Creating Context Networks in Dutch Legislation

Winkels, R.G.F.; Boer, A.W.F.; Plantevin, I.

Published in:
Frontiers in Artificial Intelligence and Applications

DOI:
10.3233/978-1-61499-359-9-155

Creating Context Networks in Dutch Legislation

Radboud WINKELS [1], Alexander BOER and Ivan PLANTEVIN
Leibniz Center for Law, University of Amsterdam, The Netherlands

AbstractThis paper describes ongoing research on automatically determining relevant context to display to a user of a legislative portal given the article they are retrieving, purely based on 'objective' criteria inferred from the network of sources of law. A first prototype is presented and a formative evaluation of it by legal expert users. Results are promising, but there is room for improvement.

Keywords. network analysis, legislative portal, MetaLex, references

1. Introduction

The official Dutch portal for national legislation (`www.wetten.nl`) allows users to search and browse all legislation as text with hyperlinks. When an article is in focus on that site, the structure (e.g. chapters and paragraphs) of the regulation is shown on the left hand side, but it does not clearly show the chapter you are in (figure 1). It is also not easy to switch to earlier versions of the same article or find out which other sources of law refer to the particular article in focus.

Figure 1. Interface on wetten.nl: On the left you can browse through all the chapters of the regulation and in the text of article 35 on the right there are 2 outgoing references visible as hyperlinks.

This paper describes ongoing research to improve the user interface on Dutch legislation by providing more and broader context to users for the article(s) they are investigating. We try to exploit the network structure that laws exhibit. Articles refer to other

[1]Corresponding Author: E-mail: winkels@uva.nl.

articles, in the same law or in other sources. This forms part of the network. Besides, earlier versions (or later, if one examines laws from the past) of the document in focus are related and some others might be through delegation of power.

Our research question is: given a particular document (article) in focus, can we determine other relevant documents purely on the basis of 'objective' meta-information? That is, we do not look at or interpret the content of the documents, nor do we use metadata added by other sources than the official owners and publishers of the documents.

Network analysis in a legal context has been done before, both on legislation (e.g. [5][1]) and on case law (e.g. [3][9]). The work on legislation however focused on analysing the entire network at the level of laws [5] or within one specific law [1]. We are looking for a small local network at a more detailed level (e.g. articles), given one starting point, possibly encompassing several laws or other sources.

In earlier research we experimented with a graphical representation of a contextual network ([8]), but one of the findings from a small evaluation was that (legal) users probably prefer a textual representation. In this paper we will present a prototype that resembles more the look-and-feel of the official Dutch portal. The idea is to provide the following types of information:

Internal General: A list of the most important texts (articles, but also chapters, etc.) in the law the text in focus is part of. Only texts that cite or are cited by other texts will appear in this list.

Internal Focus: A list of texts that belong to the same law as the consulted text and are citing the text in focus or are cited by it, ordered by importance.

External Focus: A list the texts that are citing the text in focus or that are cited by it, but belong to a different law than that of the consulted text, ordered by importance.

Versions Focus: A list of different versions of the text in focus.

Except for the first type, the lists change when the user changes his or her focus. The first type of information only changes when the user moves to another law.

We will first explain how we construct a context network from the legislation given a particular document in focus. Next we will present a small formative evaluation of the context provided by expert users and we end with some conclusions and suggestions for further work.

2. Creating a Context Network of Law

The "Metalex Document Server" (MDS) contains all regulations from the Dutch portal in CEN MetaLex XML[2] and as RDF linked data ([4]). This format is much more suitable for our purposes than that provided by the official portal. MDS currently contains more than 290 million RDF triples (August 2013), and this number is growing every day since every change to the `wetten.nl` site is added to the triplestore.

Regulations can be identified by their BWB-ID, e.g. the Foreigners law has ID "BWBR0011823"'. This ID can also be used to find a document at the `wetten.nl` site by entering it in the url; the entire law can be found at: `wetten.overheid.nl/`

[2] `http://metalex.eu/`

BWBR0011823/.[3] 'BWB' stands for "Basiswettenbestand" (basic law file), the content management system for all Dutch regulations that underlies the portal. An 'R' following 'BWB' indicates that the document is a regulation, a 'V' indicates a treaty ('verdrag'). The 7-digit number does not carry a specific meaning. The opaqueness of the BWB identifier is unfortunate, but hard to avoid, as the title of a regulation may change over time and cannot be used.

