The GUC Goes to TREC 2004: Using Whole or Partial Documents for Retrieval and Classification in the Genomics Track

DRAFT

Kareem Darwish and Amgad Madkour
Department of Information Engineering and Technology
The German University in Cairo
5th District, New Cairo, Cairo, Egypt
{kareem.darwish,amgad.madkour}@guc.edu.eg

Abstract

We were interested in examining the relative effect of using parts of the documents, different combinations of parts of the documents, or whole documents on retrieval and classification. We were also interested in the effect of MeSH terms on retrieval. Our experiments show that indexing titles, abstracts, and MeSH terms for adhoc retrieval yielded statistically significantly better results than any other part or combination of parts, with abstracts outperforming any other individual part of the documents. In the triage sub-task, using whole documents for training a classifier outperformed using titles, abstracts, diagram captions, MeSH terms, and windows of text around gene names. However, training a classifier using the combination of titles, abstracts, and MeSH terms produced results comparable to using whole documents.

1 Introduction

The overarching theme of our experiments in the adhoc retrieval and classification tasks was the determination of the relative effectiveness of different parts of the documents or a combination of parts on retrieval and classification.

For the adhoc retrieval task, we examined the effect of inclusion and exclusion of the Medical Subject Headings (MeSH) terms in the indexed documents on retrieval effectiveness and we compared the relative effectiveness of the MeSH terms compared to the title and abstract fields of the MEDLINE citations. MeSH are hierarchically structured controlled vocabulary terms developed by the National Library of Medicine (NLM) to classify, mostly manually, documents such as the MEDLINE citations.

For the triage sub-task, we were interested in comparing the relative effect of training a classifier using the full text of a document or alternatively selected parts of the documents. The selected parts of the documents that we experimented with included titles, abstracts, MeSH terms, diagram captions, small windows of text surrounding genes and gene products, and combinations of different parts in the document. We also report here on experiments that we performed for the annotation sub-task.

The rest of the paper will be organized as follows: section 2 provides background on some results reported in the literature, sections 3 and 4 describe the experimental setup and the results of the experiments respectively, and section 5 concludes the paper.

2 Background

Considerable amount of research has focused on the effect of manually assigning MeSH terms to documents in IR applications. Srinivasan [8] examined the effect of the inclusion and exclusion of MeSH terms on information retrieval using a collection produced by Hersh et al, which will be referred to hence forth as the Hersh collection [4]. The Hersh collection has 75 topics and 2,344 medline citations, which include the titles and abstracts of the articles along with manually assigned MeSH terms. Srinivasan's experiments showed that the inclusion of the MeSH terms in documents statistically significantly improved retrieval effectiveness. She also demonstrated that the inclusion of MeSH terms can be combined with blind relevance feedback and thesaurus-based query expansion to produce even better retrieval effectiveness. She used a statically generated thesaurus for query expansion. Aronson et al examined the effect of including and excluding MeSH terms in documents on retrieval and confirmed the results obtained by Srinivasan's [1]. The experiments of Aronson et al showed that MeSH terms statistically significantly improved retrieval effectiveness and that using thesaurus-based expansion further improved retrieval effectiveness. Aronson et al used the UMLS Metathesaurus for query expansion [1]. French et al examined the effect of augmenting user queries with automatically generated MeSH terms. The MeSH terms were selected using the Entry Vocabulary Indexes technique which employs a probabilistic mapping between natural language text and controlled vocabularies [3]. French reported that augmenting queries with a small number of MeSH terms statistically significantly improved retrieval effectiveness.

In the 2002 KDD Cup competition, held in conjunction with SIGKDD, the main task focused on identifying articles for FlyBase, a publicly available database on the genetics and molecular biology of Drosophila (fruit flies), containing experimental evidence of gene expression of gene products, which might then be manually curated [5]. Some of the problems they faced in the task included the casual mention of genes, the mention of mutated gene expressions (as opposed to natural expressions of genes), and the ambiguity between the gene and its transcript [10]. The ClearForest and Celera, the group reporting the best results for the task, utilized constrained pattern matching of diagram captions to identify which documents contain experimental evidence. Their justification for using diagram captions relies on the fact that curators who manually select papers look mainly at the diagrams in the paper to ascertain the presence of experimental evidence [10]. In the absence of diagrams from the provided text only documents, diagram captions provide an indication of the content of the diagrams. For the same task, Shi et al. trained a Naïve Bayes Classifier based on the distance between a gene name and keywords suggesting experimental evidence [6].

