Enhancing Information Accessibility of Scientific Publications with Text Mining and Ontology

Weijia Xu   Amit Gupta   Pankaj Jaiswal   Crispin Taylor   Patti Lockhart

Texas Advanced Computing Center
University of Texas at Austin
Austin, Texas USA
{xwj, agupta}@tacc.utexas.edu

Department of Botany and Plant Pathology
Oregon State University
Oregon, Portland, USA
jaiswalp@oregonstate.edu

American Society of Plant Biologists
Rockville, Maryland, USA
{ctaylor, plockhart}@aspb.org

Abstract— **We present an ongoing effort on utilizing text mining methods and existing biological ontologies to help readers to access the information contained in the scientific articles. Our approach includes using multiple strategies for biological entity detection and using association analysis on extracted analysis. The entity extraction processes utilizes regular expression rules, ontologies, and keyword dictionary to get a comprehensive list of biological entities. In addition to extract list of entities, we also apply natural language processing and association analysis techniques to generate inferences among entities and comparing to known relations documented in the existing ontologies.**

Keywords—component*;* ***Information systems applications; Ontology; Text Mining; Association Analysis***

I. Introduction

Due to its technical depth and rich, informational content, a journal article often requires that readers, domain experts, and curators invest significant amounts of time and effort to fully comprehend and make intelligent use of its content. This can be especially true in emerging areas, where novel ideas and new terminologies may be presented without precedent. As new technologies accelerating scientific discovery and more content becomes available online, the number of new articles that must be read and understood continues to rise. There are over 22 millions references indexed by MEDLINE. Therefore, there is a pressing need to develop computational methods and tools that can facilitate the readers' understanding of the content of the publicaiton.

To address this challange, we present software developments from an ongoing project, DIVE, which features auto extraction of informational vocabulary, web based access and curation tools. The framework implements several strategies in entity extraction, including using regular expression rules, ontology and a keyword dictionary. The results of the extracted biological entities are then stored in a database and made accessible through an interactive web application for curation and evaluation by authors and other domain experts. Additional text mining and associaiton analysis can be run on extrated entities to help readers understanding of the paper. The system can benefit the entire life cycle of the digital publication, from initial manuscript submission to publishing the article and presenting information to readers. New information defined and verified by experts may also be injected to other information resources.

II. Methods and Implementaionts

A. *Entity Detection Workflow*

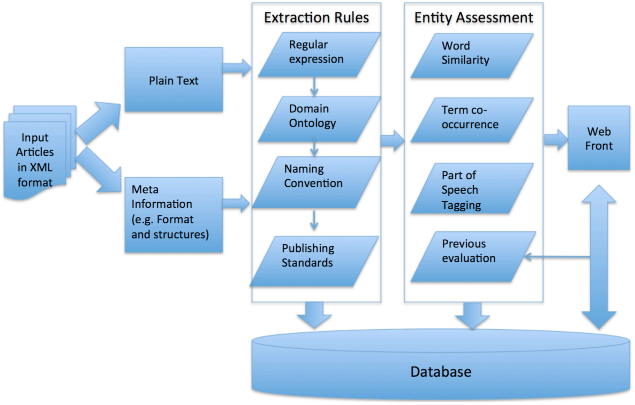

Figure 1. Processing workflow overview

Figure 1 shows the overview of our processing workflow. There are three major steps in processing the documents: text extraction, entity candidate extraction, and candidate assessment.

1) Text Extraction

The text extraction process the input structured document tagged by JATS [1]. During this step, the input document will be processed into two data structures for textual data and structural data. The textual information data includes as a list of string representation of the body of text included in the journal articles. The structural data includes the metadata information presented at the input document, such as section mark, special formatting mark etc. A mapping is maintained between the textual value and metadata information by their global positions in the original documents. This dual data structure allows for efficient text processing of the publication content while still being able to easily retrieve the meta structure around a particular set of words during the subsequent steps of processing.

2) Entity Candidate Detection

We implemented a rule-based approach for processing the text and structure in order to identify informational vocabulary candidates. The detection rules can be defined based on various heuristics and requirements such as publishing requirements, naming conventions, and domain ontologies.

New rules can be added on demand over time. Currently, there are four types of rules implemented in the DIVE, regular expression rules, word dictionary, publishing convention, and ontology rules.

