Data and text mining

Figure mining for biomedical research
Raul Rodriguez-Esteban1,* and Ivan Iossifov2
1Systems Biology, Pfizer Inc., Cambridge, MA 02139 and 2Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11742, USA

ABSTRACT

Motivation: Figures from biomedical articles contain valuable information difficult to reach without specialized tools. Currently, there is no search engine that can retrieve specific figure types.

Results: This study describes a retrieval method that takes advantage of principles in image understanding, text mining and optical character recognition (OCR) to retrieve figure types defined conceptually. A search engine was developed to retrieve tables and figure types to aid computational and experimental research.

Availability: http://iossifovlab.csilu.edu/figureme/

Contact: raul.rodriguez-esteban@pfizer.com

1 INTRODUCTION

Biomedical articles are not just text. They include tables, graphs, charts, pictures and multimedia files. If we consider supplementary information, databases and web sites, an article’s content might be scattered in numerous locations and formats. While the properties of an article’s text have been dilated upon, the focus of the present study is on the figures and tables. Figures store experimental data often not available anywhere else. Due to their limit in number and costly production, figures are the subject of carefully curated content, including protocols or descriptions of biological processes. Tables store experimental results and data summaries in an aggregate manner, and they are sometimes represented using images or embedded within figures, blurring the difference between tables and figures. Tables can be mined for data sharing with technologies like the semantic web (Prkv et al., 2007).

Finding figures or tables with tools not designed for the task is not trivial. Three previous studies have considered the issue of figure retrieval: Sub-cellular Location Image Finder (SLIF; Murphy et al., 2004), BioText (Hearst et al., 2007a, b) and Yale Image Finder (YIF; Xu et al., 2008a). Both BioText and YIF perform text queries across different article sections, and YIF additionally searches through text embedded in the image. SLIF retrieves fluorescence micrograph images from a single-journal corpus. Micrograph images have regular properties that allow high-retrieval performance; therefore the design may not be flexible enough to be adapted to other figure types.

Figures convey information in similar ways by using a recognizable language, which is arguably part of the scientific language. The degree of coherence and articulation reflects the grammar that regulates the representation of layout and pictorial elements (e.g. text, colors, lines, shapes). Despite the regularities, there exists ample variability in figure design to hamper efforts at categorization. A figure may belong to several figure types (or none) by virtue of its variegated facets or its constituent sub-figures. This variability prevents applying simple descriptions to define figures. Rafkind et al. (2006) proposed a classification system that divided figures into five sets according to their coherence and frequency, similar to the division by Shatkay et al. (2006). We propose classes driven by research needs and we show that the flexibility of our methods allow for a less restricted approach than Rafkind et al. (2006) used for classification.

2 METHODS

The object of the study was a set of 80 949 articles from the digital archive PubMed Central (PMC). Some 233 395 images and 86 625 tables were extracted from this set and indexed using the open source Lucene technology1. The text indexed came from tables, table captions, figure captions and text within figure images. Additionally, the full text of articles was indexed separately to allow for more focused searches, e.g. searching for figures about the protein p53 within articles mentioning colon cancer.

Four figure types were selected in consultation with computational and experimental biologists in the area of systems biology. The types selected were gel, pathway, structure and time. Gel figures were defined as those depicting gel electrophoresis experimental results. Pathway figures were defined as diagrams representing interactions involving at least one protein. Structure figures described or depicted primary, secondary, tertiary and quaternary protein structures. Time figures plotted or listed data values over time.

A machine learning algorithm was trained to automatically annotate figures with the types defined. First, sets of randomly selected figures were manually annotated (Table 1). Then, figure descriptors were generated drawing from both text and image. The text was extracted from figure captions and from within figure images. Extracting text embedded in figure images required preliminary processing (Kou et al., 2007; Li et al., 2008; Xu et al., 2008b). This included cleaning non-textual elements using properties that characterize horizontal text objects: alignment, height–width proportion, character separation and character connectedness. Text with angle different from horizontal was not considered. After optical character recognition (OCR) using ABBYY FineReader 9.0, the text extracted and the caption text were tokenized and encoded separately with a method called set-of-words which consists in representing every distinct text token as a feature with value one. The figure image was processed to generate several feature sets (Kalva et al., 2007; Ritter and Wilson, 1996): color histograms, shape distributions and texture descriptors (similar to Rafkind et al., 2006). Color histograms were built from gray-scale, black-and-white, RGB and HSV color space frequencies. Shape distributions were measured with
Table 1. Figure type frequency within the annotated samples

<table>
<thead>
<tr>
<th>Figure types</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gel</td>
<td>353/44147</td>
<td>8.5</td>
</tr>
<tr>
<td>Pathway</td>
<td>204/10029</td>
<td>2.0</td>
</tr>
<tr>
<td>Structure</td>
<td>139/4005</td>
<td>3.5</td>
</tr>
<tr>
<td>Time</td>
<td>287/2001</td>
<td>14.3</td>
</tr>
</tbody>
</table>

The number of samples annotated was different for each type. The frequency of figure types changed as shown by Percentage; time figures being the most frequent and pathway figures the least. The Percentage column represents the same values as the Frequency column.

3 RESULTS AND DISCUSSION

Table 2 shows the performance of the trained algorithms. Retrieval of time figures yielded the lowest value, reflecting the difficulty of the task. Feature sets were of varying importance for each type, reflecting their different properties. For time figures, text within the figure was crucial because axis labels or measure units can be highly dependent. For class assignment, text immediately before and after the figure, and from full text, did not improve performance (data not shown). Table 3 shows query examples. Search options allow filtering for single journals and full text.

The tool presented can take the role of a specialized image repository, with the added capability of query searching. The repository of pathway drawings BioCarta contains 354 reference pathways, and the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000) contains 345. In comparison, we therefore directly comparable between sets with different class distribution.

Sobel operators. Texture descriptors were created from spatial gray level dependence statistics. All features were binarized using median values. Text and image features were combined in one vector per figure.

Support vector machines (SVMs) were selected as machine learning method (Chapelle et al., 1999). SVM implementations (Joachims, 1999) handle weighted cost functions, which can better deal with unbalanced datasets (i.e. datasets with classes distributed non-uniformly) (Morik et al., 1999). While classifiers are frequently designed to improve class-assignment accuracy, retrieval performance depends on precision and recall. In retrieval, maximizing accuracy is an optimal strategy only for balanced datasets. To deal with this problem, several strategies have been proposed (Chawla et al., 2004), such as increasing the penalty for misclassification of positive samples. In SVMs, this penalty can be reflected in the cost-factors for positive and negative samples (C+, and C−, respectively). For example, for C+ = 2C−, a misclassified positive example is as pernicious as two misclassified negative examples. The smaller the proportion of positive examples the larger C+ needs to be to compensate. A typical choice for the ratio C+/C− is the ratio of negative training samples over positive training samples.

SVM training and test sets were generated by 10-fold random sampling without replacement. Precision, recall, micro-averaged $F$-measure (van Rijsbergen, 1979) and area under the ROC curve (Fawcett, 2006) were used for evaluation. While $F$-measure is the commonest metric for retrieval performance, the area under the ROC curve is insensitive to class skew and therefore directly comparable between sets with different class distributions.

REFERENCES


http://www.biocarta.com