Artificial Intelligence in Medicine 49 (2010) 135–143

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Figure classification in biomedical literature to elucidate disease mechanisms, based on pathways

Natsu Ishii [a,*], Asako Koike [b], Yasunori Yamamoto [c], Toshihisa Takagi [a,c]

[a] Department of Computational Biology, Graduate School of Frontier Science, The University of Tokyo, 5-1-5 Kashiwano-ha, Kashiwa, Chiba 277-8568, Japan
[b] Central Research Laboratory, Hitachi Ltd., 1-280 Higashi-koigakubo, Kokubunji, Tokyo 185-8601, Japan
[c] Database Center for Life Science, Research Organization of Information and Systems, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan

ARTICLE INFO

Article history:
Received 15 March 2009
Received in revised form 26 March 2010
Accepted 29 March 2010

Keywords:
Figure classification
Text mining
Supervised machine learning
Disease-related pathways
Multi-factorial disorders

ABSTRACT

Objective: As more full-text biomedical papers are becoming available in digitized form online, there is a need for tools to mine information from *all* parts of such papers. Because the figures and legends/captions in biomedical papers provide important information about research outcomes, mining techniques targeting them have attracted a great deal of attention. In this study, we focused on pathway figures that illustrate signaling or metabolic pathways, because many of these are important in understanding disease mechanism(s). We developed a figure classification system based on textual information contained in biomedical papers to provide an automated acquisition system for such pathway figures.

Materials and methods: We used full-text journal articles available on PubMed Central as our data set. We used several supervised machine learning methods, such as decision tree and a support vector machine, to classify figures in the data set. We compared the classification performance among the cases using only figure legends, using only sentences referring to the figure in the main text of the article, and combining figure legends with sentences referring to the figure in the main text of the article.

Results: Compared with previous related work, a sufficiently high performance was achieved with the figure legends alone. The performance with the sentences referring to the figure in the main text was actually lower than that with the figure legends alone, indicating that focusing on the main text alone is inadequate. The combination of legend and main text clearly had an effect, but including the prior and following sentences in addition to the sentence referring to the figure dramatically improved the performance.

Conclusions: We developed an automatic pathway figure classification system based on both figure legends and the main text that has quite a high degree of accuracy. To our knowledge, this is the first attempt to address a figure classification task using legends and the main text, and it may provide a first stage for achieving efficient figure mining.

1. Introduction

The importance of text mining techniques is unquestionable given the exponentially growing number of biomedical papers. Many efforts have been made to cope with the flood of textual information in this domain, but most have focused on abstracts, such as MEDLINE. As more full-text biomedical papers are becoming available in digitized form online, there is a need for tools to mine information from *all* parts of a paper, including figures, figure legends, and tables.

Biomedical papers often contain many figures. Futrelle et al. [1] counted the number of words contained in figure legends/captions and in sentences in the main text and found that 50% of the content of typical biological papers was figure related. Notably, because figures and their legends/captions, collectively called "legends" hereafter, in biomedical papers provide important information about research outcomes, mining techniques targeting them have attracted a great deal of attention. Liu et al. [2] developed a figure legend indexing and classification system, FigSearch. They defined schematic representations of protein interactions and signaling events in biomedical literature as a figure type of interest, took a supervised machine learning approach to classify figures, and performed indexing of figures using figure legends. However, they did not use the main text referring to the figures. Shatkay et al. [3] used graphical features of images combined with the text of

* Corresponding author. Tel.: +81 4 7136 3973; fax: +81 4 7136 4100.
E-mail addresses: natsui@cb.k.u-tokyo.ac.jp (N. Ishii), akoike@hgc.jp (A. Koike), yy@dbcls.jp (Y. Yamamoto), tt@k.u-tokyo.ac.jp (T. Takagi).

0933-3657/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2010.04.005

PubMed abstracts for biomedical document categorization to determine which documents were relevant to a given annotation task performed by Mouse Genome Database curators. However, their aim was not to analyze individual figures but to characterize each document based on the figures included. In other words, they were not interested in what each figure represented. Accordingly, they used neither figure legends nor textual information in the figures for their document classification. Murphy and co-workers [4–6] dealt with fluorescence microscopy images, which contain information about the distribution of proteins and other biological macromolecules inside cells. They set as a goal the building of a knowledge base system that could interpret such images in online journals. They used graphical features of images to find and interpret figures of interest from journals and textual information in legends to separate multiple images contained in one figure, but the main text was not used.