CEN MetaLex distinguishes the source of law as a published *work* from its set of *expressions* over time, and the expression from its various *manifestations*, and the various locatable *items* that exemplify these manifestations, as recommended by the IFLA Functional Requirements for Bibliographic Records (cf. [6]). References in legislation go from a specific version (expression) to the work level of the entity it refers to. Suppose article 35 of the foreigners law of September 1st 2010 refers to article 37 of that same law, then the reference leads from expression 2010-09-01 to the work art. 37 (see figure 2).

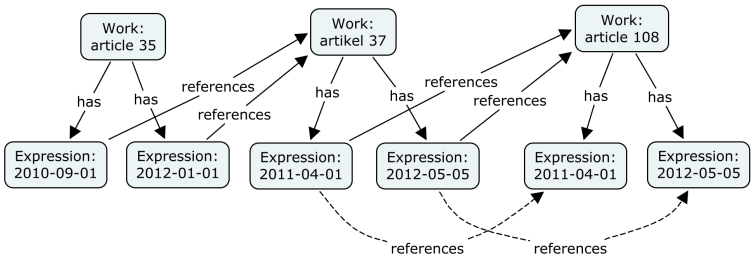

Figure 2. Expressions of works can reference other works. Sometimes new versions of citing and cited works are created at the same time, in which case the expression 'virtually' cites an expression, because the work it cites has only one valid expression during its life time. This is the case for the expressions of citing article 37 and cited article 108 in this example.

Since the current official legislative portal does not provide access to all versions of regulations, let alone at a level of granularity lower than entire regulations, we need some way of constructing a version history. To uniquely identify the *text* of an article, a hash code is generated and attached to it. The hash can be shared across bibliographic expressions with identical manifestations. These bibliographic expressions have an `owl:sameAs` relation with such a hash, which can be used to determine whether specific expressions are identical or not on the manifestation level (the XML contents). Note that articles and higher level parts of laws are not hashed and as a result, cannot be compared directly. At that level, different expressions can be considered to be different versions, even though they may actually have identical contents.

The construction of the citation network and its analysis are performed 'off line', as the process takes some time. For our present purposes, the network is constructed for only six laws from the tax domain.[4] These laws were selected by the tax authorities because they are closely related - together, they are important for the tax treatment of inheritances.

[3] It will display the current version of the "Vreemdelingenwet 2000", in Dutch.

[4] The same six laws we used for our research on extracting concepts and definitions as reported in [7].

The citation network is constructed in three steps. The desired citations are retrieved from the MDS through SPARQL queries first.[5] The XML files containing the results are then parsed to create identifiers for the nodes in the network and finally, the network is created by adding the nodes and the edges.

Only citations within the six laws are resolved and retrieved; citations for citing or cited entities that belong to other laws than the six used, are included in the network, but these entities are not further analysed. As a result, completeness of the network is guaranteed only for entities that belong to one of the six laws. The network is built by adding all entities as nodes and their relations as edges to the graph. If multiple identical citations exist – which is possible if different versions of an article cite the same entity, or if an article cites the same entity more than once – no additional edges or weights are added. The resulting network consists of 7,992 nodes and 13,496 edges.

During the creation of the network, three additional details are saved. The first is a list of the short titles for all laws that appear in the network and the BWB identifiers that belong to them. These titles are retrieved by querying the SPARQL endpoint during the parsing process. The second is a list of 'human' descriptions for the nodes in the network. The 'human' description for `BWBR0011353/artikel/4.14` for example, is `artikel 4.14`. The third is a list of work level URIs for all nodes in the network. For cited entities, these URIs are identical to the URIs listed in the SPARQL results. The work level URI for a reference is obtained by discarding the unwanted parts of the *part of* structure. The network and the additional saved data can be loaded fast enough to not hinder the user experience of the web application.

3. Analyzing the Citation Network

The citation network of the six laws is analyzed to yield one degree and two centrality measures for each node in the network. The degree measure is the *in degree*, which is the number of incoming citations for a given node. The first centrality measure is *degree centrality*, which is the fraction of nodes in the network a given node is connected to. The second centrality measure is the shortest-path *betweenness centrality*. The betweenness centrality of a node is the sum of the fraction of all-pairs shortest paths that pass through that node.

