Biomedical Document Triage Based on Figure Classification

Nawei Chen*, Dorothea Blostein, Hagit Shatkay
School of Computing, Queen’s University

ABSTRACT

The annotation task in model organism databases is to assign attributes, such as Gene Ontology (GO) codes, to biological entities, such as genes and proteins based on the evidence found in documents or other resources. Document triage precedes an annotation task; it identifies relevant documents that can support the annotation process. Annotation in organism databases involves manual efforts of database curators. Efforts are being made to automate parts of the annotation task. Contests at KDD Cup 2002 and TREC Genomics 2003-2005 formally define subtasks related to annotation. The subtasks are based on document classification and extraction of text from the documents. Here we investigate the analysis of figure images, to complement the analysis of text. Figures are often content rich and concisely summarize the most important results in a paper. We automate document triage by representing image features of each figure as a list of labels. A document is represented by combining the labels of all the figures in the document. We test the method using part of the data from the triage task in TREC 2004.

Keywords: biomedical literature mining, biomedical image classification, biomedical image clustering, document categorization, document triage

1. INTRODUCTION

Model organism databases are important research resources for biologists. Annotation, in this context, is the assignment of attributes to biological entities in the databases. Biomedical publications are usually used as evidence for the annotation of biological entities. As illustrated in Figure 1, annotation is task specific. For example, one annotation task in a biomedical database is to annotate genes with Gene Ontology (GO) codes, while another task is to annotate genes with gene expression information. Annotation is preceded by document triage, which identifies the subset of documents that have evidence and can support the annotation.

Figure 1. Document triage and annotation in a biomedical database. Document triage decides whether a document is relevant for annotation. Triage is task specific. Annotation extracts different types of information in different tasks. The supported information plus the link to the original document are stored into a biomedical database.

Often, document triage and database annotation are performed manually: curators are employed to examine documents and find experimental evidence to annotate biological entities. Efforts are underway to automate parts of the triage and the annotation procedures. We discuss this in the context of the following databases and annotation tasks:

* chenn@cs.queensu.ca
• Mouse Genome Informatics (MGI)$^2$ is the model organism database for the mouse. Among other tasks, MGI curators annotate genes with Gene Ontology (GO) codes, and with gene expression data from the evidence found in biomedical articles or other resources$^3$.
• FlyBase$^4$ is an annotated database on the genetics and molecular biology of the Drosophila (fruit fly).
• Part of the challenges in the KDD Cup 2002$^5$ was a set of triage and annotation tasks, defined to stimulate research in automation of these tasks. Task 1 was to automate the document triage for FlyBase by identifying papers that contain wild-type Drosophila gene expression information.
• The categorization tasks in TREC 2004 genomics track$^6$ had two sub-tasks: triage and annotation, defined to simulate the task of curators for the MGI system. The annotation task is to assign GO codes to genes. The triage task is to classify a document as relevant or irrelevant for GO annotation, given a set of labeled documents as training data.

We focus on the use of figures for document triage. Figures are often content rich and concisely summarize the most important results or methods used in a paper. Figures, specifically fluorescence images from biomedical articles have been used to predict protein sub-cellular localization$^7$. Previous researchers have used figure captions for triage and annotation. For example, Regev et al. used figure captions to identify papers that contain experimental evidence for gene expression products and ranked first in task 1 of KDD Cup 2002$^8$. They state that curators who manually review papers look mainly at the figures in the paper to ascertain the presence of experimental evidence. As a second example, Darwish and Madkour$^{12}$ used text extracted from figure captions for the triage task in TREC 2004. As far as we know, the figure images themselves have not been considered for triage and annotation yet. In this paper, we explore the possibility of using figures for the biomedical document categorization (also called document triage in this paper).

2. USING FIGURES FOR THE TRIAGE TASK

Document triage can be viewed as a binary classification task. The input is a set of full-text documents. A document must be either classified as positive (to be used in annotation) or negative (not to be used in annotation). To automate the task, we train a classifier using a set of labeled training documents (Section 2.1), and then apply the classifier to other documents (Section 2.2). Our basic idea is to create an image-based description for each document, and then apply a naïve Bayes classifier. This approach is adapted from Duygulu et al.’s work on image annotation$^9$. Duygulu et al describe an image using a small vocabulary of blobs, which are labels assigned to the clusters of all the segmented image regions in a collection of images.

2.1 Train the classifier

Our experiments are done using the triage data from TREC Genomics 2004, as described in Section 3. The training data consists of documents that have been labeled as positive or negative by human experts. We further label the subfigures in the training data, as illustrated in Figure 3, and described in Step (3) below.

In our system, training documents are processed using six steps as follows:

(1) **Figure extraction.** The full-text documents we use are in XML format, obtained from TREC Genomics 2004. The data can be downloaded freely from the TREC 2004 Genomics track web site$^6$, after registration. We extract captions and links to the figures from the XML documents, and then download the figure images themselves from the publisher’s web site. A sample document is shown in Figure 2(a). One of the extracted figures is shown in Figure 2(b). We use about 4,000 figures in training and testing as described in Section 3.

