Motivation

Many text search tools rely on approaches to expand queries typed by the user and increase the recall of text searches. Because of the rich dictionary used in the biomedical literature, traditional approaches such as stemming are of limited use. We have developed a new solution to this problem for our biomedical search engine Twease.org.

Approach

Our approach consists of five main steps:

1. Corpus dictionary analysis
2. Estimation of a statistical stemming model
3. Parse each query words into prefix/potential stem/suffix
4. Score each word similar to a query word with the stemming model
5. Rank word variants by score

Impact on text searches

We have compared the recall of text searches when queries are expanded with morphological word variants, and are not. For this evaluation, we have used a protein interaction dataset collected by Albert and colleagues by mining the literature [1]. For each of the 2,789 interactions in this dataset, we created queries such as:

"SMRT" "HDAC3" (affinity | assemble | associate | attach | bind | complex | contact | couple | dimerize | dissociate | dock | interact | link | precipitate)

Each interaction lists one or more articles that describe the interaction. In our evaluation, we consider these articles relevant to the query generated from the interaction. Using these relevance judgements, we calculate precision and recall at each rank in the results returned by the search engine (result returned top has rank 1, second rank 2, and so on).

References


For more information

We distribute the implementation of our method under the Gnu General Public License to maximize its use in the biomedical community. http://icb.med.cornell.edu/crt/twease/ See also the Twease search engine at Twease.org