The three measures can be used to determine the relative importance of nodes in the network, each measure resulting in a different expression of importance with higher values representing greater importance. The in degree assigns a higher importance to entities that are cited more, while the degree centrality makes entities that are both citing and cited a lot, more important. Nodes with high betweenness centrality values often connect larger, otherwise unconnected groups to each other (cf. findings of [1]). They play a important role in linking other nodes together.

The prototype portal should allow users to switch between these importance measures easily so they can determine which one makes most sense to them (if any).

Of these three measures, betweenness centrality is the most time expensive calculation. While *in degree* and *degree centrality* are computed near instantaneously, computing *betweenness centrality* requires around 6 minutes.[6] If the Brandes algorithm is used,

[5] MDS provides a SPARQL endpoint at `http://doc.metalex.eu:8000/test/`.

[6] To calculate the betweenness centrality, the citation network is converted to an undirected graph first.

the required time increases with the number of nodes V and edges E linearly, with $O(VE)$ [2], since the citation network is unweighted. The calculation requires $O(V+E)$ space, 55MB for the citation network. The results of the calculations are saved and loading them is near instantaneous.

4. Time Travelling

As stated in the introduction, we also would like to offer users all other available versions of (parts of) laws when viewing a specific version. Retrieving all versions (expressions) for a given expression or work is straightforward. The triple store can be asked to return all entities that *realize* the given work:

```
PREFIX mo: <http://www.metalex.eu/schema/1.0#>
SELECT DISTINCT * WHERE {
    ?e mo:realizes <work>    }
```

Note: <work> is replaced by the work level URI of the given work.

To present a list of these versions, sorted by date, some additional steps are necessary, since the results cannot be sorted by date through the SPARQL query.

To find which versions (expressions) differ on the manifestation level, not only the expressions for the given work must be retrieved, but also their respective hashes (see Section 2). Next, the latest expression for each unique hash is kept. If there is just a single manifestation for all expressions of the same work, the version the user was viewing is kept. This comparison can only be done for parts of articles that contain text, since hashes are only created for such entities. From articles upwards in the structure of laws, no hashes are available. If we want to check whether the content of expressions of articles or higher level entities actually differs, the manifestations should be parsed and compared, which is difficult due to the metadata present in the XML files of the official portal.

5. The Prototype

The three parts necessary for our alternative portal described above – creating the network, analysing it and collecting versions – have been built in three Python modules. For the actual interface of the application a Django server application was built around these modules. The application has been designed for easy installation and maximum compatibility. The client side of the application is written in HTML5 and uses the jQuery JavaScript library. All recent versions of the major web browsers are supported, except Internet Explorer. The interface adapts its width and height to the available screen resolution and as a result fits on most devices' screens, from the iPad to desktop monitors.

The web application can be seen at `justinian.leibnizcenter.org/wetten`. A Dutch manual is accessible through a link in the interface. A screenshot is presented in Figure 3.

Frame 'A' consists of two parts. The top part contains a menubar, the bottom part is a button to show all available versions of the consulted text. Under *Wetten*, links to load each of the six laws for which the network was analyzed are present. Under *Opties*, the user can specify the number of texts to show in the windows in Frames 'D' & 'E' and

