Building Scholarly Data Forest

Marko Požega, Dario Poljak, Kristina Kocijan

Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
{mpozega1, dpoljak1, krkocijan}@ffzg.hr

Abstract. In this paper, we will demonstrate syntactic analysis and visualization of scientific data, namely references from scientific papers. Our main goal is to build a parser which could extract references from scientific papers, convert them to XML format, send to custom visualization algorithm and present in a web interface as a *ReferenceTree* for a single author. For this process, we use several different technologies such as NLP software NooJ, programming languages PHP and JavaScript in combination with HTML5. Our main problem was dissimilarity in reference styles between articles. Thus, our parser was designed to recognize different reference source (book, paper, web page) in APA, MLA and Chicago reference styles. As for the visualization idea, we have chosen the concept of presenting an author as a tree, the publication years as the main branches, the articles/books as twigs and references used in each article/book as the leaves. The books are grouped on the left side of the tree while the articles are grouped on the right side. With final output, every processed author should have a unique tree (preferences of references) and could be compared with the rest of the scientific forest.

Keywords. Scholarly data, network visualization, contact trees, egocentric networks, NLP, APA, MLA, Chicago reference style, science mapping, *ReferenceTree*.

1 Introduction and related work

When we talk about the tree view of a network type data it is very often that we are talking about the trees that are using connected nodes with either a top-down or right-to-left orientation. Sometimes this so called 'tree view' has a rather circular shape network or matrix. Although informative breadth-wise, such visualizations are usually very modest in the depth of information they are able to show. It is quite recently that a new tree view visualization has been proposed [1] with a (real) tree shape visualization called ContactTrees (since they were originally designed to show person's social ties). Such trees have an egocentric approach with the ability to show multilevel aspects of social interactions in just a glance [1, 2, 3, 4, 5] that may be of help to sociologists as well as data managers as suggested by [2].

Our work is very much inspired by the work presented in [2]. We applied a similar approach in building our scientific reference trees which we present here as a new tool

for science mapping as defined in [6]. However, our main concern is to show which papers an author has cited throughout his/her academic career, rather than to visualize scientific disputes among different authors or their co-publishing behaviors.

In the sections that follow we will explain in more details steps involved in building *ReferenceTrees* starting with the data and an NLP tool we used for building syntactic grammars for automatic recognition and classification of references and finishing with the more detailed description of a tree. We will conclude the paper with some additional future work ideas.

2 Recognizing the references

We built syntactic grammar for reference recognition with an NLP tool - NooJ[1] constructed by Silberztein [7]. NooJ provides a graphical editor for building powerful syntactic grammars (graphs) that are well suited for our purpose. It allows us to create functional but also visually understandable grammars (Fig. 1). Each graph uses nodes that can be NooJ or regular expressions, plain text or even variables. It also uses the strength of a transducer and enables us to produce customized output such as XML like notation of the data needed for our *ReferenceTrees* (Fig. 3).

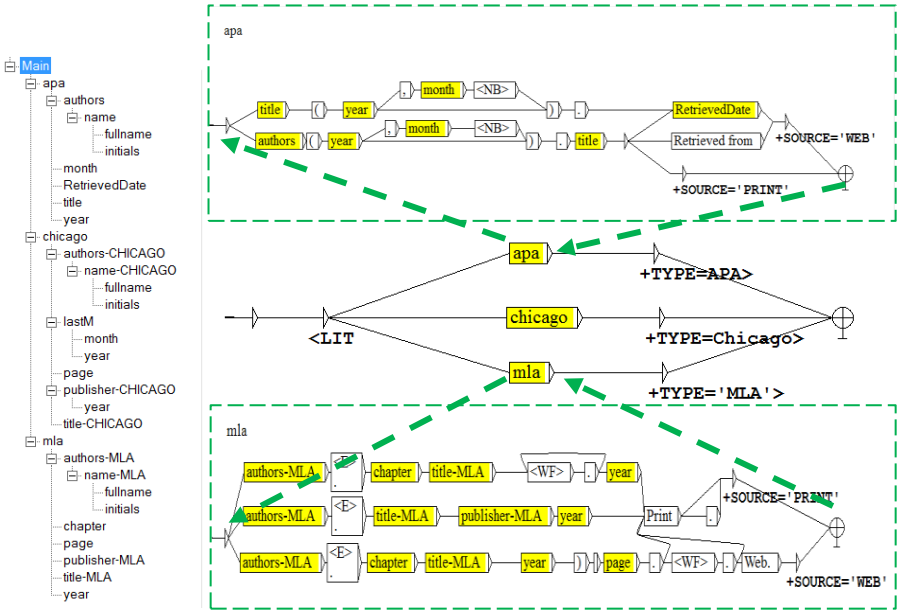

Fig. 1. The structure of the Main Graph with APA and MLA subgraphs

The main syntactic grammar is built with many smaller subgraphs (month, year, page, etc.) some of which are reused at several positions making the grammar easier and

[1] NooJ can freely be downloaded from http://www.nooj4nlp.net/

faster to write and maintain (Fig. 1). Not all three reference styles use as simple and concise a grammar as the one we built for APA. They actually grew in the complexity and required some additional nodes in order to perform according to the requirements of each style (compare Fig. 1 and Fig. 2).

