Serbian Legislation as a Network

Dragan Vidaković, Stevan Gostojić, Aleksandar Kovačević
Department of Computing and Control Engineering, University of Novi Sad, Novi Sad, Serbia
{vdragan, gostojic, kocha78}@uns.ac.rs

***Abstract*—The recently introduced concept of a legislation network is a promising approach to addressing the challenging issues of tackling with and quantifying the complexity of the legislation. In this paper, we described the process of crawling and scraping Serbian legislation, applying natural language processing methods to detect references in legislation and applying network science methods (e.g. node centrality and community detections) to quantify network properties. The quantified network properties were attributed with corresponding legal interpretations.**

***Keywords*—legislation network; legislation; network science; natural language processing; computational legal studies;**

I. INTRODUCTION

As legislation increases in size and complexity, finding relevant legislation becomes a challenging task even for experts. Analyzing large number of documents and their relationships is a complex and labor-intensive task. Due to the nature of legislation, where one piece of legislation contains references to other pieces of legislation, the legislation system can be viewed as a network. Natural language processing (NLP) techniques and network science techniques can be used to identify important legislation and similar legislation and thus facilitate more effective retrieval and browsing of legislation.

The specific problem our work addressed can be divided into three subproblems:

1. crawling and scraping legislation and metadata,
2. using NLP methods to detect references in legislation, and
3. applying network science methods, such as node centrality and community detection, on a legislation network and giving it legal interpretation.

The rest of this paper is organized as follows. In section II, several related works are presented and discussed. In section III, we describe the legislation collection process, development of Serbian legislation network and the algorithm for community detection. In section IV the results are presented and analyzed. Finally, we make a brief concluding mark and give the future work in Section V.

II. RELATED WORK

Taxonomy of reference types, that is valid nowadays was proposed by Berger [1]. He differentiates between four reference types: fully-explicit, semi-explicit, implicit and tacit references and proposes more than 50 parameters to distinguish between those reference types. Waltl, Landthaler and Matthes [2] proposed an extensible model, based on text mining, for distinguishing between reference types. Evaluation on German legislation corpus revealed that the model accurately detects full-explicit and implicit references but should be improved when detecting semi-explicit references. Winkels et al. [3] identified explicit references in Dutch case law with an intention to develop a recommender system where users of a legislative portal receive suggestions of other relevant sources of law, given a focus document.

Zhang and Koppaka [4] proposed a semantic-based legal citation network and discussed the semantic multi-dimensionality of legal citations from a data science perspective. Boulet et al. [5] used mathematical methods to analyze the structure of French legal citation networks. Agnoloni and Pagallo [6] investigated the relevancy of case law using a citation network obtained from the Italian Constitutional Court. Fowler et al. [7] applied network analysis techniques to find the most relevant precedents using a citation network obtained from the United States Supreme Court. Sakhaee, Wilson and Zakeri [8] identified several New Zealand legislation networks and used those networks to test several legal and political hypotheses.

III. METHODOLOGY

The legislation in force is published by the Legal Information System[1]. It represents the biggest legislation database which contains the only valid and official legislation in its original, complete and accurate form. Beside legislation text, it contains metadata and occasionally a list of references to other legislation. Metadata for each legislation contains its title, type, area, group within the area, date of enactment, publication service and date of publication, and the date of application.

In order to form a document collection consisting of legislation in force (and associated metadata) we developed a crawler and a scraper. A crawler implemented using Python bindings of Selenium[2] software-testing framework, simulates user actions by visiting the Legal Information System, filling search forms and opening result pages. Then, a scraper extracts data from an HTML document, collecting document's text, corresponding metadata and available references to other documents. Titles of all documents are extracted from metadata files and aggregated into separated file for reference extraction process.

An NLP pipeline for the extraction of full-explicit references (a reference type with full information about the referenced document) is shown in Figure 1. The input to the pipeline is an HTML document. Firstly, the text is

[1] http://pravno-informacioni-sistem.rs/SlGlasnikPortal/reg/content
[2] http://selenium-python.readthedocs.io/

converted from Cyrillic to Latin alphabet and non-English characters are replaced with their ASCII pairs (č with c, ž with z etc.). The text is then tokenized and stemmed using the Serbian language stemmer created by modifying the Croatian language stemmer[3]. Fourthly, the stems are joined using space as a separator. In the next step, full-explicit references are identified by performing a full-text search of document titles (transformed on the same way as an HTML document) over transformed text. Finally, references are aggregated accross all of the processed documents. The output of the pipeline is a set of automatically assigned numerical identifiers representing documents fully-explicit referenced from the input document.

