BIOINFORMATICS

Vol. 22 no. 14 2006, pages e446–e453
doi:10.1093/bioinformatics/btl235

Integrating image data into biomedical text categorization

Hagit Shatkay*, Nawei Chen and Dorothea Blostein
School of Computing, Queen's University, Kingston, Ontario, Canada

ABSTRACT

Categorization of biomedical articles is a central task for supporting various curation efforts. It can also form the basis for effective biomedical text mining. Automatic text classification in the biomedical domain is thus an active research area. Contests organized by the KDD Cup (2002) and the TREC Genomics track (since 2003) defined several annotation tasks that involved document classification, and provided training and test data sets. So far, these efforts focused on analyzing only the text content of documents. However, as was noted in the KDD'02 text mining contest—where *figure-captions* proved to be an invaluable feature for identifying documents of interest—images often provide curators with critical information. We examine the possibility of using information derived *directly from image data,* and of integrating it with text-based classification, for biomedical document categorization. We present a method for obtaining features from images and for using them—both alone and in combination with text—to perform the triage task introduced in the TREC Genomics track 2004. The task was to determine which documents are relevant to a given annotation task performed by the Mouse Genome Database curators. We show preliminary results, demonstrating that the method has a strong potential to enhance and complement traditional text-based categorization methods.

Contact: shatkay@cs.queensu.ca

1 INTRODUCTION

Categorization of biomedical text is pivotal both for supporting curation tasks in biological databases and for providing researchers with literature appropriate for their specific information needs. For example, curators for the Mouse Genome Database (MGD) need publications with specific contents to validate the expression of genes under certain conditions. Other examples for curation-related task include the identification of papers discussing subcellular localization in support of the annotation of proteins with Gene Ontology (GO) codes for subcellular component, or of papers discussing function—to be used as evidence for functional annotation. On the other side of the quest for information, scientists in individual labs may want to easily identify papers that are likely to be related to their own research, or may look for papers discussing a new area of interest into which they are ready to venture. Underlying all these examples is the need to identify a subset of documents, with some common topical characteristic, within a large set of documents. The latter set may include hundreds of documents returned by a broad PubMed search, or possibly thousands of documents in a certain journal, or even the millions of documents comprising the whole of MEDLINE.

In the past few years several initiatives were established to encourage and evaluate work on biomedical text categorization. The KDD'02 cup (Yeh *et al*., 2003) had a task in which documents were to be categorized as containing (or not containing) evidence for gene expression within the Drosophila wild type, in support of FlyBase curation. For the past two years the TREC Genomics track (Hersh *et al*., 2005, 2006) featured a text categorization task, in which documents were to be classified according to their evidence contents in support of assigning GO annotation to mouse genes. Part of *Task 2* of the BioCreative challenge (Hirschman *et al*., 2005) involved identifying papers that contain evidence for assigning GO codes to human proteins, in support of Swiss-Prot curation.

In all these tasks the documents were categorized based only on the text occurring in them. While participating in the KDD cup, Regev *et al*. (2002) noted that the use of figure captions proved particularly helpful for their high performance in identifying documents discussing gene expression. Following this work, figure captions were also used by participants in the TREC Genomics track (Darwish and Madkour, 2005) as part of the text-features used for categorization. The success of using figure captions is related to the fact that figures contain important cues that are typically used by database curators and annotators to quickly scan documents and distinguish relevant from irrelevant ones. FlyBase curators have indeed indicated that the experimental results shown in papers and used in support of curation, are often presented in figures and their captions (Yeh *et al*., 2003). Figures are often content rich and concisely summarize the most important results or methods used and described in an article.

Our present work is motivated by this idea, taking it one step further; namely, we investigate the use of features derived directly from the image data of the figures (as opposed to just from the text of the figure captions) for biomedical document categorization. It is intuitively clear that image and text data, especially in scientific documents, tend to complement each other. Moreover, psychological studies on the contribution of multimodal data (image, animation, text) to effective understanding in human readers, confirm the efficacy of the combination of image and text for improving the processing and understanding of information by humans, compared with the unimodal form (i.e. either text or image data alone) (Mayer and Moreno, 2002). We report here a first experiment, introducing image features into the text categorization process, and show preliminary results in applying it to a subset of the TREC Genomics data.

*To whom correspondence should be addressed.

Notably, image-based categorization of documents is an established research field (Chen and Blostein, 2006). It is applied in diverse areas ranging from digital library construction and document image retrieval to office automation. Document image classifiers differ vastly in the problems they solve, in their use of training data to construct class models, and in the choice of document features and classification algorithms. There is no single general, adaptable, high-performance image-based classifier, due to the great variety of documents, the diverse criteria used to define document classes, and the ambiguity in the class definition itself. Thus, the specific task at hand needs to be considered when choosing and applying image-based categorization methods in the biomedical domain.

