An Application of Text Categorization Methods to Gene Ontology Annotation

Kazuhiro Seki and Javed Mostafa

Laboratory of Applied Informatics Research
Indiana University, Bloomington
1320 East Tenth Street, LI 011
Bloomington, Indiana 47405-3907
{kseki,jm}@indiana.edu

Abstract

With the intense interest and fast growing literature, biomedicine is an attractive domain for exploration of intelligent information processing techniques, such as information retrieval (IR), information extraction, and information visualization. As a result, it has been increasingly drawing much attention of researchers in IR and other related communities [2, 4, 5]. This study shows a successful application of general IR and text categorization methods to this evolving field of research targeting biomedical texts.

In the post-genomic era, one of the major activities in molecular biology is to determine the precise functions of individual genes or gene products, which has been producing a large number of publications with the help of high throughput gene analysis. To structure the information related to gene functions scattered over the literature, a great deal of efforts have been made to annotate articles using the Gene Ontology (GO) terms. GO is a controlled vocabulary developed for describing functions of gene products in order to facilitate uniform queries across different model organism databases, such as FlyBase, Saccharomyces Genome Database (SGD), and the Mouse Genome Informatics (MGI) Database. GO terms are organized in directed acyclic graphs (DAG) under three top level nodes: molecular function (MF), biological process (BP), and cellular component (CC). DAG is similar to hierarchical structure but allows a child node to have multiple parent nodes. Figure 1 illustrates the structure of GO.

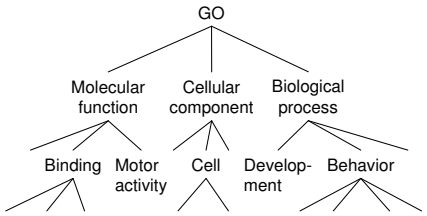

Figure 1: Structure of Gene Ontology.

Because of the large number of publications and specialized content, GO annotation requires extensive human efforts and substantial domain knowledge, which is usually conducted by experts. Thus, there is a potential need to (semi-)automate GO annotation, which could greatly alleviate manual curation. This was one of the primary objectives pursued at the Text Retrieval Conference (TREC) 2004 Genomics Track [3].

The 2004 Genomics Track consisted of two tasks: *ad hoc retrieval* and *categorization* tasks. For the former, given 50 topics obtained through interviews with real research scientists, the participants were required to find relevant documents from 10 years' worth of MEDLINE data. The latter task (which is our focus in this paper) was composed of two sub-tasks; one was called the *triage* task and the other the *annotation* task. Both tasks mimicked some parts of GO annotation process currently carried out by human experts at Mouse Genome Informatics (MGI). Figure 2 depicts the conceptual flow of the two sub-tasks.

In short, the goal of the triage task was to correctly identify whether an input article contains experimental evidence that warrant GO annotation regardless of specific GO codes. The annotation task was the next step to the

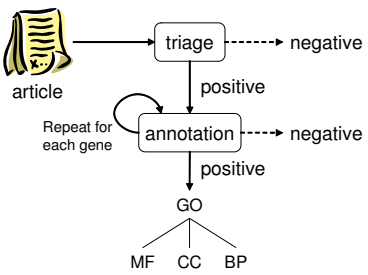

Figure 2: A conceptual flow of the categorization subtasks.

triage decision, and the goal was to correctly assign GO domain codes, i.e., MF, BP, and CC (not the actual GO terms) or not to assign them, i.e., negative, for each of the given genes that appear in the article.[1] Note that there may be more than one gene associated with an article and there may be more than one domain code assigned to a gene.

We participated in TREC 2004 and primarily worked on automatic GO domain code annotation. We approached the task by treating it as a text categorization problem and adopted a variant of *k*NN classifiers. To apply *k*NN, we first represented each input, ⟨article, gene⟩ pair, by a term vector, where terms were collected from text fragments (paragraphs) containing the target gene. To exhaustively locate the gene name occurrences, we took advantage of existing databases to automatically compile a gene name synonym dictionary and preprocessed both gene names and text to tolerate minor differences between them. In addition, we utilized approximate word matching to identify gene occurrences to deal with other irregular forms of the gene names. The collected words were then sent to feature selection using chi-square statistics, which were reused for the supervised term weighting schemes [1].

We evaluated the proposed framework on the TREC Genomics Track data sets and showed that it performed the best in the TREC official evaluation. Further analyses revealed that the flexible gene name matching used in conjunction with the gene name dictionary was notably effective. Another finding is that the result sections of articles contributed the most for GO annotation. It was also demonstrated that our framework was successfully applied to the triage task as well, producing results comparable to that of the best reported system at TREC 2004.

[1] Assuming perfect triage decision, there would not be negative cases at the annotation stage. However, there were negative instances purposefully included in the TREC data.

Acknowledgment

This project is partially supported by the NSF grant ENABLE #0333623.

References

[1] Franca Debole and Fabrizio Sebastiani. Supervised term weighting for automated text categorization. In *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, pages 784–788, 2003.

[2] William Hersh. Report on TREC 2003 genomics track first-year results and future plans. *SIGIR Forum*, 38(1):69–72, 2004.

[3] William Hersh, Ravi Teja Bhuptiraju, Laura Ross, Laura Ross, Aaron M. Cohen, and Dale F. Kraemer. TREC 2004 genomics track overview. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*, 2004.

[4] Lynette Hirschman, Jong C. Park, Jun-ichi Tsujii, Limsoon Wong, and Cathy H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

[5] Hagit Shatkay and Ronen Feldman. Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6):821–856, 2003.