Interactive Retrieval Using Weights

Jonathan Schuman and Sabine Bergler
Department of Computer Science and Software Engineering
Concordia University
1455 de Maisonneuve Blvd West
Montreal, Quebec, H3G 1M8

Abstract

Using the same interactive IR component as for TREC 2006, this submission probed the ability of a user without requisite domain knowledge to interactively set appropriate weights. The weighted keyword proximity method employed allowed a computer science undergraduate to achieve rankings above the median for all measures without the use of external knowledge sources or term expansion techniques but in an interactive fashion. This suggests that domain experts should be able to perform above the median reliably, and are expected to excel when they can use domain terminology to their advantage.

1 Motivation

BioKI is a system designed for domain experts to formulate queries for literature retrieval. The expert has the option to associate weights with the keywords and is given different contexts to view the results, in order to easily refine the query terms and weights in an interactive fashion. The system targets in particular low-frequency information, non-redundant information that may occur anywhere in a journal and that may not be present elsewhere in the document nor in other articles.

The underlying presumptions are:

1. a fully automated system will not be able to expand the query keywords beyond simple synonym and acronym extension, because redundancy-based techniques such as relevance feedback discriminate against such low-frequency information
2. an expert who is hunting for low-frequency information has considerable domain expertise and in particular, knows appropriate technical terms for efficient retrieval
3. the expert cannot, however, reliably predict whether the chosen keywords select the information sought, a quick review of the returned passages will be required to interactively reformulate the query in case of undesired results (red herrings on one side, missing specifiers on the other)
4. usually only the top 20 results are relevant to the user, who will rarely review more than the first screen.

TREC Genomics fits these presumptions only partially. While it is open to interactive systems and the queries target both, redundant and low frequency information, the fact that up to 1000 results may be submitted and that judges only get to judge among the pooled results of the most frequently returned passages will preclude a brilliant outlier passage from entering the pool. It is important to keep in mind that the pooled results measure conformity as well as accuracy, and that recall is only calculated with respect to an ill understood subset of the corpus. We still feel, however, that the TREC Genomics data and retrieval task are informative to assess interactive retrieval including low-frequency information.

2 BioKI for TREC Genomics 2007

The IR system used for our three runs in 2007 was identical to the one used in 2006. Last year's assessment (Bergler et al., 2006) reports numbers that were not representative of the system performance — we discovered a major flaw in the routine that maps retrieved passages to byte offsets. The same passages reported for TREC Genomics 2006 yielded scores as reported in Table 3. Out of 92 runs, the rescored system would have obtained rank 20 for the document retrieval task, rank 6 for passage retrieval, and rank 18 for aspect retrieval.

The results of the 2006 exercise were obtained with queries that had been refined based on the 2005 data and evaluation script, that is the developer had access to the MAP score of the query on 2005 data, which consisted of only abstracts, and thus was not identical to the full articles used in 2006, but still presented a clear advantage to an information seeker using the system without such reference material. For 2007, the task queries were changed significantly: not only do they concern different topics, but the format, too was changed. Some entity *types* were specified for list questions, as in Topic <227>: *What [GENES] are induced by LPS in diabetic mice?* Here the type [Genes] is supposed to match with all genes. Our lab does not have the capability to gather all possible [BIOLOGICAL SUBSTANCES] for instance, and thus types were ignored in the queries for two BioKI runs. For comparison, we submitted one run that attempted gazetteer-based typing of a few of the categories.

Running the same system under these new conditions allows for the relative assessment of strong and weak points. In particular, we tested whether a domain novice can in fact achieve a respectable result with a moderate amount of iterations, which we posit as a proxy to the usefulness of the BioKI design to a domain expert, who would be able to define even better targeted queries.

2.1 Data and Preprocessing

In addition to the preprocessing reported in (Bergler et al., 2006), five of the topic entity types were annotated in the corpus, using gazetteer lists compiled from the following sources:

DISEASES National Institute of Neurological Disorders and Stroke[1]

DRUGS DrugBank[2] (Approved & Experimental DrugCards – January 25, 2007)

MOLECULAR FUNCTIONS Gene Ontology[3]

MUTATIONS MutationFinder [4]

SIGNS OR SYMPTOMS MeSH[5]

2.2 Scoring Method

BioKI assigns scores to paragraphs based on the keyphrases and weights supplied by the user. The scoring method used for TREC Genomics 2006 and 2007 is based on the following principles:

- the closer the keyphrases occur together, the higher the rank
- the more keyphrases are found in the text segment, the higher the rank

The main components of this scoring measure are thus (weighted) keyphrase coverage and keyphrase proximity. The function is similar to one used in (Lawrence and Giles, 1998), which considers the number of keyphrases found, their proximity, and their frequency. In early experiments, we observed that scoring multiple occurrences of the same keyphrase lead to less relevant rankings, and so term frequency is not considered in our scoring function. Also, to allow for unequal weighting of keyphrases, coverage is calculated relative to assigned weights. The scoring function is calculated as $T(1+p)(2c)$, where $p = \frac{t-w}{t}$ and $c = \frac{\sum_{\text{matched}} k_i}{\sum_{\text{queried}} k_i}$. T is a scoring threshold, p is a proximity factor, and c is a relative keyphrase coverage factor. The proximity and coverage factors are defined in terms of w, the number of characters in the smallest span containing all matched terms; t, the number of characters in the entire tile; and k_i the weight of the i^{th} keyphrase.