To the best of our knowledge, the use of figure images themselves has not yet been considered for general biomedical document triage and for automated support of biomedical annotation and curation. Perhaps closest to ours is work by Murphy *et al*. (Huang and Murphy, 2004; Murphy *et al*., 2004), which uses image categorization for identifying subcellular localization articles. They provide an excellent in-depth investigation of a specific task: identifying and interpreting a specific type of image that is characteristic of localization experiments. While their extensive work utilizes information extraction from text to help improve image categorization and interpretation, it is not directed at the integration of text and image features for the purpose of document categorization. Moreover, the research focuses on protein subcellular localization and is not generalized to other biomedical categorization tasks.

In this paper, we explore the possibility of using figures for the document triage task in support of biomedical database curation. We describe a first attempt at using image features for biomedical text categorization, as well as at the integration of such features with the more traditional text-data. The next section outlines the methods we apply, while Section 3 describes the data set and demonstrates preliminary results of applying our integrated categorization method. Section 4 concludes and outlines future work.

2 USING FIGURES FOR DOCUMENT TRIAGE

Document triage can be viewed as a binary classification task. The input is a set of full-text documents, and each document is classified as either *positive* (relevant for annotation) or *negative* (irrelevant for annotation). To automate the task, a classifier is trained using a set of labeled training documents, and is then applied to the test documents to predict their class. Our basic idea is to create an image-based vector description for each document in both the training and the test sets. Once a vector description is created, traditional classification methods can be applied to the data. In this paper we focus on the simple naïve Bayes classifier, although more advanced methods are likely to yield improvement. The image-description approach is adapted from work by Duygulu *et al*. (2002) on content-based image retrieval. Duygulu *et al*. segment images into regions, cluster similar regions across the different images into what they call ''blobs'', and thus create and use a small vocabulary of characteristic segments for representing images. Through most of this section, (2.1, 2.2), we describe our image feature extraction and the document representation in terms of image features. The last part of the section (2.3) provides a brief description of a first integrated framework for combining image features and text data for biomedical document classification. Our experiment and results using a subset of the TREC Genomics 2004 data are described in Section 3.

2.1 Document descriptors via image features

As with any supervised text categorization task, the training data consists of documents that have been manually labeled by human curators as *positive* or *negative*. Typically in text categorization, the documents are then represented as weighted vectors of terms or of words. (For reviews see: de Bruijn and Martin, 2002, Shatkay and Feldman, 2003.) In the heart of our approach is the representation of documents *as vectors of image features* rather than of text features[1], which we describe in detail below.

Before delving into the details, in a nutshell the method comprises five main steps: First, figures are *extracted* from the full-text documents. As single figures often display multiple pictures, they are broken in a *segmentation* step into subfigures. These subfigures are then *classified* into several high-level types of images that we have defined. These three steps are shown in Figure 1. Within each class, *clustering* is then applied to refine the grouping of images by specific contents. Each subfigure is assigned an identifier coding its class and its cluster. In the final step, each document is then *represented* as a vector over the space of subfigure-identifiers as features (similar to the vector space over terms or words typically used in text). We discuss these steps in detail below.

a) Figure extraction. This step starts with full-text XML documents. Captions and links to the figures are extracted from the XML format, figure images are downloaded from the publisher's web site. A sample document is shown in Figure 1(i). One of the extracted figures is shown in Figure 1(ii). For the training and tests described here we used a total of about *4,400* figure images, of which *1,900* came from the training and *2,500* from the test documents.

b) Figure segmentation. As evident from Figure 1(ii), each image may consist of several subfigures. Each image is thus segmented into its subfigures using an approach based on connected components analysis (Gonzalez and Woods, 2002). Such analysis is performed on thresholded black-and-white images, where connected components are regions of neighboring foreground pixels. The connectedness is defined based on eight-neighbors of each pixel. Figure 1(iii) demonstrates the results of such segmentation. We note that this is not a fool-proof procedure, and errors are expected to occur. In the data described here, we identified a total of about *26,500* subfigures (*∼11,000* in the training and *∼15,500* in the test set).

c) Subfigure classification. The subfigures identified in step *b* may illustrate various types of data and be organized in a variety of layouts. As pointed out by Murphy *et al*. (2002), there are no uniform standards for figure organization in the scientific literature. As shown in Figure 2, we have identified several prominent types of figures in the scientific literature and use these types for categorizing subfigures. Obviously this ''ontology'' of image types is neither complete nor perfect, but has proven to be a useful first step for the limited scope in which it is used here.

Subfigure classification forms the basis for creating labels that are later used to represent image features in each figure. Currently, at

[1]We note that for combining text and figures we do use *both* text and image features.