3 Experimental Setup

3.1 Ad-hoc Retrieval Task

For the ad-hoc retrieval task, we submitted one official run and conducted a group of post-hoc runs as follows:

3.1.1 Official Run

For the official run, we used PSE which is an open-source retrieval engine that uses OKAPI BM-25 weighting formula [2]. We indexed the title, abstract, and MeSH fields of the documents and removed extraneous fields such as author names and publication dates. Before indexing the documents, we normalized the case of all the tokens and removed all stopwords based on the stopword list used by PubMed, but we employed no stemming. For the queries, we used the title and need fields of the 50 topics. The queries were processed in the same manner as the documents.

3.1.2 Post-hoc Runs

For the post-hoc retrieval runs, we used the Lemur toolkit exclusively using the default settings of OKAPI BM-25 weighting formula. Lemur is a language modeling and information retrieval toolkit developed jointly between Carnegie Mellon University and University of Massachusetts at Amherst. When using blind relevance feedback in Lemur, we used the default settings in which a query is augmented with the 20 best terms from the top 5 retrieved documents. For the queries, we used the same queries as those from the official run.

Our initial attempts to index the entire collection in one index all failed for a reason that we are still investigating. Therefore we split the collection into 5 sub-collections and indexed and searched each sub-collection independently. Since the collection was distributed into 5 document files, each document file was used as a sub-collection. After searching each sub-collection separately, the ranked lists for each query were combined into a single list and the documents in the list were sorted based on the scores reported by Lemur. Splitting the collection should not affect term frequencies or document length normalization (as the average document length is fairly consistent across sub-documents), but the splitting might affect document frequencies. The splitting is likely to have an effect on blind relevance feedback, because the feedback is done on each sub-collection separately. The effects of splitting need to be further investigated.

The runs we performed had two main aims:

1. We wanted to reexamine the effect of excluding MeSH terms on retrieval effectiveness. To do so, we indexed the collection once using the title, abstract, and MeSH fields, which we will refer to as TWM, and one more time with the title and abstract fields only, which we will refer to as TW.
2. We wanted to examine the effect of indexing three different parts of the documents, namely titles (T), abstracts (W), and MeSH terms (M) on retrieval effectiveness. Even though indexing terms from specific parts of the documents to the exclusion of other parts was expected to yield lower retrieval effectiveness than indexing the full documents, examining the effect of different of parts of the documents may indicate which part contributes the most number of valuable terms, which may consequently be used to improve a processes such as blind relevance feedback by skewing term selection. We examined the impact of the titles, the MeSH terms, and the weightiest 20 terms from the abstract field (term weighting was determined using the OKAPI BM-25 formula). The title field represents a natural language summary of the document which was generated by the original author; the MeSH field represents a controlled vocabulary representation of the document which was generated by domain professionals; and the

most valuable terms are automatically generated summaries. The indices and subsequent runs will be referred to as T, M, and W for the title, abstract, and MeSH term respectively.

All in all the collection was indexed 5 times and was searched with and without blind relevance feedback.

3.2 Classification Tasks

3.2.1 Triage Task

For the triage task we submitted 5 official runs and conducted a series of additional (unofficial) runs. For all the runs we used SVM Light with either a linear or a polynomial kernel, and we trained SVM Light with all the default parameters using all the provided positive and negative training examples. For all the runs, official and unofficial, the only text processing that we performed was case normalization and we performed no stemming or stopword removal. We opted out of performing stemming and stopword removal because experiments we performed on the training set indicated that neither of them improved filtering effectiveness. We augmented the full length documents with manually assigned MeSH terms from PubMed.

We performed the following runs:

3.2.1.1 Official runs

We submitted 5 official runs. In four of the runs, SVM Light was trained using the diagram caption from the papers. The use of diagram captions was inspired by their use in the 2002 KDD Cup by ClearForest and Celera [10]. Based on some preliminary experiments we performed on the training data, results suggested the effectiveness of captions for training. As a baseline run, we used whole documents for training. Listed below are the names of the runs, the thresholds used, and the part of the documents on which SVM light was trained.

Name	Kernel	Threshold	Train on
GUClin1260	Linear	-0.973	Captions
GUClin1700	Linear	-1.000	Captions
GUCply1260	Polynomial	-0.894	Captions
GUCply1700	Polynomial	-0.932	Captions
GUCwdply2000	Polynomial	-0.950	Whole Document

Some preliminary experiments we performed on the training examples provided some direction to the choice of threshold, but the experiments provided a direction rather than definite threshold choices. Therefore, our choices of threshold were somewhat arbitrary.