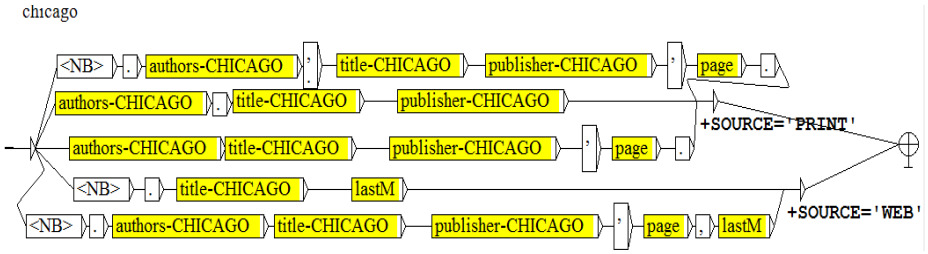

Fig. 2. Subgraph for recognition of Chicago style reference

Our grammar has been trained on the *University of Pittsburgh* and *The Purdue Online Writing Lab* sets of data pertaining APA, MLA and Chicago citing styles. Both sites explain various ways of citing works such as books, articles and websites with examples for each of the citing styles. We finished the testing phase when our grammar reached the f-score of 1. After the parsing, the concordance window provides us with the references found but also with the XML-like output (Fig. 3) that consists of attribute=value sets. This kind of output can easily be exported and managed through other programs.

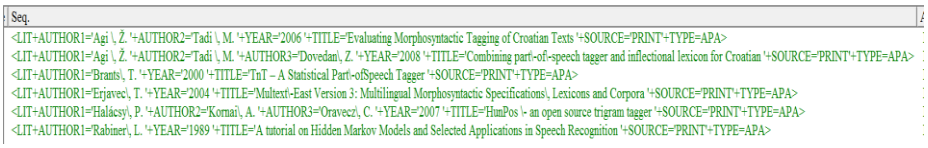

Fig. 3. NooJ XML output

3 Building the trees

Our website[2] uses PHP 5 powered by Apache and JavaScript with addition of free JavaScript vector library called Raphaël made by Dmitry Baranovskiy [8]. The site is divided into three main sections: homepage (basic information and user instructions), the core of the site (uploading documents and generating the tree), and the public *ReferenceTrees* section (displaying trees that the authors have made public). We will describe here the middle section, i.e. the tree structure and generation of the tree.

As already noted in the introduction, we imagine our visualization as a realistic tree (Fig. 4) that gives an overview of all the books (left half of the tree) and articles (right half of the tree) written by an author. Each main branch (both sides) stands for a year when the book/article was published. On each main branch there are twigs representing

[2] URL: www.ikstudenstkiprojekti.ffzg.hr/ReferenceTrees/index.php

a specific book/article. Twigs presenting papers written only by an ego are positioned on the upper side of the main branch, while the papers written in co-authorship are on the lower side. References used in a book/article are shown as leaves of each twig. Leaves are color coded depending on a type of a reference (book, article or a web site). The tree sections are animated giving more information when selected. If an author has used the same reference in more of his/her papers, all of the matching references are highlighted upon the selection of any one of them. Although at this point, only the individual trees may be explored, we feel that this is the first step in building and exploring author reference networks.