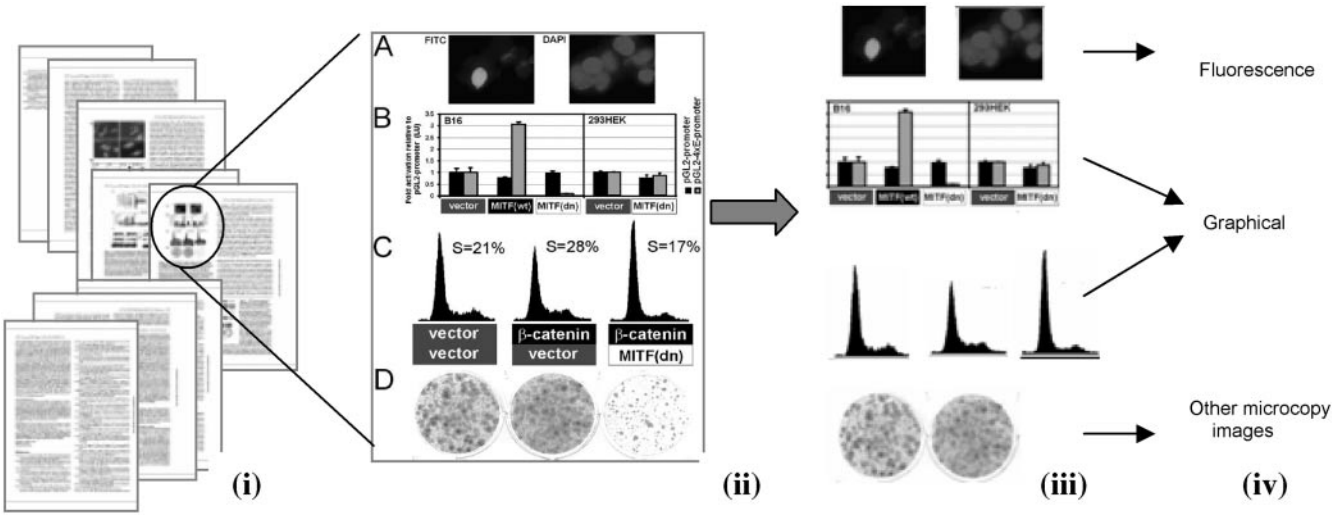

Fig. 1. (i) A sample input document with PubMed Identifier 12235125 (Widlund *et al*., 2002). (Figures reproduced with permission of the Rockefeller Univsrsity press.) The document has nine pages and *six* figures. (ii) Extract all the figures from the document and save as image formats, such as JPEG or GIF. One of the extracted figures is shown enlarged. (Corresponds to step *a* below.) (iii) Figure segmentation based on Connected Components analysis. Subfigures are extracted from each figure. Connected components whose bounding box areas are too small are discarded since they are most likely characters used to label figures. The example document has a total of 39 subfigures. (Step *b* below.) (iv) Subfigure classification using a hierarchical scheme as defined in Figure 2. (Step *c* below.)

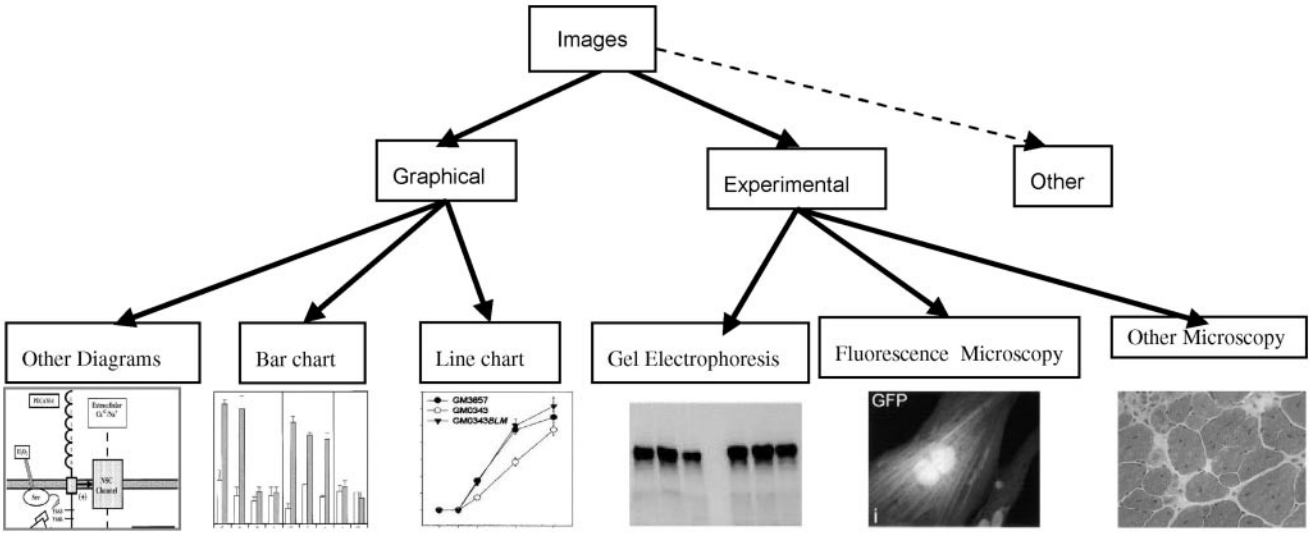

Fig. 2. The hierarchical image classification scheme for subfigures. A sample image is shown for each class. At the top level, images are classified into *Graphical* and *Experimental* images. Other types of images found in publications include photographs such as pictures of mice, author images, etc. 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 *Other Diagrams*. In our experiments, *Graphical* images are not further classified. We focus on classification of *Experimental* images into *Gel Electrophoresis*, *Fluorescence Microscopy*, and *Other Microscopy* images.