[1] http://ninds.nih.gov/disorders/
[2] http://redpoll.pharmacy.ualberta.ca/drugbank/
[3] http://www.geneontology.org/
[4] http://mutationfinder.sourceforge.net/
[5] http://www.nlm.nih.gov/mesh/

2.3 Query Formulation

Queries were formulated by selecting the main noun phrases in each topic as keyphrases. Whenever possible, named entities included in the topic that could serve as a discriminating factor (e.g. organisms, diseases) were weighted more heavily than other keyphrases.

Queries for each topic were iteratively refined by examining the preliminary returned passages. Passages were examined to assess the relative importance of keyphrases and effect corresponding weight changes, and for occurrence of additional terms that could be added to the query[6]. On average, topics were finalized after 3.75 iterations.

For the latter task, weighting was heavily leveraged for term 'fishing'. Terms of questionable benefit to the query, either from the topic text itself or found within preliminary passages, could be easily assessed by assigning them low weights and observing the context in which they occur in the subsequent iteration. For example, in the query for topic <214> (see Figure 1) 'growth cone' and the variants for 'guidance' were found in only a few passages during early iterations; but adding them to the query with low weights showed more of their occurrences, allowing for quick confirmation of their suitability to the query.

Topic 227: *What [GENES] are induced by LPS in diabetic mice?*
"induced::10" "LPS|lipopolysaccharide::60" "diabetic::40" "mice|mouse::10"

Topic 214: *What [GENES] are involved axon guidance in C.elegans?*
"axonal|growth+cone::20" "guidance|guiding|pathfinding|navigation::10" "C.elegans|C.+elegans|Caenorhabditis+elegans::100"

Table 1: Weighted keywords for topics 227 and 214

Aside from term 'fishing' and expansion, in each iteration weights were modified in order to refine the ranking of returned passages, and to eliminate irrelevant passages at the lower end. This was achieved primarily by establishing a partial ordering of keyphrase relevance, and attempting to suppress passages that did not have the most relevant terms. For example, in the query for topic <227> (see Figure 1) 'LPS' is specified as the most discriminating term, followed closely by 'diabetic'. Passages containing only these two terms rank quite highly (just below passages containing all terms, if any), while passages omitting one of these terms but containing all others are unlikely to rank within the top 1000 submitted passages. Further, passages omitting both 'LPS' and 'diabetic' score below the internal threshold and are not nominated at all.

This aspect of the iterative process revealed shortcomings in the expressibility of the weighting method, where undesirable keyphrase combinations could be pushed down in ranking, but not completely suppressed. Looking again at the query for topic <227>, passages containing only 'induced' and 'LPS' were observed to be predominantly irrelevant in early iterations. While some weight refinement could be used to disfavor such passages, no combination of weights is possible that would score these below the internal threshold while maintaining sensible top ranks, given the current scoring function. This issue cropped up in several of the topics, where it seems something akin to a weighted Boolean search may be better suited to trimming off irrelevant passages from the bottom ranks. This however, we view as a TREC-specific issue, as in BioKI's native task, users would never view such low ranked passages.

For both biokiP and biokiS runs, the same queries were used; the only difference between the runs was span delimitation during postprocessing. The shortest span containing all occurring keyphrases was returned for biokiS, while the same span was expanded to paragraph boundaries for biokiP. Our third run, biokiST, also used essentially the same queries. The only difference being the eight topics that had entity types annotated in the corpus (201, 202, 205, 210, 219, 222, 229, 232). For each of these topics, an additional meta-keyphrase, matching any term annotated for the given entity type, was added to the query. No changes were made to the weights for biokiST.

[6]Only terms found within preliminary passages were used for term expansion – no external sources were used.

Table 2: Ranks of TREC Genomics 2007 runs

	Passage2	Aspect	Document
biokiST	16	13	21
biokiS	18	12	20
biokiP	33	3	19

3 Results

Our three runs fared similarly. Predictably, biokiP outranked the other two for the aspect score, since it includes the most words, and thus aspect triggers. It also obtains the best document score, but since our three runs placed in successive ranks on the document score, this is not indicative. biokiST predictably obtains the highest rank for Passage2. The respective ranks are summarized in Table 2.

