

FigSearch: Using Maximum Entropy Classifier to Categorize Biological Figures

Fang Liu^{1,2} (fang.liu@klinmed.uio.no)
Tor-Kristian Jenssen² (tkj@pubgene.com)
Vegard Nygaard¹ (vegardny@radium.uio.no)
John Sack³ (sack@highwire.stanford.edu)
Eivind Hovig¹ (ehovig@radium.uio.no)

¹Dept. of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway

²PubGene AS, Forskningsveien 2A, P.O.BOX 180 Vinderen, 0319 Oslo, Norway

³Stanford University, HighWire Press, 1454 Page Mill Road, Palo Alto, CA 94304

Abstract

Figures in scientific papers represent an intuitive and concise way of knowledge presentation. With more attention being paid on full-text mining in bioinformatics, we initiated an effort of studying figures in full articles. FigSearch is a prototype figure legend indexing and classification system, using both text-mining and supervised machine learning.

We defined schematic representations of protein interactions and signaling events as an interesting figure type. A maximum entropy classifier was used in categorizing each figure, by assigning an estimated likelihood, as being relevant/non-relevant according to our definition. One advantage of the maximum entropy principle is that it provides a probability of decision, instead of a binary assignment.

In our pilot study, FigSearch showed satisfactory performance in a preliminary validation by domain experts. Such a system can be useful in applications such as for a publisher's website, in bio-picture gallery constructions, or as an aid for other complicated text-mining projects.

1. Introduction

Document classification has been studied using many different methods, including naïve Bayesian[1], maximum entropy[1,2], support vector machine[3,4]. It has been applied to several aspects of biomedical research, such as database construction[4], gene/protein functional annotation[1], and clinical records data mining[2]. In this era with dramatic data flood, machine

learning aided document classification can considerably facilitate researcher to efficiently locate information of interest.

Text mining in bioinformatics is in the process of moving from analyzing abstracts towards full text. As a component in articles, figures encompass a wealth of information being presented in both graphical and textual form.

We present here a study using a maximum entropy classifier to select a specified type of figures, in this case, schematic illustrations of protein interactions and signaling events, from a set of figures collected from a corpus of full-text biological articles. We used the graphical part of figures in a manual training process, and applied text-mining and classification techniques on the textual figure legends. To a significant extent, this accelerated the experts' manual work, while the results showed positively the linkage between picture and text.

2. Methods

First, an XML parser was developed in-house to extract figures from full-text papers. Manual classification was performed on a randomly chosen training set. It is worth noting that this manual training process was done purely by reading the graphical elements, based on the assumption of strong associations between graphical and textual information. Thereafter, we

executed a text-mining procedure on figure legends, including stemming, removal of stopwords, and elimination of very infrequent words.

We chose the words having the most non-uniform distribution frequency between different classes as feature words, and recorded its distributional unevenness as a weight (λ_i) to each feature (f_{w_i}). The higher a weight, the more significant a feature word would represent the interesting class. Then, every figure legend was mapped to a feature vector (f_w) representation:

$$f_w(i) = \begin{cases} \lambda_i, & \text{if } f_{w_i} \in \text{figure legend;} \\ 0, & \text{otherwise;} \end{cases}$$

where $i = 1, 2, \dots, N$, N is the total number of feature words.

We calculated the probability of each legend with a logarithm model, as the following:

$$P = \frac{\sum_i \log(f_w(i))}{\|Z_{f_w}\|},$$

where $Z_{f_w} = \sum_{i=1}^N \log(\lambda_i)$ is the normalization factor

to make sure P ranged from 0 to 1. An analog probability rather than a yes/no decision was given to each figure. The threshold was determined by investigating the classification performance on the training dataset.

A gene/protein name dictionary compiled as in PubGene[5] was adopted to generate figure legend to gene/protein name(s) indexes. This provided the FigSearch system the ability that a user can efficiently search type-specific figures containing gene(s)/protein(s) of their interest.

3. Results

We determined a threshold as 0.53 after examining recall, precision, and effectiveness of classification performance under different thresholds (on the training set). At this threshold, the training set gave 70.5% recall, 86% precision and 77.4% effectiveness (given equal weight on

recall and precision). Eventually, the classifier identified 414 figures as relevant according to our class definition, out of the whole set of 50,258 figures. Then, among the 414 figures, we found 167 as highly relevant, indicating an over 30 times better performance of our system than a random experiment ($p\text{-value} < 8.75e-11$).

We integrated the XML parser, document classifier, and gene/protein indexing module into a prototype application, called FigSearch. The system with an easy-to-use web interface is accessible at

<http://pubgeneserver.uio.no/figsearch/>.

4. Discussion

Maximum entropy takes advantage of feature distribution characteristics in the training set, and is suitable for classification tasks having a discriminant feature set. In this preliminary study, FigSearch demonstrated the feasibility of utilizing the graphical content to aid a text-mining task. The rapidly growing volume of open-access full-text articles may be included in refining our system and extending it to future applications, for instance to other types of figures to be identified.

References

- [1] S.Raychaudhuri, J.T.Chang, P.D.Sutphin, and R.B.Altman, "Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature", *Genome Research*, 2002, v12, pp.203-14
- [2] S.V.Pakhomov, A.Ruggieri, and C.G.Chute, "Maximum entropy modeling for mining patient medication status from free text", *Proc AMIA Symp*, 2002, pp.587-91
- [3] B.J.Stapley, L.A.Kelley, M.J.E.Sternberg, "Predicting the sub-cellular location of proteins from text using support vector machines", *Pac Symp Biocomput*, 2002, pp374-85
- [4] I Donaldson, et al. "PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine", *BMC Bioinformatics*, 2003, v4, pp11-23
- [5] T.K.Jenssen, A.Laegreid, J.Komorowski, E.Hovig, "A literature network of human genes for high-throughput analysis of gene expression, *Nature Genetics*, 2001, v28, pp21-8