the first level, images are classified into *Graphical*, *Experimental* and *Other* classes. For the *Experimental* class, we currently define only three subclasses: *Fluorescence Microscopy*, *Gel Electrophoresis*, and *Other Microscopy*. These three subclasses are visually distinct and correspond to clearly different experimental settings. Obviously, more classes should be defined to accommodate other types of experimental imaging. *Graphical* images can also be partitioned into subtypes. For instance: *Line Chart*, *Bar Chart* and *Other Diagrams*. However, in the experiments described here graphical images are *not* further partitioned.

In order to train a classifier to categorize subfigures under this classification scheme, we manually labeled a few hundred subfigures in each class (500 *Graphical* subfigures, 500 *Fluorescence Microscopy*, 300 *Gel Electrophoresis*, and 300 *Other Microscopy*). We use two Support Vector Machine (SVM) classifiers: one at the root level to classify the images into *Graphical* vs.

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 one of four class labels: *Graphical*, *Fluorescence Microscopy*, *Gel Electrophoresis*, or *Other Microscopy*. Examples of subfigure classification results are shown in Figure 1(iv). Using a stratified 10-fold cross validation, the first level classifier for separating *Graphical* from *Experimental* subfigures demonstrates about *95%* accuracy, while the second classifier that separates the three types of experimental subfigures demonstrates a level of *93%* accuracy. Note that this is *not* the ultimate categorization task discussed in this paper; rather, it is a preprocessing step used towards representing images that appear in scientific papers.

To facilitate classification by SVM, subfigures must be represented as feature vectors. The following 46 features are used for representing subfigures in this stage:

- *Statistics based on gray-level histograms*. The histograms represent the distribution of pixels in the subfigures according to their gray-level. Four statistics are derived from the histogram: the first three moments (mean, variance, and skewness) as well as the entropy of the gray-scale distribution (Gonzalez and Woods, 2002).
- *Haralick's texture-features* (Haralick *et al*., 1973), based on the co-occurrence of pixels within the subfigure. The co-occurrence matrix provides information about co-occurring pixels of specific values, orientation and distance. Six features are derived from the matrix including, among others, *contrast* (variation in gray level), *correlation* (likelihood of co-occurrence for specified pixel pairs), and *homogeneity* (formally described as *Inverse Difference Moment*).
- *Edge direction histogram* (Jain and Vailaya, 1998), originally used for shape-based retrieval. Edges are detected in the subfigure, using Canny's edge detector (Canny, 1986). A histogram which bins together edges sharing a similar direction is then formed. Our implementation uses a bin granularity of 10°, resulting in a histogram of 36 bins. The bin sizes (i.e. the number of edges in each of the bins) are used as features.

The image feature vectors are normalized before classifying them. Classification is done using Weka's (Witten and Frank, 2005) implementation of Support Vector Machines, with the radial basis function kernel.

d) Subfigure clustering into finer groups. In the previous step subfigures were classified into one of four coarsely-defined classes. In the relatively small training set (256 documents) described here alone there were about 11,000 subfigures. As it was expected that the four broad manually defined classes, while intuitively clear, are unlikely to provide sufficient discrimination among thousands of subfigures, we use unsupervised clustering to refine the grouping of similar and related images into tight subsets. Since the number of subfigures assigned to the *Fluorescence Microscopy* class is about 4 times larger than the number of subfigures assigned to each of the other two classes, the Fluorescence Microscopy class is sub-clustered into 20 clusters, while the other classes are sub-clustered into 10 clusters each. Clearly, a different number of clusters may be used, and may yield different results. We have chosen the current numbers based on the total size of the image set used here, the total number of sub-figures stemming from it, and based on several experimental runs. We expect to test more methodically in future studies how the number of clusters affects the classification performance. While this is an interesting point whenever clustering is concerned, it is not a central issue for the work presented here.

```
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
```

Fig. 3. The document shown in Figure 1(i), represented using only subfigure identifier terms.

The clustering step groups together images with similar characteristics. In this study, we use the simple k-means algorithm, as implemented in Weka (Witten and Frank, 2005). The features considered are the same ones used for the subfigure classification described in step *c* above. As this is a first study on the use of images for biomedical text categorization, we have not yet explored the range of possibilities for representation, classification and clustering, and expect to do so in the future. A discussion of the variety of methods for document image classification techniques is given in a previous survey (Chen and Blostein, 2006).