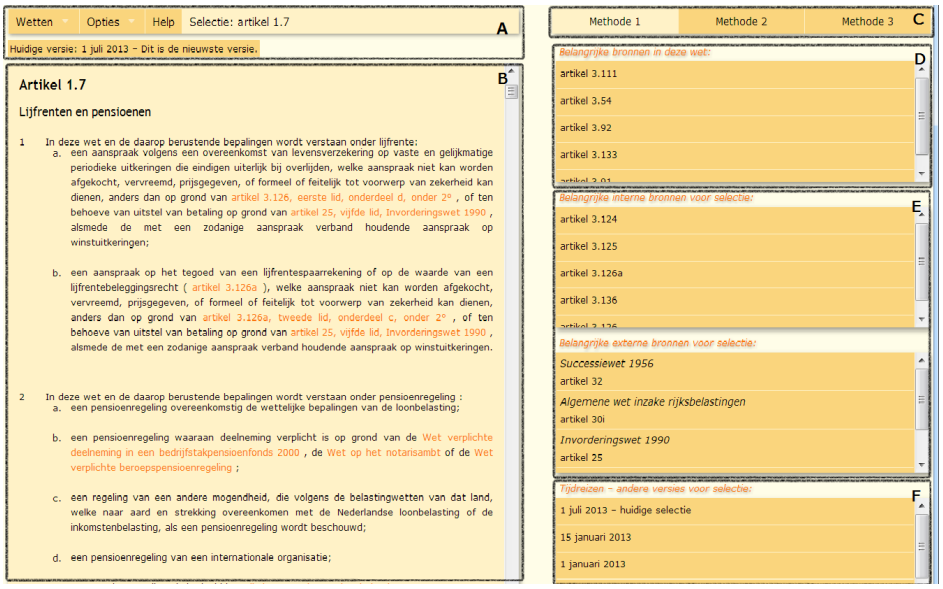

Figure 3. The prototype with art. 1.7 of the Income Tax Law in focus in frame 'B'.

whether to enable version history on the sub-article level. More on this later. *Help* is a link to the online manual for the interface. The rightmost part of the menubar shows the text currently in focus, which is *artikel 1.7* in Figure 3.

Frame 'D' gives a list of the most important texts in the current law. The desired importance measure can be set using the three buttons in Frame 'C'. The methods are in the same order as discussed above: in degree, degree centrality, betweenness centrality. When hovering over the buttons, the names of these methods are shown. The contents of the texts listed in 'Frame D' can be loaded by clicking on the list items and are displayed in a split-screen type of view below Frame 'B'.

When a text is in focus, as in Figure 3, Frame 'E' shows texts that cite the focus text or are cited by it, in the same law (upper part of the frame), or in other laws (lower part of the frame). Both lists are sorted by the same selected method as used for Frame 'D'.

Finally, Frame 'F' shows different versions of the text in focus. If this text is within an article, only versions that actually differ in content are shown. This is achieved by using the hashes as described above.

Two shortcomings of the prototype are worth mentioning here. The first one is that three types of structural parts of legal texts cannot be displayed after having been selected from the three lists in Frames D & E. These parts are of the types *afdeling*, *paragraaf* and *titeldeel*. They cannot be brought into focus either. This is due to the fact that these parts do not have expressions associated with their works in the RDF triple store. This, in turn, is due to the way these elements are represented in the `wetten.nl` XML data. The problem could be solved by modifying the conversion process of the Metalex Document Server. Currently, the user is notified of this problem when trying to load one of these parts. If the user wants to focus on such an element, the parent chapter is used as focused text instead.

The second shortcoming is that the ordering of entities in Frames D & E is only based on the network of the six laws analyzed for this project. Only by building and ana-

lyzing the complete citation network for all laws could the completeness and ordening of these lists be guaranteed. The time travel (version history) functionality however, works for all laws as it is independent of the citation network.

5.1. Formative Evaluation

The interface has been tested by three employees of the Dutch tax authorities (*Belastingdienst*), who have provided feedback.

The employees were asked to try out the various functions of the application for at least the *Wet inkomstenbelasting 2001* and to subsequently answer questions on relevance and completeness of the three lists resulting from the network analyses, and questions on time traveling and general user experience. Their findings are summarized below:

List of important texts in consulted law: the items listed were found to be important, but their utility depended on the specific law and user: important texts might not be particularly useful for persons who simply never have or need to consult them. Texts from other sources (the parliament, for instance) would have been desirable, but those are not part of the citation network of the Dutch legislation on the official portal and the Metalex Document Server.

Important texts for focus: the two lists were found quite useful, as all cited and citing texts could be seen at a glance. However, one user noticed that when traveling back in time, irrelevant texts appeared. This is due to the fact that texts citing or cited by later versions of the focused text are shown, since only a single network has been built for all versions of the text. To solve this issue, ideally multiple networks for various points in time should be built (and analyzed as well). A less perfect solution would be to filter the results by time. This way all future texts could be filtered out, but the importance measures will still be based on the entire network. Currently the dates are not part of the network, so this would require a small change of the system.