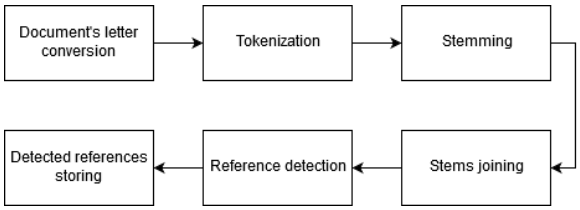

Figure 1. Reference extraction NLP pipeline

Serbian legislation network can be viewed as a directed graph with legislation as nodes and full-explicit references between legislation as links. Python language software package NetworkX [9] was used for creation and manipulation of the network, and calculation of basic network scientific measures.

We used Latapy and Pons algorithm [10] based on random walks to detect communities in Serbian legislation network. This algorithm simulates many short random walks on the network and computes pairwise similarity measures based on these walks. Similarities are later used to aggregate nodes into communities. Time complexity of Latapy and Pons algorithm (O(|V|^2 log |V|) makes it fast and applicable to complex networks.

IV. Discussion

In order to evaluate the accuracy of the reference extraction pipeline, full-explicit references were manually detected in ten randomly selected documents. The accuracy of the reference extraction process was 90%. Since the legal documents are well structured, and have unified referencing, ten is representative number for the evaluation.

In brief, developed Serbian legislation network contains 5,391 nodes and 17,343 edges. The highest degree nodes, representing documents with most references, are shown in Table 1. The average node degree is 3.217.

Document	In Degree	Out Degree	Degree
Law on the Government	1,482	10	1,492
Law on planning and construction	284	35	319
Labor Law	290	20	310
Law on environmental protection	282	20	302
Customs tariff Law	289	0	289

Table 1. Highest degree nodes

[3] http://nlp.ffzg.hr/resources/tools/stemmer-for-croatian/

The diameter of a network, representing the greatest distance between any pair of nodes is 14. The average distance between two nodes (network path length) is 4.7. The network density (the ratio of the number of links and the number of possible links) is 0.0006. The global clustering coefficient (the measure of the degree to which nodes tend to cluster together) is 0.0985. Serbian legislation network contains a cycle consisting of 7 documents, shown in Figure 2.

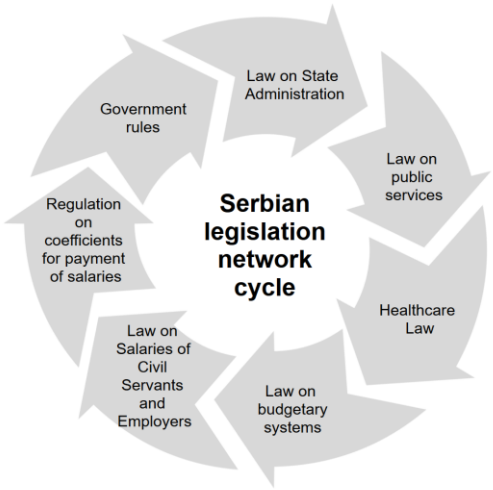

Figure 2. Serbian legislation network cycle

In order to determine the relative importance of particular legislations, we calculated five node centrality measures: PageRank [11], Katz Prestige [12], eigenvector centrality [11], closeness centrality [13], and degree centrality [13], shown in Table 2 (X denotes rank greater than 20). Eigenvector centrality computes the centrality for a node based on the centrality of its neighbors, giving greater score to nodes connected with other high-scoring nodes. As we can see from the Table 2, the top ranked nodes by eigenvector centrality measure are highly ranked by Katz Prestige, PageRank and closeness centrality measures as well. Only degree centrality measure deviates from the others.

Document	Eigenvector	Katz Prestige	PageRank	Closeness	Degree
Law on Administrative Procedure	1	1	3	1	8
Company Law	2	2	1	6	6
Labor Law	5	5	5	3	3
Law on Obligations	3	3	6	8	X
Constitution of the Republic of Serbia	15	19	7	5	7
Law on the Government	X	X	2	17	1
Data Secrecy Law	8	10	13	10	X
Law on State Administration	16	X	X	2	18

Table 2. The most highly ranked documents

Community detection algorithm detected 141 communities in the network, shown in Table 3. We displayed top 8 communities containing 78% of network nodes. To label a community, we counted document's

area keywords inside the community. Keywords with the highest count were chosen as the label of the community.