To summarize this stage, subfigures within each of the four classes that were formed in step *c* are clustered into finer groups. The clustering results are used to assign a cluster label to each subfigure, which together with the class label serve to characterize each subfigure in every document.

e) Document representation as an image-based feature vector. In steps *c* and *d* each subfigure has been assigned both a class name and a cluster number. Combined, this information forms a label characterizing each subfigure in terms of its class and cluster. For example, the top left subfigure in Figure 1(iii) is assigned the label *F17*, where *F* stands for *Fluorescence Microscopy* and *17* stands for *cluster 17* among the 20 clusters of *Fluorescence Microscopy* subfigures. The labels of all the subfigures in each document are taken as new kinds of terms used to represent each document based only on its image features. A feature vector is then constructed from the description, similar to the way weighted term vectors are built from text. For example, the description of the document shown in Figure 1(i) is shown in Figure 3 (before vectorization and term weighting is performed). In this description, G represents *Gel Electrophoresis*, F represents *Fluorescence Microscopy* and E represents *Other Microscopy*, while ''graphics'' denotes subfigures that are non-experimental *Graphical* images. This image description was created by concatenating the labels of 39 subfigures, comprising the six figures in the whole article.

The corresponding vector representation under a simple term-frequency weighting scheme is shown in Figure 4. This is a 41-dimensional vector, as there are 10 *Gel Electrophoresis* clusters, 20 *Fluorescence Microscopy* clusters, 10 *Other Microscopy* clusters, and a single *Graphical* class that is not subclustered. In this case each number in the vector represents the number of times the respective feature occurred in the representation shown in Figure 3.

2.2 Image-based classification with naïve Bayes

Given the image-based description created in step *e* above, each document is further converted into an *n*-dimensional feature vector,

E0	E1	E2	…	E5	E6	E7	E8	E9	F0	…	F6	…	F9	…	F16	F17	F18	F19	G0	G1	G2	G3	G4	G5	G6	G7	G8	G9	graphics
<0	1	1	…	2	0	1	0	0	0	…	1	…	1	…	1	2	0	2	1	2	0	1	1	0	2	1	0	0	19 >

Fig. 4. The vector representation for the document shown in Figure 1(i) and Figure 3, using term-frequency weighting. The feature labels are listed above their weights. In the weight vector, '. . .' indicates a sequence of consecutive 0's.

where n is the total number of distinct image-based terms (where a *term* is a descriptor such as ''graphics'' or ''*E7*'' above). For each article, every such term is weighted according to its frequency in the article, using MALLET's (McCallum, 2002) default weighting scheme.

Once the feature vectors are formed, we build a naïve Bayes Classifier using all the training documents, to distinguish *positive* articles (relevant for curation) from *negative* ones (irrelevant for curation). Naïve Bayes is a simple and popular classification method; given its simplicity and ease of implementation, it performs well in practice (Mitchell, 1997). The naïve Bayes classifier is built by obtaining statistics from the set of labeled training data. A document D, represented by its feature vector $(d_1, \ldots, d_n)$, where in our case d_i is the weight of the i^{th} subfigure-identifier term, is assigned to the class C that maximizes the likelihood: $\Pr(D \mid C) = \prod_{i=1}^{n} \Pr(d_i \mid C)$.

Expressing the conditional probability $\Pr(D \mid C)$ as a product of simpler probabilities is based on the (naïve) assumption of conditional independence among the features, given the class. We use the MALLET toolkit (McCallum, 2002) for feature vector creation and for the naïve Bayes classification of documents. We note that although MALLET was originally built for text processing and categorization, we use here image-derived features (as shown in Figure 3) rather than text features as input to MALLET.

The representation and training steps given above, when applied to the training data, result in clusters and classifiers for subfigures (steps *c, d* above), which allow each document to be represented based on its image contents (steps *a-e* above). More importantly they yield a naïve Bayes classifier for categorizing documents, using their image-based representation. Given a *new input document*, we classify it by executing the following procedure: First, the document goes through steps *a-c*, namely, its figures are extracted, segmented and its subfigures classified, in a way similar to the preprocessing applied to the training data. Then each subfigure is assigned the cluster label of its nearest neighbor in the training set, using the results of training step *d*. An image-based description is created containing a list of labels of all the subfigures in the document, similar to training step *e*. Then a feature vector is computed and fed into the naïve Bayes classifier described above. This classifies the input document as *positive* or *negative* based on its relevance to the curation task at hand.