(2) **Figure segmentation.** As evident from Figure 2(b), each figure may consist of several subfigures. We segment each figure into its subfigures. A bottom-up segmentation approach, based on Connected Components (CCs) analysis, is used for this purpose. Sample segmentation results are shown in Figure 2(c). Unavoidably, errors do occur during the segmentation process. Usually a figure has mixed types of subfigures and has no standard layout.
Figure 2. (a) A sample input document with PMID (PubMed Identifier) 12235125. The document has nine pages and six figures. (b) Extract all the figures from the document and save as image formats, such as JPEG, or GIF. One of the extracted figures is shown in enlarged version. (c) Figure segmentation based on Connected Components (CCs) analysis. Subfigures are extracted from each figure. The CCs whose bounding box areas are too small are discarded since they are most likely characters used to label figures. (d) Subfigure classification using a hierarchical classification scheme defined in Figure 3. Unavoidably, there are errors in figure segmentation and subfigure classification partly due to lack of standard layout and mixed types of subfigures in a figure.

(3) Subfigure classification. We classify the subfigures and use the classification results as a basis for creating labels that capture image features in each figure. We define a hierarchical classification scheme explained below and shown in Figure 3; we plan to refine this scheme in the future. Currently, at the first level, images are classified into Graphical and Experimental classes. For the Experimental class, we define three subclasses: Fluorescence Microscopy, Gel, and Other Microscopy. The reason is that the three subclasses are visually distinct and obtained in different experimental settings. In order to train a classifier to use this classification scheme, we manually labeled a few hundred subfigures in each class. We use two SVM (Support Vector Machine) classifiers: one at the root level to classify the images into Graphical and Experimental images, and the other at the second level of the classification hierarchy to further classify Experimental images into one of the three subclasses. Thus every subfigure is assigned to one of four class labels: Graphical, Fluorescence Microscopy, Gel, and Other Microscopy. Image features used by the SVM classifiers include statistics based on gray-level histograms, concurrence texture features, edge direction histogram, and features based on run-length analysis of binary images. Example of classification results are shown in Figure 2d.

Figure 3. A hierarchical image classification scheme. A sample image for each class is given. At the first level, images are classified into Graphical and Experimental images. Other types of images found in publications include photographs of people, pictures of mice, etc. In our current work, we manually pre-filter the extracted subfigures to remove such Other images. At the second level, Experimental images are classified into Fluorescence Microscopy, Gel Electrophoresis, and Other Microscopy images. Graphical images are classified into Line Charts, Bar Charts, and Diagrams. In our experiments, Graphical images are not classified further; we focus on classification of Experimental images into Gel, Fluorescence, and Other Microscopy images.
(4) **Subfigure clustering.** Next, we perform clustering to define fine-grained image classes automatically. In Step (3), all the training subfigures were classified into one of four classes. There are about 10,000 subfigures in the training data. We felt that for better accuracy we should partition the classes into subclasses, as four manually defined classes may not provide sufficient discrimination among thousands of subfigures. We thus use clustering for obtaining subclasses of *Experimental* subfigures. *Graphical* images are not clustered further. Since the number of subfigures belonging to the *Fluorescence Microscopy* class is significantly larger than the other two classes, they are clustered into 20 clusters. Subfigures belonging to the other two classes are clustered into 10 clusters each. Currently, we choose the number of clusters heuristically. Different number of clusters may be used. In the future, we may test how the choice of number of clusters affects the classification performance.

Clustering should group together images with similar characteristics. The choice of image features is critical for the effectiveness of clustering. In the current clustering, we use all the gray-level image features used in the subfigure classification described in step (3). More discriminant features for each class may be used as well. Further discussion of document image classification techniques is given in our survey paper\(^\text{10}\). To summarize, subfigures within each subclass of *Experimental* are clustered in this step; the clustering results are used to assign a cluster label to each *Experimental* subfigure.

(5) **Create an image-based feature vector of each document.** Using the classification and clustering results from steps (3) and (4), we assign a label to each subfigure. For example, the top left subfigure in Figure 2(b) is assigned the label F17, where F stands for *Fluorescence* and 17 stands for cluster 17 in the clustering of *Fluorescence Microscopy* subfigures. The labels of all the subfigures in one document are combined to create an overall description of a document based on image features. Then a feature vector is extracted from the description. For example, the description of the document shown in Figure 2(a) is:

```
 graphics graphics graphics F19 graphics graphics E2
 F17 F9 F19 F16 graphics
 graphics graphics graphics G6 G7
 graphics G1 graphics G3 graphics
 F17 G0 graphics graphics graphics graphics E7 F6 G6 E5 graphics
 E1 graphics E5 G1 G4 graphics
```

In this description, G represents *Gel*, F represents *Fluorescence* and E represents *Other Microscopy*. The image description is created by combining the labels of 39 subfigures, drawn from six figures in this document.