3.2.1.2 Unofficial runs

For the unofficial runs, we tried a variety of setups in which we varied two factors, namely:

1. Which part of the document SVM Light was trained on. We tried whole documents (WD), titles (T), abstracts (W), MeSH terms (M), captions (C), title + abstract (TW), title

+ abstract + MeSH terms (TWM), title + abstract + MeSH terms + captions (TWMC), a window 5 words preceding and following each mention of a gene or a gene product (W5), and lastly a window of 10 words preceding and following each mention of a gene or a gene product (W10). Genes and their products were recognized using a modified version of YAGI, which is short for Yet Another Gene Identifier [9].

2. The SVM Light cut-off thresholds to find the most effective threshold. We varied the value of the threshold between -0.90 and -1.10 with increments of 0.01.

   A polynomial kernel was used for all the unofficial runs.

3.2.2 Annotation Task

For the annotation task, we submitted 5 official runs and did not perform any unofficial runs. We followed three paths in performing the runs as follows:

1. GUCbase: We performed a baseline run in which all (document, gene name) pairs were annotated with all three possible annotations, namely Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).
2. GUCsvm0 and GUCsvm5: For these two runs, we trained SVM Light for each of the three possible annotations using the GO sub-tree entries. The GO sub-tree entries corresponding to one of the annotations were used as positive examples while the entries in the two other GO sub-trees were used as negative examples. We only used the name and definition fields from the GO entries with case normalization and with no stemming or stopword removal. Then, we concatenated all the paragraphs that mention the gene that we wish to classify from the (document, gene name) pair. We used YAGI to identify genes and we used carriage return to detect paragraphs. We then classified all the concatenated paragraphs using SVM Light. If the lumped paragraphs containing the gene name had a score of 0.0 or -5.0 for the GUCsvm0 and GUCsvm5 respectively for a particular classification, then we generated (document, gene name, classification) tuple.
3. GUCir30 and GUCir50: For these two runs, we used Lemur to index all the entries of the GO with only case normalization and no stemming or stopword removal. Again we used the OKAPI BM-25 weighting formula in Lemur. We used the concatenated paragraphs containing the gene from the (document, gene name) pair as queries. Upon searching the GO entries using our queries, we added the sum of the logs of OKAPI BM-25 scores for the top 30 returned documents for each of the BP, MF, and CC classifications separately. For example, if 5 of the returned entries belonged to the BP sub-tree of GO, then the score of BP is just the sum of the logs of the returned scores for these 5 entries. If the score of any of the classifications was greater than a threshold of 30 and 50 for the GUCir30 and GUCir50 runs respectively, then the (document, gene name, classification) was generated. If the score of all 3 different classifications was less than the threshold then only the classification with the highest score was generated.

4 Results and Discussion

4.1 Ad-hoc Retrieval Task

The results in mean average precision of the ad-hoc retrieval tasks were as follows:

Setting	Without feedback	With feedback
Ad-hoc runs		
Official	0.331	-
Post-hoc runs		
TWM (Title, abstract, and MeSH)	0.331	0.327
TW (Title and abstract)	0.307	0.308
T (Title only)	0.105	0.108
M (MeSH only)	0.079	0.080
W (Most valuable 20 abstract terms)	0.168	0.185

For the official run, our average precision was higher than or equal to the median for 45 out of 50 topics and was the highest for 7 topics out 50. Also, the mean average precision for the official run was identical (to the third significant figure) to that of the equivalent TWM Lemur run. This indicates the collection splitting likely had little effect on retrieval effectiveness. However, blind relevance feedback did not improve retrieval effectiveness for any of the Lemur runs, except for the W run, which is inconsistent with results that we locally obtained on the OHSUMED collection and results reported in the literature. This might indicate that splitting the collection had an adverse effect on blind relevance feedback.

In comparing the TWM and TW runs, the inclusion of the MeSH terms in the documents statistically significantly improved retrieval effectiveness. Statistical significance is indicated if the *p* value of a paired two tailed *t*-test is less than 0.05. This result is consistent with previously reported results in the literature [1, 8].

In comparing the title, abstract, and MeSH fields, the abstract field contributes the most number of valuable for retrieval. This can be clearly seen from comparing the "T" run to the "TW" run and from comparing the "T" run to the "W" run. In both comparisons, we can see that using abstracts or a summary of the abstracts statistically significantly increased retrieval effectiveness over the use of titles or MeSH terms.

