collection. The authors identify a common problem with clinical decision support, where a physician searches for general terms like "treatment" and documents in the document collection instead have more specific terminology like "chemotherapy." The authors point out that query expansion through statistical analysis, a common approach, often causes irrelevant documents to match,

because “chemotherapy” might often co-occur with “lung cancer” even though the original query might not have anything to do with it.
Liu and Chu propose a knowledge-based query expansion mechanism, where a form of outside knowledge links terms together, that then get appended to the query. In their paper “Knowledge-based query expansion to support scenario-specific retrieval of medical free text,” Liu and Chu use UMLS (Unified Medical Language System) as their outside knowledge source.
I will experiment with knowledge-based query expansion to eliminate frequent words without relevant contribution that through context can be exchanged with query terms that are more specific.

The paper by B. Koopman et al combines two previous approaches to medical information retrieval, that of graph based retrieval and the attempt of grouping words into concepts.
In graph based retrieval models, terms are not retrieved like a bag-of words, but rather a graph is created where words make up the nodes and edges make up the connections between words. The importance of a word in a query is determined by the connectivity of the word within the graph. PageRank is one example of a graph based retrieval model.
B. Koopman’s paper makes use of the natural language processing system MetaMap, which is base on UMLS which was used by Liu and Chu. MetaMap is a project by the U.S. National Library of Medicine to map individual words to concepts. MetaMap makes use of knowledge sources outside of the document collection to create categories. MetaMap had previously been used for a Bag-of-Concepts model for medical IR, but B. Koopman combines the graph based approach with the Bag-of-Concepts approach with a significant improvement over term-based-tfidf, concept-based-tfidf and term-graphs.
I will make use of the idea that using concepts can improve relevancy and I will build upon MetaMap.

3) Proposed Work

a) Problem analysis:

I plan to rely on innate structure present in effective transfer of ideas to reduce the document pool that is used to obtain a ranked list of relevant documents. And i plan to expand on the idea of stop words by broadening the approach to “stop categories”.

A well crafted transfer of ideas has an introduction, a body, and a conclusion, similar to the standard essay form. The introduction primes the audience with the problem domain to be encountered in the body and the conclusion synthesizes the ideas presented in the body. I analyzed example patient histories, and most appear to follow a similar pattern: Each begins with a description of the main symptoms or actions taken. In general subsequent sentences appear to be getting ever more specific.

Since the introduction is an expose of the ideas found in the body, its vocabulary is predictive of the vocabulary likely to be encountered in the body.

If I cluster the collection vocabulary into overlapping categories with MetaMap and generate a list of potential categories for a patient history from its introduction, I hypothesize that due to the predictive properties of vocabulary in the introduction, it is possible to create a ranking for the relevancy of categories by analyzing the vocabulary of the body in respect to the categories generated from the introduction. Each word in the body that matches a category from the introduction makes that category more relevant. Categories that are present in the body but were not introduced in the introduction are not considered. This will result in a reduction of the document pool by focussing on categories that, inferred by the structure of a patient history, are more relevant.

I hypothesize that this approach can reduce the amount of false positives by narrowing the document pool to a few highly relevant categories.

Assumptions:

- Effective communication can be modeled by the standard essay form
- Patient histories are a form of effective communication
- Using the entire document collection will result in a high number of false positives.
- Categories generated by MetaMap are sufficiently broad to allow for many overlaps, but categories are also sufficiently general to avoid blanket coverage.

Regular tfidf based queries and concept based graph search as proposed by Koopman et al make terms like "Year" and "age" part of their query. These terms can be found in almost all patient histories, since they describe the age of the patient, but out of context these words are either too general ("15", "year", "old", "girl"), or too specific ("15 year old girl"). As described by the Liu paper, allowing general terms into the query results in a multitude of irrelevant documents that will get considered for retrieval. One approach could be to turn these common words into stop-words for the application of clinical decision support, but I argue that the context of a patient history can provide information that will make it possible, with the use of a training set, to generate a list of “Stop Categories,” which, just like lists of stop words that include frequently occurring terms, are categories that cover terms that are too broad or too frequently occurring to be helpful in the retrieval of relevant documents. I propose to remove categories or words covered by Stop Categories to increase precision.

b) Proposed implementation

Preparation:

An inverted index is generated for the entire document collection.
Query text segmentation: introduction and body sections are mapped to sentences. A tunable algorithm maps several words or sentences to the introduction, the rest is mapped to the body.