Community label	Number of nodes
State regulation and army	1687
Public incomes, security, ecology and environment	985
Decentralization and development	502
Labor, employment and taxes	334
Commerce	288
Social insurance and health care	188
Monteary system and finances	129
Agriculture	105

Table 3. The most numerous communities

The graphical representation of the Serbian Legislation network was obtained by using PyGraphistry[4] visual analytics library for big graphs visualization and manipulation. Graphical representation with colored communities and enlarged high degree nodes is shown in Figure 3.

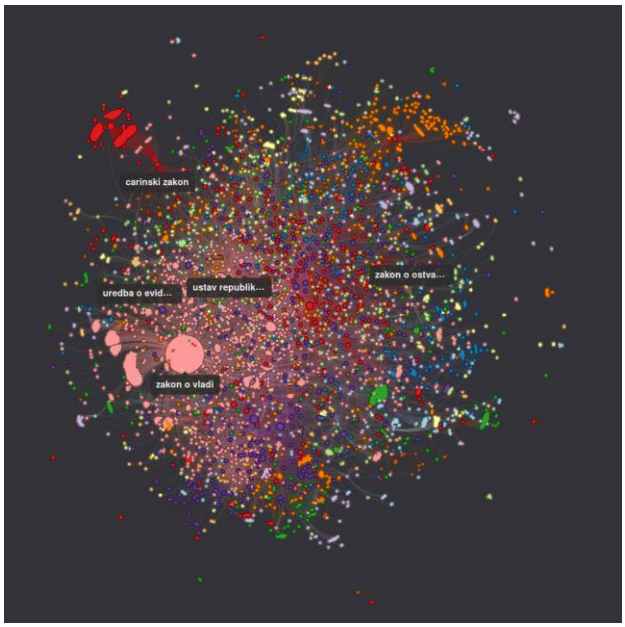

Figure 3. Serbian legislation network

V. CONCLUSION

In this paper, we collected valid legislation from Serbian Legal Information System. We used natural language processing techniques to extract full-explicit references and develop Serbian legislation network. Community detection algorithm and network science measures were applied and discussed.

Our future work includes the application of case studies for validating the usability and usefulness of the proposed legislation network.

REFERENCES

[1] A. Berger, "Die Erschliessung von Verweisungen bei der Gesetzesdokumentation", in *Informationssysteme*, vol. 3, 1971

[2] B. Waltl, J. Landthaler, F. Matthes, "Differentiation and Empirical Analysis of Reference Types in Legal Documents", *JURIX 2016 – 29th International Conference on Legal Knowledge and Information Systems*, pp 211-214, 2016

[3] R. Winkels, A. Boer, B. Verdebregt and A. van Someren, "Towards a Legal Recommender System", in *Frontiers in Artificial Intelligence*, vol. 271, pp 169-178, 2014

[4] P. Zhang and L. Koppaka, "Semantics-based legal citation network", in *Proceedings of the 11th international conference on Artificial intelligence and law*, Stanford, California, pp 123-130, 2007

[5] R. Boulet, P. Mazzega and D. Bourcier, "A Network Approach to the French System of Legal codes – Part I: Analysis of a Dense Network", *CoRR*, vol. abs/1201.1262, 2012

[6] T. Agnoloni and U. Pagallo, "The case law of the Italian constitutional court, its power laws, and the web of scholarly opinions", in *the 15th International Conference on Artificial Intelligence and Law (ICAIL)*, pp 151-155, 2015

[7] J. H. Fowler et al., "Network analysis and the law: Measuring the legal importance of precedents at the US Supreme Court", *Political Analysis 15.3*, pp 324-346, 2007

[8] N. Sakhaee, M. C. Wilson and G. Zakeri, "New Zealand Legislation Network", *JURIX 2016 – 29th International Conference on Legal Knowledge and Information Systems*, pp 199-202, 2016

[9] A.A. Hagberg, D.A. Schult and P.J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA USA, pp 11-15, 2008

[10] P. Pons and M. Latapy, "Computing communities in large networks using random walks", *ISCIS'05 – Proceedings of the 20th international conference on Computer and Information Sciences*, pp 284-293, 2005

[11] P. Bonacich, "Power and centrality: A family of measures", in *American journal of sociology*, pp 1170-1182, 1987

[12] L. Katz, "A new status index derived from sociometric analysis", *Psychometrika 18.1*, pp 39-43, 1953

[13] C. Ni, C. Sugimoto and J. Jiang, "Degree, Closeness, and Betweenness: Application of group centrality measurements to explore macro-disciplinary evolution diachronically", in *Proceedings of ISSI*, 2011

[4] http://graphistry.github.io/pygraphistry/index.html