The regular expression rules utilize common naming conventions to identify biological entities, such as gene name, protein name, molecule structures, chemical compound, etc. Each rule can be defined as a regular expression and used for matching the candidate word. The word dictionary rule consists of a pre-defined list of words that should be included or excluded in the candidate lists. The publication content is searched against the list at run time. The publishing convention rules are used to identify words that are in special format, such as in italic, or in a particular component of the publication, such as a figure legend. The enclosing tags of the candidates are used to define each rule. Additional rules can be added by specifying additional tag values or by using naming conventions to detect entities like species names. The ontology rules utilize five biological ontology including gene ontology [2], plant ontology [3] plant trait ontology [4], plant environment condition ontology [5] and Chemical Entities of Biological Interest (ChEBI) [6].

3) *Entity Candidate Assessment*

By applying the extraction rules listed above, a set of entity candidates can be detected from the input document. Some candidates might be detected by multiple rules. Different detection rules also have different accuracy. Ontology file and dictionary based approaches have the highest certainty. Candidates only identified by other rules need further validation. We currently implemented two automatic validation mechanisms. One is based on the previously validated results; the other one is based on co-location with other confirmed entities. However, the primary method of validation is by domain expert evaluation through the web interface, which is detailed in the following section.

B. Association Analysis

The data association analysis can be used to generate inferences between values from two or more fields of the data in a given condition using FP-Growth algorithm[8].The analysis starts with selecting and aggregating subset of data specified by the input parameters as a list of records, also known as *transactions*. The analysis algorithm will scan the selected data set to compute the frequency of each value, also referred as an *item*, and store the frequency value and co-occurrence with other *item, collectively referred as itemset,* in a tree structure, named frequent pattern tree (FP-tree). Then the frequent item sets can be identified from the FP-tree to generate inferences among subset of values.

III. PRELIMINARY RESULTS

Figure 2 shows top 20 inference rules based on all ontology terms extracted from the collection. Each label indicates a frequent item set found in the collection. The directional arrow indicates an inference on co-occurrence between two item sets. The shade of the directional arrow indicates the confidence level of the rule. Such visual representations of inferred association between diverse entity types could tremendously aid a researcher in forming insights. This also has potential to be a similarity metric between articles that could help editors gauge the novelty of a new article submission.

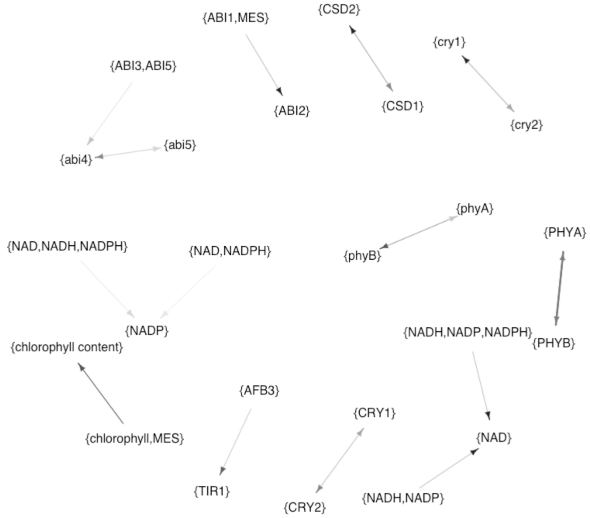

Figure 2 Top 20 inference rules from association analysis.

We are continuing working on evaluating the performance of the entity extraction over large data set and improving its accuracy. We are gathering feedback from domain researchers and publishing professionals for further entities candidate evaluations. We are also working on comparing the inferences from association analysis with known relationships documented in the existing ontologies.

ACKNOWLEDGMENT

DIVE is partially supported by CyVerse (NSF award DBI-0735191 and DBI-1265383) and the Gramene, a Comparative Plant Genomics Database (NSF award IOS-1127112).

REFERENCES

[1] National Center for Biotechnology information. *Journal Article Tag Suite*. http://jats.nlm.nih.gov/, 2013.

[2] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis A.P. "Gene Ontology: tool for the unification of biology." *Nature genetics* 25, no. 1 (2000): 25-29.

[3] Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., Zapata, F. (2005). Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comparative and Functional Genomics*, *6*(7-8), 388–397. http://doi.org/10.1002/cfg.496

[4] Arnaud, E., Cooper L., Shrestha, R., Menda, N., Nelson, R.T., Matteis, L., Skofic M. (2012) Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes in *KEOD*, pp. 220-225.

[5] Plant Enviroment Condition Ontology, http://bioportal.bioontology.org/ontologies/PECO#

[6] Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, *36* (Database issue), D344–D350.

[7] Cooper, L. and Jaiswal, P. (2016) The Plant Ontology: A Tool for Plant Genomics. *Plant Bioinformatics: Methods and Protocols*, 89-114

[8] Han, J. Pei, J. and Yin, Y. "Mining frequent patterns without candidate generation," in *ACM Sigmod Record*, 2000, vol. 29, no. 2, pp. 1–12.