Document pool reduction:

1) Category generation
   Words in the beginning of a patient history will be used to generate a list of relevant categories. Stop words are removed, synonyms for concepts are found and a list of categories is generated.

2) Each of the categories in the list will get scored according to how predictive it is or what vocabulary is used in the subsequent patient history. Only documents that are clustered in the highest scoring categories will be evaluated

Stop Categories:

- The 2015 dataset for the clinical support track is run through MetaMap and used to manually generate a list of categories that span words that are not relevant to a query.

4) Experimental Methodology:

a) Evaluation metrics:

All test runs use the same set of 30 queries, taken from the NIST 2014 Clinical Decision support track website. When needed, training sets are generated from 2015 data.. Relevance data for analysis of the quality of the retrieved documents is provided by NIST and will be used to evaluate my proposed hypotheses. The first 1000 documents returned are are taken into consideration and compared to the relevance data.

Ranked retrieval evaluation:
I will be comparing test runs using trec_eval, a utility provided by the National Institute of Standards and Technology to evaluate recall for participants of their medical support track.

Baselines are established through the following two tests retrievals:

1) **TB-TFIDF**: bag of words term-based-tfidf
2) **CB-TFIDF**: category-based-tfidf, phrases are generated with the MetaMap tool and used undergo the same bag of words tfidf search.

Six test runs are compared to the baseline results:

1) **CB-DPR-TFIDF**: Category based document pool reduction: Document pool reduction by creating categories from terms used in the introductory part of the patient history. The entire patient history, with the summary prepended, is used to generate phrases with MetaMap. Each phrase has a category assigned by MetaMap and categories get scored by how many tokens belong to them. A percentage of categories is discarded.
2) **KBQE-TFIDF**: Instead of eliminating categories not seen in the introduction, the introduction is increased in weight by term duplication.
3) **SC-TFIDF**: Stop Categories are generated and eliminated from the query..

In comparing the six experimental runs I am focussing on four measurements made by the trec_eval utility:

Number of relevant documents (num_el_ret): The total number of relevant documents returned.
Mean average precision (MAP): Average of all precision values obtained for the top k documents when a relevant document is retrieved.
Precision at 10 (P@5): Precision of top 10 results
Precision at 30 (P@20): Precision of top 30 results. This is the primary metric taken into consideration when trek participants are scored.

b) Data:

A total of four experiments where conducted. Each experiment had a few parameters that could be manipulated. Runs that were part of the same experiment have the same color in the left-most column. Shades of green and orange indicate which experiment had the best and worst outcomes. Lighter shades of green signify a better outcome, shades of orange indicate poor performance when compared to other results.

run	num_rel_ret	map	P@10	P@30
Base 1: TFIDF	1620	0.1115	0.2667	0.2267
Base 2: MetaMap and TFIDF	1659	0.1196	0.27	0.2333
Tfidf Categories Removed	1655	0.1168	0.27	0.2378
Tfidf Categories Predicted 0.1,15	1528	0.1069	0.2633	0.2378
Tfidf Categories Predicted 0.4,15	1477	0.0976	0.2767	0.2289
Tfidf Categories Predicted 0.5,15	1359	0.0891	0.26	0.2089
Tfidf Categories Predicted 0.1,20	1543	0.1074	0.2567	0.2289
Tfidf Categories weighting 1.	1333	0.0832	0.2167	0.1844
Tfidf Categories weighting 2	1527	0.1003	0.2767	0.2278
Tfidf Categories weighting 3	1628	0.1138	0.2733	0.2433
Tfidf Categories weighting 4	1631	0.1117	0.27	0.2456
Tfidf Categories weighting 5	1619	0.1107	0.27	0.2356
Tfidf Categories weighting 6	1621	0.1143	0.2667	0.2411

5) Results and Analysis:

First experiment: Category based document pool reduction (CB-DPR-TFIDF)

Document pool reduction by creation of categories from terms used in the introductory part of the patient history. Two parameters where manipulated, the number of tokens counted towards the introduction, and the percentage of categories discarded after priorities have been assigned. I recorded four measurements that appear to be indicative of what trends with the data. The number of retrieved documents appears to decrease, as the percentage of categories discarded increases. It appears that a slight improvement in precision could be achieved with small percentages of categories discarded. This effect was canceled out when more words where used for category generation. Optimum values for the parameters seemed to be in the range of 10 percent of categories discarded and 15 words counted towards category selection. This resulted in a 2% improvement in precision at 10 when compared to the baseline. Tests run without prepending the summary result in very similar measurements.