(6) **Train a Naïve Bayes Classifier.** Given the image-based description created in Step (5), we build feature vectors and train a Naïve Bayes Classifier using all the training documents. We use the MALLET toolkit for feature vector creation and document classification\(^\text{11}\).

2.2 Executing the classifier

The results of the training phase (Section 2.1) are the clusters for each of the three *Experimental* subclasses (Step 4) and a classifier (Step 6). Given an input document, we classify it using the following procedure: First, the document goes through steps (1) - (3), the same figure-based preprocessing as in the training phase. Then each subfigure is assigned the cluster label of its nearest neighbor in the training set using the results of training Step (4). An image-based description is created containing a list of labels of all the subfigures in the document, similar to training Step (5). Then a feature vector is computed and fed into the Naïve Bayes classifier from training Step (6). This classifies the input document as positive or negative.

3. EXPERIMENTAL RESULTS

We use the data set from the categorization task of TREC Genomics Track 2004. The documents for the task consist of articles from three journals over two years. Year 2002 articles are designated as training data and year 2003 articles as test data. Only full-text documents from the *Journal of Cell Biology (JCB)* provided in TREC Genomics Track 2004 are used in the current experiment. The distribution of training and test set is shown in Table 1.
The results are shown in Table 2 using the abbreviations: TP (True Positive), FP (False Positive), FN (False Negative) and POS (number of positive samples in the test data). We use the same evaluation metrics as were used to evaluate the triage subtask in TREC2004. As reported by Hersh et al. \cite{14}, the primary evaluation metric in the TREC2004 triage subtask is a normalized Utility value, calculated as $U_{\text{norm}} = \frac{(20\times TP) - FP}{20\times POS}$. The constant 20 serves to weight the evaluation toward positive answers: the cost of missing a document that should be triaged is much greater than the cost of including a document that does not need to be triaged. Hersch et al. discuss that the ideal approach for determining this constant would involve interviewing MGI curators and formally determining utility. Time did not allow this, so informal reasoning was used to derive the value of 20.

Our results do not compare directly with the TREC2004 Triage task results, because we use only a subset of the training and test documents. The whole training set of triage task data has 5,837 documents and the test set has 6,043 documents. All 59 of the TREC 2004 Triage runs were based on full-text documents, including figure captions, but not including any analysis of figure images. Our results are opposite, making use only of figure images, and making no use of text. As shown in Table 2, our results are roughly comparable to the average results in TREC 2004 runs. This is encouraging, indicating that a combination of figure and text analysis may yield good results in the future.

<table>
<thead>
<tr>
<th></th>
<th>Utility</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our system</td>
<td>0.3074</td>
<td>0.2791</td>
<td>0.3529</td>
<td>0.3117</td>
</tr>
<tr>
<td>Mean of 59 runs in TREC 2004 triage subtask (from Table 6\cite{14})</td>
<td>0.3303</td>
<td>0.1381</td>
<td>0.5194</td>
<td>0.1946</td>
</tr>
</tbody>
</table>

Table 2. Classification results, using the evaluation metrics from the TREC 2004 triage subtask\cite{14}. Results from the TREC 2004 Triage runs are shown for an informal comparison. Due to the efforts involved in obtaining figure images, we only used a fraction of the test and training documents used in the TREC Triage task. Our testing used 34 positive and 325 negative documents, whereas the TREC 2004 Triage testing used 420 positive and 5623 negative documents.

4. CONCLUSION

The current research is a preliminary exploration of the possibility of using figures for document triage. We believe that a refined classification scheme for subfigures is important for improving the result. We plan to add the rejection ability to our system so that the Other type of images can be automatically rejected without the need of manually filtering as mentioned in the caption of Figure 3. Feature selection is crucial to improve subfigure classification and clustering performance. In our future research, we will investigate how a human curator uses figures in judging whether a document supports annotation or not, and how figures are used during the annotation process. Observing how humans handle the task will provide further ideas on how to automate (parts of) the task.

We are going to do research on using figures to automate more tasks currently performed by human experts. For example, we are interested in using figures for the annotation task in the TREC 2004 genomics track. We are also interested in combining the analysis of text and figures for document triage and annotation tasks. Figures and figure captions are information rich portions of documents. They usually present important experimental findings. Figure images often play a key role in understanding the papers’ results. Therefore, combining image analysis with text analysis is expected to help resolve ambiguity and improve the effectiveness of the tasks that are traditionally done using only text.
ACKNOWLEDGMENTS

We gratefully acknowledge the financial support provided by NSERC, Canada’s Natural Sciences and Engineering Research Council.

REFERENCES

2. Mouse Genome Database (MGD), Mouse Genome Informatics Web Site, the Jackson Laboratory, Bar Harbor, Maine. World Wide Web (URL: http://www.informatics.jax.org/, December 2005).