Experiments on Genomics Ad Hoc Retrieval

Miguel E. Ruiz

School of Informatics
Department of Library and Information Studies
State University of New York at Buffalo
meruiz@buffalo.edu

Abstract

This paper represents te results of the State University of New York at Buffalo (School of Informatics) in the TREC 2005 Conference. We participated in the Genomics ad hoc retrieval task. Our approach used the SMART system for indexing the large collection of MEDLINE documents. For this purpose we used a distributed retrieval approach and divided the large collection into 5 non overlapping sub collections. We tried several approaches on the training topics to select the best run possible. Our results perform slightly above the median system in the conference. We also paired with the NLM team to contribute a run for their fusion approach.

Keywords: Genomics; Information Retrieval; Distributed Information Retrieval; Vector Space Models.

1 Introduction

The SMART system created by Gerald Salton (1971) and his collaborators has been a classical tool for research groups to perform information retrieval experiments in the last 30 years. For this year we decided to use it for indexing and retrieving the large collection of MEDLINE documents for the genomics ad hoc retrieval task. The main challenge was to deal with the limitation of the current version of Smart publicly available.

Session 2 presents a description of the data for the genomics ad hoc retrieval task, section 3 presents details of the processing of documents and query as well as the system set up. Section 3 presents our results and a brief analysis. Section 5 presets our conclusion.

2 Genomics Ad hoc Retrieval Task

The Genomics ad hoc retrieval task has been designed to explore realistic information needs proposed by scientists who work in the area of genomics (Hersh, 2005). This year one of the main changes with respect to other ad hoc retrieval tasks is that the topics are structured according to 5 different templates:

- Standard methods or protocols for doing some sort of experiment or procedure.
- Role(s) of a gene involved in a disease.
- Role of a gene in a specific biological process.
- Interactions between two or more genes in the function of an organ or in a disease.
- Mutations of a given gene and its biological impact or role.

The document collection contains 4,591,008 MEDLINE records which is a 10-year subset of the actual MEDLINE database. The major problem that we faced when trying to index this large document set was that the indexes created by Smart were larger than 2 GB and hence the indexing process will fail due to system restriction on the maximum size of a direct access file. To solve the problem we decided to divide the collection into smaller sub-collections so that the index files produced by smart would be smaller than 2 GB. For this purpose we decided the collection in 4 subsets of 918,202 and one last subset of 918,200 documents. The main idea was to have roughly equal sub-collections. Once we have these 5 sub collections, the problem was transformed into a distributed retrieval task with non-overlapping homogeneous collections.

2.1 Document and query processing

For processing the documents we converted the original MEDLINE records into a simple smart document format as presented in Table 1. We used only information coming from Title, Abstract, R field and MeSH Terms. Since each of these fields will be processed differently we created a representation that stores the information in four indexes or ctypes. This means that we are using a generalized vector space that has four dimensions (text from title and abstract, chemical compounds, MeSH terms and extracted bigrams). Text from title and abstract was indexed together in a single dimension (ctype 0) and the text was passed through a stemming algorithm that takes care of the plurals (remove_s), and stop words from a revised list were discarded. The R field contains terms that are usually enzymes or other chemical names. These terms were not stemmed since they are usually assigned as controlled vocabulary which includes not only names but also codes for substances. MeSH terms were represented in a separate index (ctype 2) and they were represented by the stems of the words in the term, a non-stemmed terms that concatenated all words in the terms, and word bigrams generated from each term. The main idea was to allow partial and exact matching of the MeSH terms. We also took into account the indication of whether a terms was considered a main term in the document (these terms appear with a "*" in the original MEDLINE record). Main MeSH terms were repeated three times in the record to increase term frequency in the document (this is equivalent to boosting its importance with respect to other terms that are not considered main terms). Finally a simple algorithm that extracts word bigrams that do not include stop words was used to generate the terms for ctype 3.

Table 1. Mapping from MEDLINE to Smart simple document format.