Our runs placed in the top third, with only the biokiP rank below average. Overall, the scores represent a small loss in performance relative to the field compared to last year's adjusted scores. The two basic reasons for this drop are firstly, the developer had training data in the 2005 results for the 2006 queries, but had to revise the queries purely based on reviewing the returned results for the 2007 data. Not being a domain expert was a greater handicap in 2007.

Secondly, the TREC Genomics 2007 query format is less suitable for our approach. While the queries for 2005 and 2006 were regular questions that pinpointed entities and relationships explicitly, the list questions of 2007 replaced an explicit entity with a generic term that is not suitable as a keyword. Since we only expanded 8 types with partial lists of possible tokens, this left many queries with a lack of focus. Topic <200>*What serum [PROTEINS] change expression in association with high disease activity in lupus?*, for instance, was expanded to the query *"serum::10" "expression::10" "lupus—SLE::61"*, which yielded a score well below the median. The gist of the question, change and high activity, is not expressible in keywords and demonstrates thus the limits of the approach. Similarly, for topic <216>*What [GENES] regulate puberty in humans?*, was expanded to the query *"regulate::10" "puberty::50" "humans::10"*, which performed below the median except for the passage score.

Interestingly, some of the topics where BioKI performed very well, did not have many named entities, but had common nouns that served well as keywords. BioKI is (among) the top performing system(s) for topics <225>*"induce::10" "clpQ—hslV::100" "expression::10"*, <232>*"inhibit::10" "HIV+type+1|HIV+1::50" "infection::10"*, and <205>*"[SIGNSORSYMPTOMS]::10" "anxiety::10" "coronary+artery+disease|CAD|coronary +heart+disease|CHD|coronary::61"*, which include at most one named entity and rely heavily on supporting *explicitly expressed* common nouns (such as *inhibit*, *coronary*, *induce*).

3.1 TREC Genomics 2006

Our official submissions for 2006 were converted from BioKI's native output format (the actual text of the nominated passage) to byte offsets by mapping passage text into a secondary corpus provided to all TREC Genomics participants by Martijn Schuemie, a fellow TREC participant. This secondary corpus came preprocessed, and with annotations for each sentence specifying the start and end offsets of the sentence within the original corpus. Unfortunately, we did not anticipate differences between BioKI's internal preprocessing and that used to generate the secondary corpus. As a result many text passages retrieved by BioKI could not be mapped into the corpus, incurring a loss of about 10%. Also, an unknown number of passages were mapped into incorrect byte offsets.

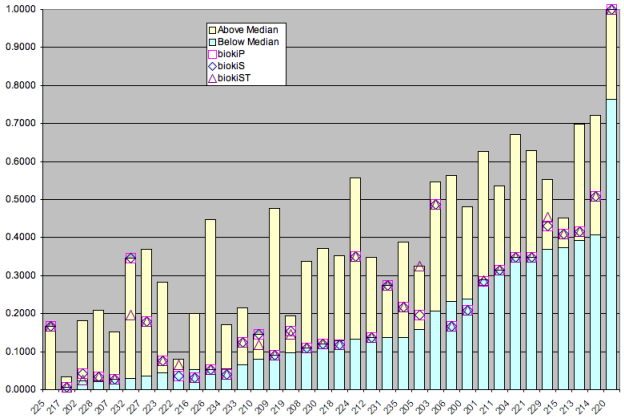
We have recently rebuilt the postprocessing stage using our own secondary corpus to map into byte offsets without incurring loss or inaccuracy. The secondary corpus contains the same preprocessed text over which the system runs, but maintains byte offset information for each word. Taking the same native BioKI output used to generate our official 2006 submissions, we have rerun the postprocessing and evaluations. Note that no changes were made to any IR components, nor were any parameters changed. The native format output files from our official submission were used as is, with only changes in postprocessing. The results of the rescoring exercise are significantly better than the officially scored results, they are reported in Table 3.