2.3 Integration with a simple text classifier

As a first attempt at integration of text data with image features, we use the simplest and most widely used and readily available text for biomedical documents, namely only the title and the abstract of the articles as they appear in PubMed. The titles and abstracts of all the articles contained in both the training and the test set were tokenized to obtain a dictionary of terms consisting of single words (unigrams) and pairs of consecutive words (bigrams), where words were stemmed using the Porter stemmer (Porter, 1997) and standard stop-words removed. Rare terms (appearing only in a single document) as well as very frequent ones (occurring in more than 10% of the documents) were also removed. The remaining terms, along with their frequencies within each of the documents were used to create, for each article, a representation similar to the one shown in Figures 3 and 4, only in this case the features are the actual text-terms. The abstracts of articles in the training set were then used, as described in Section 2.2 to train a naïve Bayes classifier using the MALLET toolkit (McCallum, 2002). We note that both the preprocessing and the classification schemes here are basic ones, and will be extended in the very near future.

The integration scheme for combining the text and the image classifiers consists of a simple OR combination, where a document is considered as relevant for the triage task if either the text-based classifier or the image-based classifier identified it as relevant. This strategy is based on the observation that the triage task stressed the importance of retrieving as many relevant documents as possible, even at the cost of drawing in false-positives (more detail is given in the next section).

3 EXPERIMENTS AND RESULTS

3.1 Experimental setting

We test our method on a subset of the data that was used for the categorization task in the TREC Genomics Track 2004 (Hersh *et al*., 2005), and specifically focus on the *triage* task. The *triage* task aimed to classify documents as *relevant* or *irrelevant* for supporting GO annotation by curators for the Mouse Genome Informatics (MGI) resource at the Jackson labs. The original dataset consisted of full-text articles from three journals: *The Journal of Biological Chemistry (JBC), The Journal of Cell Biology (JCB)*, and *The Proceedings of the National Academy of Science* (*PNAS*), over the period of two years, 2002 and 2003. The 2002 articles (a total of 5,837) were designated as the training set for the task, while those from 2003 (6,043 such articles) as the test set. The true triage decisions were provided by MGI.

In the experiments described here, we use only documents from the *Journal of Cell Biology (JCB)* as provided in TREC Genomics 2004. It is important to note that image data was *not* included in the TREC data set. Given the non-trivial time and effort needed to obtain the image data, download and process it, and given that this is the first study to use biomedical image data for biomedical literature categorization, we wanted to first validate the feasibility of the task and establish a well-defined pipeline, before embarking on the more ambitious task of utilizing the full amount of available data. The distribution of training and test data used here is shown in Table 1.

We train a classifier based on the images from the 256 training documents, and test it on the 359 test documents. A simple text-based classifier is trained on just the abstracts and titles of the same set used for training the image-based classifier, and tested on the

Table 1. The distribution of positive and negative documents in the training and test data sets

	Positive documents	Negative documents	Total figures extracted	Total subfigures extracted	Total documents
Training JCB'02	26	230	1,881	10,920	256
Test JCB'03	34	325	2,549	15,549	359

abstracts and the titles of the same test set as used in the image case. Finally, an integrated classifier assigns a document as *relevant for curation* if *either* of the two first classifiers tagged it as *relevant*. To evaluate our results, we use the same metrics used to assess the triage subtask in the TREC 2004 Genomics track. The primary evaluation metric for the triage subtask, as defined by Hersh *et al*. (2005), was the normalized *Utility* value, defined as:

$$U_{norm} = \frac{(20 \cdot TP) - FP}{20 \cdot Pos}$$

In this formulation, *TP* is the number of true positives (documents that were relevant for curation according to MGI, and identified by the classifier as relevant), *FP* is the number of false positives (documents identified by the classifier as relevant, but not considered as such by MGI), and *Pos* is the total number of articles that are relevant according to MGI. The constant *20* was introduced by Hersh *et al*., and serves to bias the evaluation to favor high recall (that is, including as many positive examples as possible). It reflects the notion that missing a relevant document that should be curated is considered much more costly than including an irrelevant document. Hersh *et al*. (2005) indicated that the ideal approach for determining this constant would involve interviewing MGI curators and formally determining utility, but they used a simplified approximation for the time being. Other measures include the standard *precision*, *recall,* and *F-score* (combining recall and precision). The formulae for these last three measures are as follows, where we again use the abbreviations *TP* (True Positive), *FP* (False Positive), *FN* (False Negative):

$$\text{Precision: } \frac{TP}{TP + FP} \qquad \text{Recall: } \frac{TP}{TP + FN}$$

$$\text{F-score: } \frac{2 \cdot recall \cdot precision}{recall + precision}$$

3.2 Results

Table 2 summarizes our results from training and testing over the JCB dataset (as shown in Table 1).

It is important to note that while our results are in the same utility range as that obtained by TREC — and the combined utility of the integrated system may look even higher than that achieved by the average TREC run — our numbers (the top three rows) *do not* compare directly with the TREC 2004 Triage results (the bottom row), because we use only a *subset* of the TREC training and test documents. The bottom row is provided not for comparing our classifiers with those of TREC, but rather to provide a ''ballpark'' range for what one may expect to see in such results, and to demonstrate that our results fall in this range. Meaningful comparative analysis can only be made among the numbers presented in the top three rows.