4.2 Triage Task

The normalized utility measures for all of our official and unofficial runs are as follows:

Name	Kernel	Threshold	Best* Normalized Utility
Official Runs			
GUClin1260	Linear	-0.973	0.343
GUClin1700	Linear	-1.000	0.385
GUCply1260	Polynomial	-0.894	0.305
GUCply1700	Polynomial	-0.932	0.360
GUCwdply2000	Polynomial	-0.950	0.517

Unofficial Runs			
WD	Polynomial	-0.980	0.551
T	Polynomial	-1.020	0.378
W	Polynomial	-1.020	0.449
M	Polynomial	-1.010	0.414
C	Polynomial	-1.000	0.431
TW	Polynomial	-1.010	0.464
TWM	Polynomial	-1.000	0.546
TWMC	Polynomial	-1.010	0.505
W5	Polynomial	-1.000	0.451
W10	Polynomial	-1.000	0.481

* Best Normalized Utility obtained by adjusting the thresholds for unofficial runs only

The results show that training the classifier using the full length documents yielded the best results. However, using the title, abstract, and MeSH term fields yields comparable results. This suggests that using only these fields, which are a part of MedLine citations, is as effective for document classification as full length journal articles. Due to the fact that clearing copyright issues often complicates obtaining full length documents, this result is significant. Also, captions did not perform as well as previous research suggested [10].

4.3 Annotation Task

The summary of our runs are as follows:

Name	Precision	Recall	F-Measure
GUCbase	0.188	1.000	0.317
GUCsvm0	0.237	0.741	0.360
GUCsvm5	0.205	0.935	0.337
GUCir30	0.221	0.840	0.350
GUCir50	0.230	0.808	0.358

Although runs showed good recall, they generally suffered from poor precision. Further investigation is required to rectify the poor precision.

5 Conclusion

The paper examined the relative effect of using different parts, combinations of parts of the documents, or whole documents on retrieval and classification.

For the adhoc retrieval task, we compared the effect of including and excluding MeSH terms on retrieval effectiveness and showed that the inclusion of MeSH terms statistically significantly improved retrieval effectiveness. Although this result is consistent with the work reported by Srinivasan and Aronson in the literature, contrary to their results blind relevance feedback did not offer any statistically significant improvement in retrieval effectiveness. This could be due to the fact that we split the document collection affecting our document frequencies. From the experiments presented in the paper, indexing a summary of the abstract field of the documents yielded statistically better retrieval effectiveness than the title or MeSH terms fields. Employing different weights to different portions of the documents or skewing term selection in blind relevance feedback can perhaps maximize the effect of more valuable portions of a document and can potentially lead to better retrieval effectiveness.

For the triage sub-task, we compared the use of titles, abstracts, diagram captions, small windows of text around genes and gene products, and combinations of the different portions to the use of whole documents. The use of the combination of the title, abstract, and MeSH term fields yielded results comparable to the use of whole documents.

For the annotation sub-task, we described our experimental procedure and reported the results we obtained. Our results generally suffer from poor precision. We need to investigate methods of improving the annotation task.

6 References

1. Aronson, A., T. Rindflesch. Query Expansion Using the UMLS® Metathesaurus®. Proceedings of AMIA '97 Annual Fall Symposium: 485-9, 1997.

2. Darwish, K. and D. Oard. Probabilistic Structured Query Methods. SIGIR 2003: 338-344, 2003.

3. French J., A. Powell, F. Gey, and N. Perelman. Exploiting a Controlled Vocabulary to Improve Collection Selection and Retrieval Effectiveness. Proceedings of the tenth international conference on Information and knowledge management: 199 – 206, 2001.

4. Hersh w., C. Buckley, T. J. Leone, D. Hickam. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. SIGIR 1994: 192-201

5. KDD Cup
http://www.biostat.wisc.edu/~craven/kddcup/

6. Min Shi, David S. Edwin Rakesh Menon, Lixiang Shen, Jonathan Y.K Lim, Hang Tong Loh. A Machine Learning Approach for the Curation of Biomedical Literature. KDD 2002

7. Robertson S., D. Hull. The TREC-9 Filtering Track Final Report. TREC-9: 25-40, 2000.

8. Srinivasan P. Optimal document-indexing vocabulary for MEDLINE. Information Processing and Management 32(5): 503-514, 1996.

9. YAGI Tool (Yet Another Gene Identification Tool)
http://www.cs.wisc.edu/~bsettles/abner/yagi.html

10. Yizhar Regev and Michal Finkelstein, ClearForest and Celera, Information Extraction from Biomedical Articles. KDD 2002