Here, the focus was on pathway figures among the variety of figures contained in biomedical papers, because multiple proteins in a specific pathway are often related to the causes of multifactorial disorders, and it is important to obtain information about pathways to understand the mechanism(s) of such disorders. For example, PSEN1, SNCA, and APOE are all genes involved in Alzheimer's disease, and they are in the same pathway with APP (amyloid beta A4 precursor protein), which has been reported to accumulate in the brain of Alzheimer's patients [7]. When there is a new experimentally determined candidate gene, one may predict that a gene is related to the same mechanism of a certain disease if it is in the same pathway as a known one. Otherwise, one may predict another mechanism of the disease by considering the function of that gene in another pathway. Indeed, pathway information is important in understanding diseases. The need for systems to retrieve pathway information is high, but construction of manually curated pathway databases (e.g., KEGG [8]) is generally laborious and time consuming. Furthermore, according to our investigation, not all information in typical pathway figures is always contained in either the legends or the main text; many pathway figures contain more gene/protein names than the legend or the main text. A complementary relationship among the figure, the legend, and the main text is especially strong in such pathway figures, so it is necessary to mine information efficiently from all of them to understand pathways.

As a first stage toward such an efficient pathway figure mining process, we focused on figure classification to automatically select pathway figures that met our definition (i.e., directed graphs composed of more than two steps that represent signaling or metabolic pathways). Unlike the document classification task mentioned above [3], the classification of individual figures requires a more sophisticated method, in that one needs to consider the specific text regarding each figure rather than the document as whole. Furthermore, a document generally contains both desired and undesired figures. Thus, it is quite difficult to efficiently extract information about figures of interest. Moreover, the graphical features employed by Shatkay et al. [3] or Murphy and co-workers [4–6] could not be readily applied to our task because describing our pathway definition in graphical features is difficult with the current state of image processing techniques. Rafkind et al. [9] used graphical features for figure classification, but their classification was rather broad, and they did not define a detailed category such as pathways. Thus, we used textual information in figure legends and in the main text referring to the figures to automatically choose pathway figures. As described in the next section, we took a machine learning approach to classify figures.

Our ultimate goal is to develop a figure finding system, which we call FigFinder, to retrieve figures relevant to a user's query (gene/protein or chemical compound names) by mining information contained in figures, their legends, and the main text in an integrated manner. The classification system introduced here may be the first stage in developing such a system.

2. Materials and methods

2.1. Full-text articles

We chose five journals available on PubMed Central [10]: Biochemical Journal (2004–2005), BMC Developmental Biology (2001–2005), BMC Molecular Biology (2000–2005), PLoS Biology (2003 to June 2005), and Proceedings of the National Academy of Sciences of the United States of America (November 1996–2001, 2003, 23 November 2004 to 15 February 2005). The total number of articles was 16,471, in which there were 75,350 figures in JPEG format and related legends. We converted each HTML full-text paper to XML format using an internally developed XML converter.

2.2. Positive and negative data

According to our pathway definition described in the previous section, we manually checked the 75,350 figures and identified 375 pathway figures to be positive data. Another 11,251 figures other than pathway figures were randomly selected as negative data. This is because too small a proportion of positive data takes a long time to learn and often leads to improper learning. Figure samples that were included in positive data are shown in Fig. 1. Fig. 1(a) represents a model for the role of Akt in IL-2 signaling in a normal cell. However, mutations in some genes in this pathway induce certain diseases. For example, a mutation in Jak1 causes acute leukemia, or lymphoma; mutations in c-myc, bcl-2, and Akt trigger various cancers. Fig. 1(b) is the interaction of EGF with EGFR and the downstream events of NF-κB activation in breast cancer cells. Fig. 1(c) shows target genes of β-catenin-T-cell factor/lymphoid-enhancer factor complex and the related cellular processes in human colorectal carcinomas. As can be seen in these figures, many signaling pathways are related to diseases, even if they are not necessarily stated to be disease related. In contrast, Fig. 2 shows samples that appeared similar to pathway figures but were considered to be negative because they did not meet our pathway definition. Fig. 2(a) is composed of fewer than two steps. Fig. 2(b) represents not pathways but interactions and complexes. Fig. 2(c) is not a directed graph. Negative data also contained diagrams, fluorescence microscopy images, and gel photographs. For each figure in both the positive and negative data, we obtained the legend from the XML-formatted full-text paper. For all the data we used, the list of article IDs in PubMed Central along with information about which figures we used is available at http://marine.cb.k.u-tokyo.ac.jp/~natsui/.

2.3. Feature word selection

We selected feature words to represent figures in our data set from among words contained in all positive and negative legends. Stopwords were first removed from figure legends using a stopword list provided by NCBI [11]. It included 132 words that appear so frequently that they are ignored in the indexing of PubMed abstracts. Next we stemmed the remaining words using the Porter Stemmer algorithm [12]. This algorithm removes suffixes from words and leaves the stem (e.g., *pathway* or *pathways* becomes *pathwai*). Then we counted the frequency of each word in the data set and excluded 4% of the low-frequency words because they are also considered to be of no value in indexing [13]. Words that were composed of fewer than three letters were also excluded to remove abbreviated words with multiple meanings (homonyms). Then we calculated chi-square statistics (CHI) and