Table 2. Classification results, using the evaluation metrics described by Hersh *et al*. (2005). Average results from the TREC 2004 Triage runs, taken from Table 6 of Hersh *et al.'s* report (2005), 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, as shown in Table 1. Our testing used 34 positive and 325 negative documents, whereas the TREC 2004 Triage testing used 420 positive and 5,623 negative documents

	Utility	Precision	Recall	F-score
Image-features system	0.307	0.279	0.353	0.312
Simple text classifier	0.315	0.647	0.323	0.431
Integrated	**0.446**	0.315	0.5	0.386
Avg. of 59 runs in TREC'04 triage task	0.330	0.138	0.519	0.195

All 59 of the TREC 2004 Triage runs were based on full-text documents[2], including figure captions, but not including any analysis of figure images. In contrast, our results for the image-based classifier makes no use of text and uses *only image data*, while the text-based classifier uses only the title and the abstracts of the documents with no other information. The combined classifier takes only the output of these two classifiers to make a categorization decision. As shown in Table 2, our results are well within the numerical range of the average results in TREC 2004 runs. This is encouraging, indicating that even with very simple features the image-based classifier can achieve a reasonable level of performance.

Most importantly, we note that the integration of the image classifier and our simple text classifier significantly improves upon the *utility* obtained by each of the individual classifiers alone. As explained in the previous section, this integration is performed by assigning the tag *relevant*, to a document if any of the two first classifiers categorized it as *relevant*. The fact that this strategy improves recall, (and in-turn utility), indicates that the two original classifiers are not strongly dependent, and use different criteria to reach their conclusions. This is an important observation, given that combining classifiers relies on the idea that an ensemble of classifiers improves performance with respect to its individual components if these components are mostly independent of each other (Sebastiani, 2002, Tumer and Ghosh, 1996). These preliminary results and the nature of both images and text in scientific documents indicate that the combination of figure and text analysis has the potential to yield good results. We expect that image data, which

[2]Notably, not all 59 runs took advantage of the full text; some participants utilized only parts of it, such as abstract, title or MeSH terms.

is a condensed form of information specific to certain types of scientific discussions, will complement the information conveyed in the natural-language text.

4 DISCUSSION AND FUTURE WORK

The research presented here is a first exploration of the possibility of using image data in support of document categorization in the biomedical domain. We note that the idea of using figures for the end goal of text classification is novel and has not been applied yet even in the general context of text categorization (i.e. outside the biomedical domain). In our current work we used a rather small data set, simple methods for segmentation, classification and clustering of subfigures, as well as a very basic text classification and integration strategy. The results of even this simple approach are encouraging and suggest that image data has much to offer in support of biomedical text categorization. A refinement of all these steps is expected to improve the end result. An important immediate step is the application of both the current and the refined methods to the full data set, and specifically to the TREC'05 categorization tasks[3]. Experiments with the GO and Allele categorization tasks of TREC'05 (Hersh *et al*., 2006) over the JCB subset, using appropriately adapted utility scaling measures, yield results similar to the ones shown in Table 2. We are already running the system on the complete data set, and are currently experimenting with categorization, clustering and feature selection strategies that are appropriate for this much larger and heterogeneous data set.

Experiments with other classifiers, aside from the naïve Bayes, as well as the application of more advanced text-categorization and the use of text from captions and other parts of the document, are natural and essential directions we are currently pursuing. Another important next step is the study of the complementary role of text and image data in biomedical text categorization. We are interested in combining the analysis of text, ontology, and figures for document triage and annotation tasks.

In our future research, we shall investigate how human curators use figures in judging whether a document supports annotation, 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) it. As noted in the introduction, Mayer and Moreno (2002) examined the role of text and diagrams in understanding scientific literature and assessed whether visual information improves recall and problem-solving skills in human readers. They observe that properly organized multimodal presentations improve human performance in understanding the presented material. Given the condensed and informative nature of scientific images, and the rapidity in which humans perceive, process, and reach decisions based on such visual cues, we expect images in biomedical text to provide an invaluable support for categorization and mining of such text. We view text- and image- based document categorization as highly complementary, rather than competing approaches.

Our current results, along with these observations and the already accepted notion that database curators strongly rely on image data in articles to support their decision, strengthen our hypothesis that utilizing images can improve document categorization. Combining image analysis with text analysis is thus expected to help resolve ambiguity and improve the effectiveness of literature mining. The preliminary results presented here, from categorizing biomedical documents using both text and image data, further demonstrate and support this idea.

There are several challenges when applying document image analysis techniques for biomedical literature mining. In contrast to the millions of abstracts in MEDLINE, the number of full-text documents is still limited. Easy-to-use electronic versions (e.g. articles in XML format), with separately accessible figures and text are available for some papers, but not for all. For other cases (e.g. articles in PDF or image format), preprocessing has to be performed to separate text and figures, and to associate figures with figure captions. This preprocessing is difficult and error prone. Moreover, training and test data based on curation decisions is not available for individual images, but only for complete documents. We are actively pursuing ways to obtain labeled images that have been used by curators to determine the relevance/irrelevance of documents. We believe that having access to such data would form a major step forward in training classifiers that utilize image data for text categorization.

