Genomics Retrieval Experiments at UTA

Ari Pirkola
University of Tampere (UTA), Finland
Department of Information Studies
pirkola@cc.jyu.fi

1. Introduction

University of Tampere submitted runs for Genomics Track ad hoc retrieval task. The first run (*uta05a*) was an automatic and the second (*uta05i*) an interactive run. The *uta05a* queries were constructed by using the original topic terms as query keys. The *uta05a* queries served as a baseline for the *uta05i* queries which were constructed by expanding the *uta05a* queries with synonyms for the topic gene names from the Entrez Gene database and by further expanding with synonymous gene names and MH terms from the top documents of an initial *uta05i* search. The mean average precision for *uta05a* was 0.2385 and for *uta05i* 0.1980.

Next in Section 2.1 we present the indexing methods and the processing of query keys. Section 2.2 considers the retrieval system and query operators. Query formulation is presented in Section 2.3. Section 3 contains the results and conclusions.

2. Methods and data

2.1 Indexing and the processing of query keys

We used the following approaches and techniques in indexing and for query keys:

- Genomics Track test collection for the ad hoc task was a subset of the Medline collection. Each record consists of several fields. We indexed the fields TI (title), AB (abstract), MH (MeSH headings), RN (Registry Number), and GS (Gene Symbol)
- Letters were normalized to lower case.
- Query keys and the words of documents were normalized using the morphological analyzer *Kstem*, which is part of InQuery retrieval system.
- Only letters (a-z) and numbers (0-9) were indexed. Hyphens and other characters in strings than letters and numbers were replaced by a space, and were not searchable.
- Strings containing both letters and numbers were decomposed into separate alphabetical and numerical strings. In searching the string elements were linked to each other by means of a proximity operator (Section 2.2).

2.2 Retrieval system and query operators used in the tests

The test system was the *InQuery* retrieval system (Allan et al., 2000; Larkey et al., 2005). InQuery provides a variety of query operators, including the Boolean conjunction operator *#band* which formed the basis of our both runs. For the #band-operator all its argument keys must occur in a document in order for the operator to contribute to the weight computed for that document. Otherwise #band contributes to the document score like the *#and-operator*. The weight of the #and-operator is computed as the product of the weights of its arguments. For the *#sum-operator*, the system computes an average of key weights.

Except for the first 10 topics (topic numbers 100-109) Genomics Track 2005 topics were organized in columns with each column representing one aspect of a topic (Hersh, 2005). In the *uta05a* queries the keys were grouped into #sum-subqueries with each subquery representing one column in a topic. For *uta05i*, subqueries were constructed based both on the topic structure and an intellectual analysis. For both query types different subqueries were combined to each other by the #band-operator. In the first 10 queries keys were linked to each other with the #sum-operator.

Phrases were searched for using the proximity operators of *#odn* (ordered window operator) and *#uwn* (unordered window operator). N was set at k+1 (*uta05a*) and k+2 (*uta05i*) where k=the number of elements in a phrase.

2.3 Query formulation

Next we describe the *uta05a* and *uta05i* query formulation and present example queries.

uta05a

Uta05a queries were Boolean queries based on the column structure of the topics (Section 2.2). Below we present an example of Genomics Track topic (# 112) and an *uta05a* query derived from it:

Gene(s)	Disease
IDE gene	Alzheimer's Disease

#band(#sum(ide gene) #sum(alzheimer disease))

Strings (gene names) containing both letters and numbers were decomposed as described in Section 2.1, and in searching the string elements were combined by means of the proximity operator of #odn. For example, for the gene name *MMS2* the proximity statement was as *#od3(mms 2).*

uta05i

The aim of the *uta05i* run was to investigate the effects of gene name synonym and MH term expansion on retrieval performance. We compare the performance of *uta05i* to that of *uta05a*. *Uta05i* query construction proceeded as follows:

Synonymous gene names for the topic gene names were retrieved from the Entrez Gene database. *Uta05a* queries were expanded with the collected synonyms, and the expanded initial *uta05i* queries were run on the test collection. Final *uta05i* queries were formulated by further expanding the queries with MH terms (including subheadings) and synonyms found in the top 20 documents of an initial *uta05i* search.

An example *uta05i* query is presented below (query # 112):

#band(#sum(#od4(ide gene) ide #od5(insulin degrading enzyme) insulinase insulysin)
#sum(#uw5(alzheimer disease genetics) #uw4(alzheimer disease)))

If the first query gave no results the #and-operator was used instead of the #band-operator (this concerns both *uta05a* and *uta05i*).

3. Results and conclusions

Table 1. Retrieval performance of the *uta05a* and *uta05i* queries.

Query type	MAP
uta05a	0.2385
uta05i	0.1980

The results of the two runs are presented in Table 1. As can be seen, for *uta05a* MAP is 0.2385 and for *uta05i* 0.1980. The results are unexpected. We expected that interactive retrieval (*uta05i*) would have given better results than automatic retrieval (*uta05a*). We will analyze the results, focusing on the question why gene name synonym and/or MH term expansion did not contribute to better retrieval performance.

Acknowledgements

This research is part of the research project *NLP-based information retrieval systems for the biological literature* funded by the Academy of Finland (funding decision 206568).

The InQuery search engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts.

References

Allan, J., Connell, M.E., Croft, W.B., Feng, F.-F, Fisher, D. and Li, X. 2000. Inquery and TREC-9. The Ninth Text REtrieval Conference (TREC-9), Gaithesburg, MD. Available at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html

Hersh, W. 2005. TREC 2005 genomics track protocol. http://ir.ohsu.edu/genomics/2005protocol.html

Larkey, L.S. and Connell, M.E. 2005. Structured queries, language modeling, and relevance modeling in cross-language information retrieval. Information Processing & Management, 41(3), 457-473.