Table 3: Adjusted Scores for TREC Genomics 2006

	Document	Passage	Aspect
Adjusted	0.3794	0.1162	0.2031
Reported (1)	0.3072	0.0419	0.2171
Reported (2)	0.3093	0.0335	0.2537

3.2 Document Retrieval Task

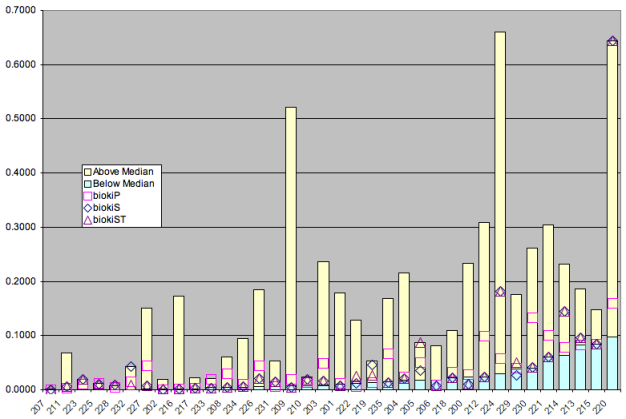
For the document retrieval task, every returned passage is mapped to the document that it occurs in, subsequent passages from the same document are not scored again. BioKI scores above the median throughout. BiokiP and BiokiS are identical, BiokiST outperforms biokiS only twice, for topics <205> and <229>.



3.3 Passage Retrieval Task

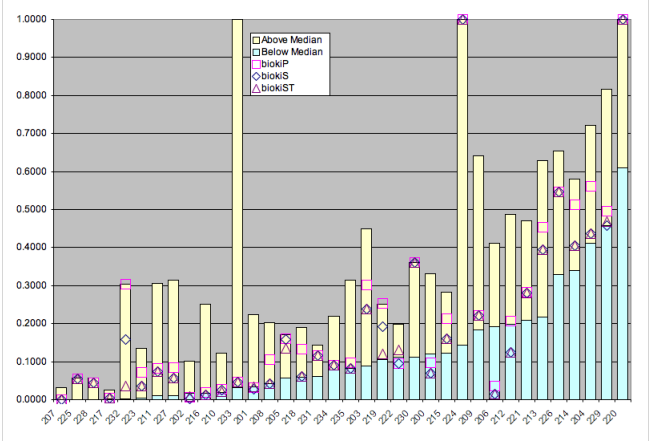
The passage retrieval scoring function changed from TREC Genomics 2006 to 2007. The original passage scoring function computed character overlap between returned passages and the gold standard, penalizing systems for extraneous text returned. Because of this measure's susceptibility to changes from non-content manipulations, for 2007 it was changed to the Passage2 score, where avarage precision is computed at each correct character of the returned passage. In comparison, the scores for BioKI are lower for the Passage2 measure than for the 2006 Passage measure.

Again, BioKI closely follows the median, with a few top performances. We observe a difference for biokiP and biokiS, since the paragraphs returned by biokiP have potentially a lot more characters than the span returned by biokiS. The difference is, however, very small.



3.4 Aspect Retrieval Task

Aspect retrieval was the surprise outcome of TREC Genomics 2006. The task is to cover as many different MeSH term aspects in the returned passages as possible. This measure is of particular interest to us, since it approximates the chance a user of an interactive system has to find inspiration in the top returned passages for reformulating the query. BiokiP, returning paragraphs and thus more text that can express different aspects, performs better than biokiS, which returns only the keyword span.



BioKI's general weighted keyword retrieval performs well on the aspect score without any fine-tuning for aspect retrieval.

3.5 Typed Queries

The performance of biokiST on the topics for which entity types were annotated is compared to biokiS in Table 4. There seem to be no regularities in the data that allow for a general assesment of these queries.

Table 4: Comparison of typed and untyped queries.

	Document		Passage2		Aspect	
Topic	biokiST	biokiS	biokiST	biokiS	biokiST	biokiS
201	0.2880	0.2840	0.0074	0.0068	0.0331	0.0283
202	0.0243	0.0426	0.0016	0.0005	0.0087	0.0021
205	0.3253	0.1955	0.0872	0.0360	0.1347	0.1591
210	0.1172	0.1461	0.0194	0.0185	0.0220	0.0262
219	0.1460	0.1537	0.0263	0.0463	0.1201	0.1912
222	0.0655	0.0357	0.0256	0.0116	0.1313	0.0952
229	0.4528	0.4296	0.0505	0.0262	0.4694	0.4569
232	0.1952	0.3466	0.0084	0.0429	0.0368	0.1583

4 Conclusion

The results support two of our core assumptions. Firstly, that the user of a generic interactive system can achieve comparable results to more specialized and more automated systems using only keywords and weights (with the possibility that domain experts will be able to obtain better results faster). Secondly, this simple approach performed close to the median even when indicative keywords were not named entities, but common nouns and verbs and no external resources were consulted. Weighted keywords in interactive systems are thus a viable alternative for expert users for a range of different query and information types.

References

S. Bergler, , J. Schuman, J. Dubuc, and A. Lebedev. 2006. Bioki, a general literature navigation system at trec genomics 2006. In *Proceedings of the 15th Text Retrieval Conference (TREC 2006)*, Gaithersburg.

S. Lawrence and C. Lee Giles. 1998. Inquirus, the NECi meta search engine. In *Seventh International World Wide Web Conference*.