[3]TREC'04 participants noted that the data for the 2004 Triage task had some limitations. Additional categorization tasks, and different utility scaling measures were defined for TREC'05.

ACKNOWLEDGEMENTS

We thank Scott Brady for his kind help. We gratefully acknowledge the financial support provided by NSERC—Canada's Natural Sciences and Engineering Research Council, CFI—the Canadian Foundation for Innovation, and by the Xerox Foundation.

REFERENCES

B. de Bruijn and J. Martin. (2002). Getting to the (c)ore of knowledge: mining biomedical literature. *Int. Journal of Medical Informatics* 67(1-3). pp. 7-18.

J. Canny. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6). pp. 679-698.

N. Chen and D. Blostein. (2006). A Survey of Document Image Classification: Problem Statement, Classifier Architecture and Performance Evaluation. *International Journal of Document Analysis & Recognition*. (In Press).

K. Darwish and A. Madkour. (2005). The GUC goes to TREC 2004: Using whole or partial documents for retrieval and classification in the Genomics Track. *Proc of TREC 2004*. NIST Special Publication. pp. 362-369.

P. Duygulu, K. Barnard, N. de Freitas and D. Forsyth. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *In Proc. of the Seventh European Conference on Computer Vision*. pp. 97–112.

R.C. Gonzalez and R.E. Woods. (2002). *Digital Image Processing*. Prentice-Hall.

R.M. Haralick, K. Shanmugam, and I. Dinstein. (1973). Texture features for image classification. *IEEE Trans. On Systems, Man and Cybernetics*. SMC-3(6). pp. 610-621.

W.R. Hersh, R.T. Bhuptiraju, L. Ross, P. Johnson, A.M. Cohen, D.F. Kraemer. (2005). TREC 2004 Genomics Track overview. *Proc. of TREC 2004*, NIST Special Publication. pp. 132-141.

W.R. Hersh, A. Cohen, J. Yang, R.T. Bhuptiraju, P. Roberts, M. Hearst. (2006). TREC 2005 Genomics Track overview. *Proc. of TREC 2005*, NIST Special Publication. pp. 14-25.

L. Hirschman, A. S. Yeh, C. Blaschke and A. Valencia. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*. 6(Suppl 1):S1. pp. 1-10.

K. Huang & R.F. Murphy. (2004). From quantitative microscopy to automated image understanding. *Journal of Biomedical Optics*. 9(5). pp. 893-912.

A.K. Jain and A. Vailaya. (1998). Shape-based retrieval: a case study with trademark image databases. *Pattern Recognition*, 31(9). pp. 1369-1390.

R.E. Mayer and R. Moreno. (2002). Aids to computer-based multimedia and learning. *Learning and Instruction*, 12. pp. 107-119.

A. K. McCallum. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

T. Mitchell. (1997). *Machine Learning*. McGraw-Hill.

R.F. Murphy, Z. Kou, J. Hua, M. Joffe, W.W. Cohen. (2004). Extracting and structuring subcellular location information from on-line journal articles: the Subcellular Location Image Finder. *Proc. of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering (KSCE-2004)*. SLIF server web site: *http://goblin.cbi.cmu.edu:8080*

M.F. Porter. An Algorithm for Suffix Stripping (1997, Reprint). In *Readings in Information Retrieval*. Morgan Kaufmann.

Y. Regev, M. Finkelstein-Landau, R. Feldman, M. Gorodetsky, X. Zheng, S. Levy, R. Charlab, C. Lawrence, R.A. Lippert, Q. Zhang, H.Shatkay. (2002). Rule-based extraction of experimental evidence in the biomedical domain—the KDD Cup (Task 1). *SIGKDD Explorations* 4(2). pp. 90-92.

F. Sebastiani. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1). pp. 1-47.

H. Shatkay and R. Feldman. (2003) Mining the biomedical literature in the genomic era: an overview. *Journal of Computational Biology,* 10(6). pp. 821-855.

K. Tumer and J. Ghosh. (1996). Error Correlation and Error Reduction in Ensemble Classifiers. *Connection Sci.* 8(3-4). pp. 385-403.

H.R. Widlund, M.A. Horstmann, E.R. Price, et al. (2002). Beta-catenin-induced melanoma growth requires the downstream target Microphthalmia-associated transcription factor. *Journal of Cell Biology*. 158(6). pp. 1079-87.

I. H. Witten and E. Frank. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. (Describes Weka: The Waikato Environment for Knowledge Analysis. http://www.cs.waikato.ac.nz/ml/weka.)

A.S. Yeh, L. Hirschman, A.A. Morgan. (2003) Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup. *Bioinformatics*, 19. pp. i331-i339