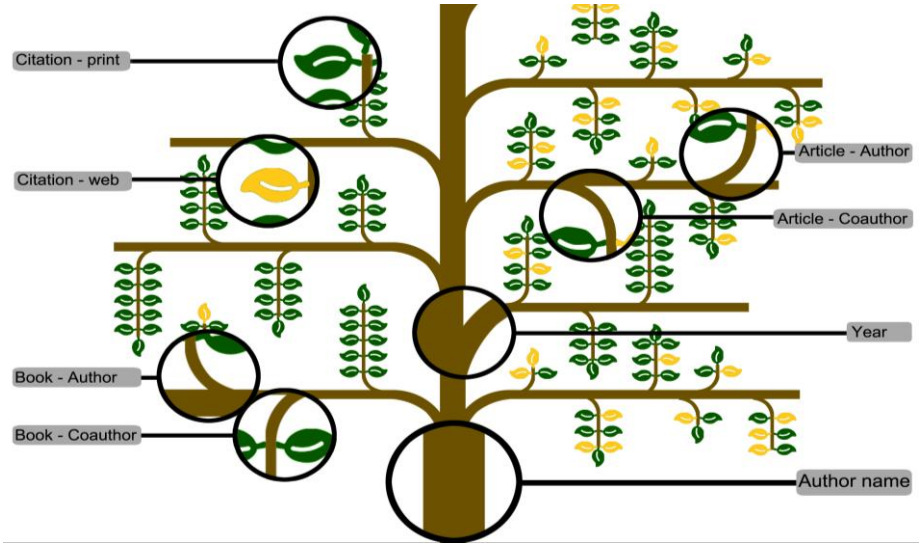

Fig. 4. Explanation of the Tree structure parts

We can explain the procedure of building the *ReferenceTree* via PHP with the following six main steps[3]:

1. Extract data from database to JavaScript/JQuery array $dataBooks
2. Calculate height of vertical tree branch using data from array $totalHeight;
3. Draw main vertical tree branch with width of 90px and height of $totalHeight;
4. Iterate through $dataBooks and for every $year draw horizontal branch on the left;
5. Repeat step 4 but use $dataPapers and draw horizontal branches on the right;
6. (a) For each $Reference in $dataBooks and $dataPapers draw vertical twigs on the year branch and calculate twig height;
   (b) With each previous iteration create and position leaf SVG DOM element from Raphael.js library using CSS and color it depending on the source type.

[3] Due to the length and complexity of real and pseudo codes used, in this paper we are only giving the main steps while the visual demo and JavaScript source code are available at: http://www.ikstudenstkiprojekti.ffzg.hr/CitationTrees/exampleTree.php.

4 Conclusion and future work

We have presented a tree-shape *ReferenceTrees* model for visualizing bibliographies used in scientific books or articles by an author (ego). We have managed to incorporate multiple dimensions (author, year of publishing, type of publication, authorship or co-authorship, number of references, source of a reference, repeated or a unique reference among all the published works) into one relatively simple representation - tree. In this process, we have taken few steps (parsed the text, extracted the data, build the trees) and used several technologies (NooJ, XML, HTML5, JavaScript, PHP) so that we can produce as complete a tool for building reference trees as possible.

We see many opportunities in advancing our *ReferenceTrees* proposed in this paper. As our future work, we are considering the ways to incorporate the size and the shape of a leaf to show some additional characteristics to our trees (information about the scientific field of the article/book, or co-reference relations and self-citations). Also, the color and the thickness of a twig may be used to show how many times that specific publication has been cited by others in our database or the language of a publication (it would be interested to see in how many languages an author publishes). This information may be further used in placing the trees with similar structures closer to one another in a scientific forest, or the forest may switch on the lights of the trees that use references belonging to specific branch of a science e.g. linguistics or even more specific e.g. morphology. Taking into account all the possibilities our *ReferenceTrees* offer, we believe that they may find their usage in digital library catalogues, or scientific social networking sites but may also give another perspective to scientific development as a whole.

5 References

1. Sallaberry, A., Fu, Y.-C., Ho, H.-C., Ma, K.-L.: ContactTrees: A Technique for Studying Personal Network Data. *CoRR, abs/1411.0052* (2014)
2. Fung, T.-L., Ma, K.-L.: Visual Characterization of Personal Bibliographic Data Using a Botanical Tree Design. In: Electronic Proceedings of IEEE VIS 2015 Workshop on Personal Visualization: Exploring Data in Everyday Life, http://www.vis4me.com/personalvis15/papers/fung.pdf (2015)
3. Fung, T.-L., Chou, J.-K., Ma, K.-L.: Comparing Characteristics of Majors Using Egocentric Botanic-trees, http://vacommunity.org/ieeevpg/viscontest/2015/entries/6.html (2015)
4. Sallaberry, A., Ma, K.-L.: Visualizing InfoVis Researchers with ContactTrees, http://web.cse.ohio-state.edu/~raghu/teaching/CSE5544/Visweek2012/infovis/posters/sallaberry.pdf (2012)
5. Sallaberry, A., Fu, Y.-C., Ho, H.-C., Ma, K.-L.: Contact Trees: Network Visualization beyond Nodes and Edges. PLoS ONE 11(1): e0146368. doi:10.1371/journal.pone.0146368 (2016)
6. Chen, C., Dubin, R., Schultz, T.: Science Mapping. In: Khosrow-Pour, M. (ed.) Encyclopedia of Information Science and Technology, Third Edition. IGI Global. DOI:10.4018/978-1-4666-5888-2.ch410. (2014)
7. Silberztein, M.: NooJ Manual. http://www.nooj4nlp.net (223 pages) (2003)
8. Baranovskiy, D.: *Raphaël* -JavaScript Library, http://raphaeljs.com [cited 2016 Jan 17]