MEDLINE Tag	Smart tag	Description	Smart ctype
PMID	.I	Document ID	----
TI	.T	Title	Ctype 0
AB	.W	Abstract	Ctype 0
SO	.S	Publication details	Ctype 0
RN	.R	Enzyme or chemical substances	Ctype 1
MH	.M	MeSH headings	Ctype 2
----	.B	Word bigrams from title and abstract	Ctype 3

The original query file (a MS Word document file) was converted to a XML format that included the fields that appear in the corresponding columns of the original topics. These XML file was then processed using a perl script that converted it to a format compatible with the document representation previously described. We also generated word bigrams for all the fields in the query.

Similarity between Query and documents was computed by using a linear combination of the scores obtained from each ctype as specified by the following formula:

$$rsv(q,d) = \sum_{i=0}^{3} \delta_i \times sim_i(q,d)$$

Where *rsv(q,d)* represents the final retrieval status value of document *d* for query *q* , δ_i is the a factor that weights the contribution of the corresponding *ctype i* in the retrieval score, and $sim_i(q,d)$ is the value computed by the smart retrieval system for the corresponding query *q* and document *d* for ctype *i*.

We use the 10 training topics to tune the parameters for the contribution of each ctype. Our experiments on the training queries show that the value of the weights for each ctype that optimized the MAP for the training topic are:

- Title and Abstract (Ctype 0): 7
- Chemical compounds (Ctype 1): 1
- MeSH Terms (Ctype 2): 2
- Bigrams (Ctype 3): 1

We also tried several weighting schemes for documents (*atn, Lnu, atc*) and for queries (*nnc, ntn, ntc, ltc, lnc, atc, atn, ann, anc, ltu*) for a total of 30 combinations of document-query weight (i.e. *Lnu.ltu*). The retrieval score was computed for each of the sub collection and the final set of retrieved documents for the whole collection was selected by combining the top 1000 results from each of the five sub collections and sorting them by final rsv score. The best results were obtained using atn.ann weighting scheme with the corresponding weights for each of the ctypes. Our baseline run using these setting obtained a MAP of 0.1713 on the 10 training queries.

2.1.1 Query expansion

We decide to explore query expansion at two levels: pre retrieval expansion and automatic retrieval feedback expansion. The pre retrieval expansion method used expansion of gene names using UMLS. These expanded gene names were generated by Susan Humphrey at the National library of Medicine using pattern matching between the terms present in the query and the gene names available in UMLS (Aronson *et al.*, 2004). These gene expansions were made available to the groups collaborating in the NLM runs. We also tried expansion of training topics by preprocessing the queries using MetaMap (Aronson, 2001). Table 2 shows the summary of our results using pre retrieval expansion.

Table 2 Results on training queries

	MAP
Baseline	0.1713
Gene-expanded	0.1700
MataMap Expansion	0.1687
Distributed Pseudo relevance feedback of gene- expanded	0.1703

One of the most surprising findings for us was that none of our attempts to expand the query (either before retrieval or using a retrieval feedback mechanism) were able to outperform our base line. Actually when we compared our base line with other runs of teams that were working with the NLM we found that our base line was quite good and only a couple of teams were able to perform better than our base line. We believe that this could be explained in part by the small number of queries and to the fact that a few of those queries had very few relevant documents and were already to hard. For the queries that had plenty of relevant document our base line was performing quite well and was hard to do much better using just query expansion.

We also tried a pseudo relevance feedback expansion mechanism. Since we were using a distributed collection (with 5 sub collections) we decided to do the pseudo relevance feedback as a process that will select the best terms to be

used to expand as a global process. In other words, we extracted all terms from each sub collection with their respective idf and weighting and then computed a common idf and selected the top 10 terms at a global level to expand each ctype. Then the terms were sent back to each collection for mapping in the respective sub collection and running the corresponding expanded query. This proved to be a very complicated process that at the end did not improve performance.

2.3 Ad hoc results

Due to lack of time we only submitted a single run that used a pre expansion using gene names. Our run is labeled UBIgenA and achieved a MAP of 0.2262 and a bpref of 0.7040. When compared with the median system it actually performs slightly above to the median system. A query by query analysis of the system shows that our official run performs above the median in 25 queries and achieves the best result in 1 query. Although not a spectacular performance it actually shows that using a distributed collection does not reduce your performance significantly. We still need to do some more analysis on the relevance feedback to check whether it has the same relative performance with respect to the base line.