Sorting methods: users did not have a preference for one of the three different sorting methods. This could be due to a high level of similarity between the results of the methods. This similarity will be discussed below. One user expressed a slight preference for *betweenness centrality*.

Time traveling: the possibility to explore the various versions of the legal sources was found helpful and fast. However, there were two shortcomings mentioned by the evaluators. The first one is the issue mentioned above where documents are shown that are only relevant to other versions of the text when time traveling. Different networks or filtering by date should be used to guarantee no irrelevant texts are presented. The second shortcoming is the limited possibility to consult future versions of legal texts. The Metalex Document server stores only versions that have already appeared on `wetten.nl`, but the `wetten.nl` portal lists for example the planned expiration dates of regulations. This information would be useful to show in the time traveling section of the interface. However, that information is not stored on the MDS and can therefore not directly be implemented in the interface. One user also mentioned that sometimes regulators change the entire content and purpose of an article, while its number remains the same. It can therefore be confusing to show all versions, if some of them can actually be considered as entirely

different works. For this reason, this method should be avoided by legal drafters, but it does unfortunately happen.[7]

Overall User Experience: the application was found to be fast, especially compared to the `wetten.nl` portal. The fact it worked fine on the iPad was an advantage as well. There were some drawbacks with the amount of information shown, the help function and the applied terminology. The users thought the application showed too much information. They would have liked an option to hide some or all of the windows displayed in the right half of the interface, since they did not require all information at all times. One user found the help function unpractical due to the manner in which it is presented: as a separate webpage. More compact popups for example, shown for individual parts of the interface without opening a new window, could be more user friendly. The terminology used in the interface was not in all cases fitting, one user thought. Additionally, one user missed the table of contents showing the structure of the law, as available on `wetten.nl`.

5.2. Similarities Between Measures of Importance

To test the hypothesis that the similarity between the results of the three methods used to assign importance to entities in the citation network could be the reason the choice between methods was found to be of lesser importance than expected, they were compared in a pairwise manner. For this comparison, two values were calculated: the overlap between results and the fraction of identical rankings. To calculate the overlap, the results for two different methods are taken and those that appear in both methods are counted and divided by the total number of results. In other words, the intersection of the two lists is taken, after which the fraction of results occurring in both lists is calculated. The second computed value, the fraction of identical rankings, consists of dividing the number of results that share the same position in two lists, divided by the total number of results.

The overlap and the fraction of identical rankings were calculated for seven lists. The first list was the top 5 of most important texts in the *Wet inkomstenbelasting 2001*. The six other lists were the lists of top 5 internal and external important texts for the three texts listed as most important in that law. These articles were article 3.111, 10.1 and 2.5.

The means of the two values for all seven lists are listed in Figure 4. From these results it becomes apparent that not only *in degree* and *degree centrality* are relatively similar – as would be expected since they both look at the number of references –, but also *degree centrality* and *betweenness centrality* have a high overlap and number of identically ranked texts. In contrast, the difference between *in degree* and *betweenness centrality* is greater. These results suggest that the relatively high level of similarity between the measures of importance could indeed explain the user's experiences. It is suspected that this similarity is due to the nature of the citation network. The network has been built for only six laws that are closely related to each other. As a result, the network would have a group-like structure, where much cited and citing texts often lie on many shortest paths, explaining the overlap between the methods.

[7] E.g. art. 8.12 of the IB2001 concerned 'children discount', 'work-on bonus' and 'work bonus' respectively through the years.

Comparison between:	Overlap	Identical positions
In degree – degree centrality	0.60 (s = 0.23)	0.23 (s = 0.18)
In degree – betweenness centrality	0.40 (s = 0.26)	0.14 (s = 0.15)
Degree centrality – betweenness centrality	0.66 (s = 0.25)	0.34 (s = 0.32)

Figure 4. Overlap of texts between the three methods, in fractions, and fraction of identically positioned texts.

6. Conclusions and Future Work

The question we set out with was whether it would be possible to determine which other legal documents to show a user of a legislative portal that has a particular document (article) in focus, purely on the basis of 'objective' network information. To answer this question, the citation network for six closely related laws was built and analyzed and a web application was created using the acquired information, while also incorporating version history functionality. The resulting interface presents information in a manner `wetten.nl` does not and to determine its usefulness, the interface was evaluated by three professional users. Based on their findings we can say the information is indeed useful, but there is room for improvement.

