Refining the Extraction of Relevant Documents from Biomedical Literature to Create a Corpus for Pathway Text Mining

Rachel Harte, Yan Lu, Stephen Osborn, David Dehoney and Daniel Chin
PPD Discovery, Inc., 1505 O'Brien Drive, Menlo Park. CA. 94025.
{harera, yl8lu}@hotmail.com, osbornsg, david_dehoney, danieljchin}@yahoo.com

Abstract

For biologists to keep up with developments in their field or related fields, automation is desirable to more efficiently read and interpret a rapidly growing literature. Identification of proteins or genes and their interactions can facilitate the mapping of canonical or evolving pathways from the literature. In order to mine such data, we developed procedures and tools to pre-qualify documents for further analysis. Initially, a corpus of documents for proteins of interest was built using alternate symbols from LocusLink and the Stanford SOURCE as MEDLINE search terms. The query was refined using the optimum keywords together with MeSH terms combined in a Boolean query to minimize false positives. The document space was examined using a strategy employing Latent Semantic Indexing (LSI), which uses Entrez’s “related papers” utility for MEDLINE. Documents’ relationships were visualized using an undirected graph and scored by their relatedness. Distinct document clusters, formed by the most highly connected related papers, are mostly composed of abstracts relating to one aspect of research. This feature was used to filter irrelevant abstracts, which resulted in a reduction in corpus size of 10% to 30% depending on the domain. The excluded documents were examined to confirm their lack of relevance. Corpora consisted of the most relevant documents thus reducing the number of false positives and irrelevant examples in the training set for pathway mapping. Documents were tagged, using a modified version of GATE2, with terms based on GO for rule induction using RAPIER.

1. Introduction

For biologists trying to keep up with the biomedical literature, there is an overwhelming amount of information to read and digest. There are currently over 12 million citations in MEDLINE [1] so it is imperative to create tools to aid the retrieval of relevant articles for a particular domain. We have applied a tool based on LSI for retrieval of relevant documents to create a corpus. Natural Language Processing (NLP) techniques have previously shown promise in the extraction of biological relationships to literature [2]. Subtle relationships may be identified that even a domain expert may not infer even if they were able to read all of the relevant literature. In this work, we created a method for building a focused corpus and used NLP to identify gene/protein interactions for pathway building. Such text mining results may be used to aid biologists in the prioritization of drug targets.

2. Methods

For a gene of interest, alternate symbols and full name synonyms were found in LocusLink at NCBI and in the Stanford SOURCE. Entrez’s “related papers” utility for MEDLINE [3], which is based on LSI, was modified to sort documents according to relevance and to visualize document relatedness. This aided the production of a corpus of documents for further text mining. A list of terms used to describe relationships was built by biological scientists using terms from the Gene Ontology (GO) [4]. This included terms representing all three categories of GO: biological process, molecular function and cellular component. Using a modified version of GATE2 [5], domain experts tagged abstracts for gene/protein relationships. Tagged documents were then used by RAPIER which applies machine learning for rule induction. These rules may then be utilized in pathway extraction from literature.

3. Results

A gene for an adhesion protein of interest was chosen. Searching MEDLINE with the gene name synonyms was refined until the minimum number were used with the least irrelevant documents being returned after inspection of search results. After querying the LSI-based tool with these keywords, 501 documents were retrieved from MEDLINE together with relevant “related” documents. Four of these “related” documents were highly connected to the other documents and formed clusters (Figure 1) that could be used to filter out irrelevant documents. Cluster r151 contained 159 documents of which half are clinical or discuss chromosome mapping and thus not useful in pathway building. To remove these abstracts, certain MeSH terms were used e.g. “Chromosome Mapping” and clinical terms which vary depending on the gene of
interest. Other keywords extracted from cluster r151 were also used to filter abstracts by title. When applied to the original query set, filtering resulted in a set of 395 abstracts.

Figure 1. Visualization of document space using the LSI-based tool. Each document is represented as a numbered node. Edges connect papers that are related. Papers related to those found by the original query are marked with "r".

Relevance scores calculated by the LSI-based tool determined a threshold below which documents were irrelevant for pathway building. This was determined by plotting the number of documents found using a score threshold of 500 to 3500 (Figure 2). All documents were collected when a threshold of 2500 was implemented (n2.5k set) and these were merged with the 395 documents collected previously.

Documents removed by filtering from the original query set were also removed from the n2.5k set and these two sets were merged. When this merged set was compared to the set of 395, it was discovered that 70 documents were unique to the latter set. Examination of the 70 unique abstracts revealed that 15 were significant and all of these had low related-document scores. Many of these were older publications, published in obscure journals or publications too new to have been referenced yet.

This resulted in a focused corpus for subsequent tagging using GATE2 followed by rule induction by RAPIER [6].

5. Conclusion

Using keyword filtering and a LSI tool we were able to optimize the retrieval of relevant documents for text mining. This resulted in reduction of 30% in the number of abstracts in the corpus, which allows the focus to be on documents relevant to the domain of interest. It also reduces the amount of time spent by biological experts in manually filtering the documents. On average, 10-30% reduction in corpus size could be achieved depending on the domain being investigated.

6. References