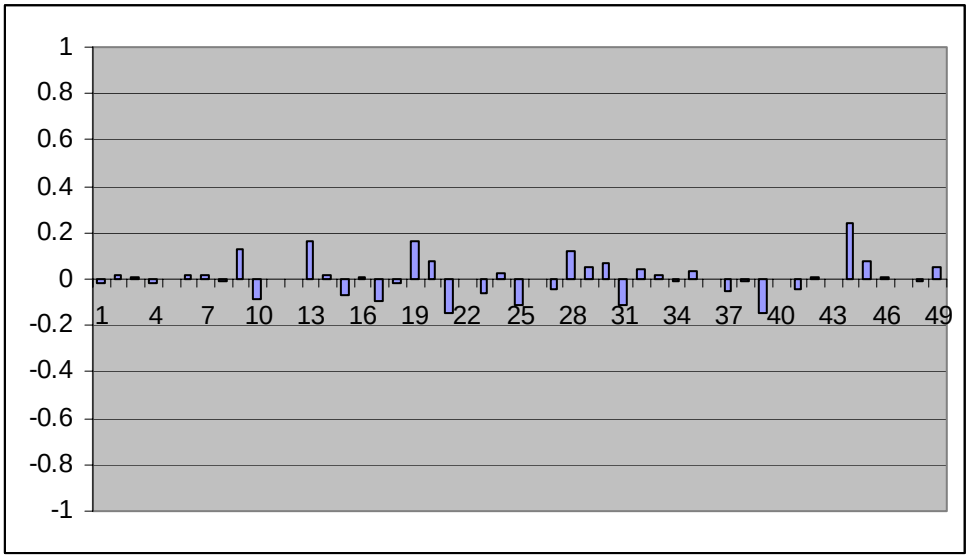

Figure 1. Query by query comparison with median system

4 Conclusions and future research

Our results in the ad hoc retrieval confirm that using a distributed index for a large collection still can achieve results that are similar to using a single index. We still need to explore the issues related to query expansion in this domain to find a way to improve results.

We plan to further explore query expansion using several methods to find whether there is a specific template that would benefit by using either pre retrieval expansion or pseudo relevance feedback.

Acknowledgements

We would like to acknowledge the support from the National Library of Medicine since part of this work was developed as a visiting researcher during the summer at the Lister Hill National Center for Biomedical Communications. We also want to thank Susan Humphrey from NLM for providing us with the gene expansion used in this work and to Alan Aronson from NLM for providing MetaMap. Thanks also to Patrick Rusch (University Hospital of Geneva) for his continuous feedback while we were working as visitor researchers in NLM.

References

Aronson A.R. (2001) “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.” *Proc AMIA Symp*., 17-21.

Aronson A.R., Demner D., Humphrey S.M., Ide N.C., Kim W., Liu H., Loane R.R., Mork J.G., Smith L.H., Tanabe L.K., Wilbur, W.J. and Xie N. (2004) “Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLNE citations.” *The Thirteenth Text Retrieval Conference, TREC-2004*, Gaithersburg, MD.

Fox E.A. and Shaw J.A. (1994). Combination of multiple searches. In *Proceedings TREC-2*, (pp. 243-249). Gaithersburg: NIST Publication #500-215.

Hersh, W. (2005). TREC 2005 Genomics Track Protocol. http://ir.ohsu.edu/genomics/2005protocol.html

Ruiz, M.E. (2005) Experiments on Genomics ad hoc Retrieval. In *Proceedings of the Fourteen Text Retrieval Conference TREC 2005*. Gaithersburg, MD.

Salton, G (1971) *The SMART Retrieval System - Experiments in automatic Document Processing*. NJ: Prentice Hall.

Singhal A., Buckley C. and Mitra. M. (1996) Pivoted document length normalization. *SIGIR 1996*, 21-29.