The information users found most useful were the lists of internal and external documents citing the article in focus or being cited by it, and the time travel functionality. The speed and general usability was found to be good as well. The users did miss functionality `wetten.nl` provides, such as a table of contents and a search function, which is not available in the interface. Implementing this functionality fell outside the scope of this study.

Regarding the three different sorting methods, we expected *in degree* to be the most appropriate option. The users however, did not prefer one over another. This could be due to the similarity between the results of these three methods for this network, therefore other measures of importance that (potentially) go beyond network analyses might be necessary. For other (larger and less coherent) networks, users may prefer one method over others. It should be stressed that notions of importance could vary a lot from one (type of) user to another.

The interface created for this project is a prototype for further research into the use of CEN MetaLex and RDF linked data. The application as a whole is highly extensible and the use of MetaLex means that it can easily be adapted for other sources of law, as long as the MetaLex standard is used.

6.1. Future Work

- Alternative types of networks could be used. The present network is unweighted, but weights assigned to the citations could be used when determining the importance of texts. For example, the number of times a text cites another text could be used as weight for an edge (cf. [8]).
- Only a single network was built for this project, which resulted in irrelevant texts appearing in the lists of citing and cited texts when time traveling. Creating and analyzing a network on a regular – e.g. monthly – basis would eliminate these irregularities.
- The network was built for just six laws, but it would be interesting to investigate whether similar results are obtained with a much completer network.

- Other measures of importance could be investigated. Alternative centrality measures could be used, but the importance of texts could also be determined by looking into their contents or using other properties not originating from the citation network. It would also be interesting to look into the possibility of defining user-dependable measures of importance, where importance is also determined by the preferences of specific users. Machine learning could be applied as well, where the application would learn from the choices of the user.
- Version history and time traveling could be further explored. The actual differences between versions could be determined and shown to user. It would also be useful to show known information about future versions of texts (e.g. expiration dates).
- A higher number of more detailed evaluations by test users would be desirable. Moreover, other methods of evaluation could be used in addition to user tests. For example, the articles considered to be the most important ones according to (law) textbooks could be compared to the results returned by the application (a method we used for evaluating importance of case law based on network analysis, [9].
- Finally, integrating texts from other sources than Dutch legislation, such as case law and legal doctrine, would be interesting, as those documents also refer to legislation. This information however, is not available on the Metalex Document Server or on the official portal. Also, the addition of functionality already available on `wetten.nl`, such as a table of contents and a search functionality should be considered.

Acknowledgments

We would like to thanks the people from the Tax Administration for evaluating the prototype.

References

[1] Romain Boulet, Pierre Mazzega, and Danièle Bourcier. *Network Analysis of the French Environmental Code*, volume 6237 of *Lecture Notes in Computer Science*, pages 39–53. Springer, 2009.

[2] Ulrik Brandes. A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[3] James H. Fowler, Timothy R. Johnson, Sangick Jeon, and Paul J. Wahlbeck. Network analysis and the law: Measuring the legal importance of supreme court precedents. *Political Analysis*, 15(3):324–346, 2006.

[4] Rinke Hoekstra. *The MetaLex Document Server - Legal Documents as Versioned Linked Data*, pages 128–143. Springer, 2011.

[5] P. Mazzega, D. Bourcier, and R. Boulet. The network of french legal codes. In *ICAIL 2009*, pages 236–237, 2009.

[6] K. G. Saur. Functional requirements for bibliographic records. *UBCIM Publications - IFLA Section on Cataloguing*, 19:136, 1998.

[7] R. Winkels and R. Hoekstra. Automatic extraction of legal concepts and definitions. In *JURIX 2012*, pages 156–165. IOS Press, 2012.

[8] Radboud Winkels and Alexander Boer. Finding and visualizing context in dutch legislation. In *Proceedings of NAiL 2013*, 2013.

[9] Radboud Winkels, Jelle de Ruyter, and Henryk Kroese. Determining authority of dutch case law. In Katie Atkinson, editor, *JURIX*, volume 235 of *Frontiers in Artificial Intelligence and Applications*, pages 103–112. IOS Press, 2011.