run	num_rel_ret	map	P@10	P@30
Base 1: TFIDF	1620	0.1115	0.2667	0.2267
Base 2: MetaMap and TFIDF	1659	0.1196	0.27	0.2333
Tfidf Categories Predicted 0.1,15	1528	0.1069	0.2633	0.2378
Tfidf Categories Predicted 0.4,15	1477	0.0976	0.2767	0.2289
Tfidf Categories Predicted 0.5,15	1359	0.0891	0.26	0.2089
Tfidf Categories Predicted 0.1,20	1543	0.1074	0.2567	0.2289

Second experiment: Category weighting (CB-LOG-W-TFIDF)

The importance of the introductory section is scaled by query term repetition, taking advantage of the fact that repeating query terms with tifidf score result in higher valuation. Repetition was calculated with the function

$f(n) = \log_2((\text{TotalTokenNbr} * a) - (b - n))$

for values larger than 1 where n is the n is the position of the token in the patient history. I am reporting two parameters that seem to be indicative of what is happening with the data. The first measurement uses values of a = 2 and b = 16. The second measurement uses values a = 1 and b = 4. It appears that retrieval worsens as a increases. Larger a cause more repetition. It also appears that a small a with a small b, while having a worse retrieval number, increases precision at 10 by about 2 percent when compared to the baseline.

run	num_rel_ret	map	P@10	P@30
Base 1: TFIDF	1620	0.1115	0.2667	0.2267
Base 2: MetaMap and TFIDF	1659	0.1196	0.27	0.2333
Tfidf Categories weighting 1.	1333	0.0832	0.2167	0.1844
Tfidf Categories weighting 2	1527	0.1003	0.2767	0.2278

Third experiment: Percent Token Position Weighting (PTPW-TFIDF)

Since a logarithmic weighting scheme does not appear to be very effective, and since an excess number of repetitions seemed to drastically degrade the quality of retrieval results, I experimented with a constant for the number of repetitions and a percentage of token-position for a cutoff. All measurements recorded below use a constant of two. Percent of tokens doubled are: 20%, 30%, 40% and 50%. Between 20% and 30% a significant improvement in the precision at 30 could be detected. Number of documents slightly decrease as percentages increase.

run	num_rel_ret	map	P@10	P@30
Base 1: TFIDF	1620	0.1115	0.2667	0.2267
Base 2: MetaMap and TFIDF	1659	0.1196	0.27	0.2333
Tfidf Categories weighting 3	1628	0.1138	0.2733	0.2433
Tfidf Categories weighting 4	1631	0.1117	0.27	0.2456
Tfidf Categories weighting 5	1619	0.1107	0.27	0.2356
Tfidf Categories weighting 6	1621	0.1143	0.2667	0.2411

Fourth experiment: Stop Categories (SC-TFIDF)

I used 2015 track queries that where categorized by MetaMap as a training set to manually trim the query terms by considering category membership. In essence I created “Stop Categories”, categories that occur frequently but often don’t hold much information. All tokens belonging to the categories “Temporal Concept” or “Quantitative Concept” where discarded. A decrease in number of relevant documents returned and a minute improvement at precision at 30 could be observed.

run	num_rel_ret	map	P@10	P@30
Base 1: TFIDF	1620	0.1115	0.2667	0.2267
Base 2: MetaMap and TFIDF	1659	0.1196	0.27	0.2333
Tfidf Categories Removed	1655	0.1168	0.27	0.2378

5) Summary:

It appears that reducing the number of query terms with the two methodologies I propose only saw minor if any measurable improvements in precision. Some experiments degraded the retrieval results. It will take a larger training set and a fresh set of queries to confirm the minor trend that “Stop categories” are able to increase precision with tfidf. It appears there might be a decrease in the number of retrieved relevant documents that comes with the increase in precision.

The categories generated by MetaMap and used to trim the document pool by taking advantage of innate structure in communication where either too broad, or the effect I hypothesize is very small.

6) Bibliography

4) Liu, Zhenyu, Chu, WesleyW. 2007. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. In Information Retrieval, Volume 10, Number 2, Page 173, Kluwer Academic Publishers. http://dx.doi.org/10.1007/s10791-006-9020-6

5) Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2012. Graph-based concept weighting for medical information retrieval. In Proceedings of the Seventeenth Australasian Document Computing Symposium (ADCS '12). ACM, New York, NY, USA, 80-87. http://dx.doi.org/10.1145/2407085.2407096