Two Experiments in Finding Relevant Case Law

Radboud Winkels[(a)] & Alexander Boer
[(a)] Leibniz Center for Law, University of Amsterdam, Netherlands winkels@uva.nl

Abstract. Within the OpenLaws.eu project, we attempt to suggest relevant new sources of law to users of legal portals based on the documents they are focusing on at a certain moment in time, or those they have selected. In the future we attempt to do this both based on 'objective' features of the documents themselves and on 'subjective' information gathered from other users ('crowd sourcing'). At this moment we concentrate on the first method. In earlier work we have described results of experiments using analysis of the network of references or citations to suggest these new documents. Now we describe two experiments where we mix the use of network analysis with similarity based on the comparison of the actual text of documents. One experiment is based on simple bag-of-words and normalisation, the other uses Latent Dirichlet Allocation (LDA) with added n-grams. A small formative evaluation in both experiments suggests that text similarity alone works better than network analysis alone or a combination, at least for Dutch court decisions.

1. Introduction

More and more sources of law are (freely) available online in Europe and the rest of the world. It concerns both legislation and case law and possibly others like legal commentaries. Professionals may already get lost in the multitude of information, let alone ordinary citizens and (small and medium sized) enterprises. Traditionally commercial publishers provide this information, plus support for access. Legal experts write commentaries for them, editors provide links between different (types of) sources and warn subscribers for interesting new developments and case law. Now that large amount of sources of law become electronically available online, the question is whether new ways for supporting access can be developed. One stream of research may be directed towards using the 'wisdom of the crowd', have users of legal information share their collections of material, the links they see between different sources, their commentaries, etc. Another stream of research is directed at (semi-) automated linking and clustering of sources of law, analysis of the network of law to find authoritative sources or predict the change of opinion of higher courts, etc. In the OpenLaws.eu project we explore both approaches (Wass e.a., 2013).

We are developing a platform that enables users to find legal information more easily, organize it the way they want and share it with others. Part of the OpenLaws.eu service is the attempt to suggest relevant new sources of law to users of legal portals based on the documents they are focusing on at a certain moment in time, or those they have selected. In earlier work we have described results of experiments using analysis of the network of references or citations to suggest these new documents (Winkels e.a., 2013, 2014). Now we describe two experiments where we mix the use of network analysis with similarity based on the comparison of the actual text of documents.

In this paper we will first shortly explain how we suggest possibly interesting documents based on network analysis of legislation and case law. Next, we describe the two experiments in which we compute the similarity of cases and use this to suggest other case law to a user interested in a particular case. Both experiments were evaluated by users and we end with conclusions and future work.

2. Recommending Documents Based on Network Analysis

Most of the sources of law available online are stand-alone web services or databases, containing one type of documents, not linked to other sources. For instance the Dutch portal for case law – rechtspraak.nl – contains a (small) part of all judicial decisions in the Netherlands. Case citations in these decisions are sometimes explicitly linked, references to legislation are not.[1] From earlier research we know that professional users of legal documents would like to see and have easy access to related ones from other collections. E.g. when we evaluated a prototype system that recommends other relevant articles and laws to users of the official Dutch legislative portal, they told us they would like to see relevant case law and parliamentary information as well (Winkels e.a. 2013).

Another problem of existing portals and data bases is that not even all internal links are explicitly represented. This is especially true for so called relative links like "the previous article", or "the second sentence of article *x*" and incomplete ones ("that law" or "article *y*" without the law that it is part of).

If these links are not given, we can try to find them automatically. For inter-legislation links we have shown this can be done very effectively for the Dutch case (de Maat e.a, 2006) and others for other jurisdictions (e.g Palmirani e.a., 2003 for Italy; Tran e.a., 2013 for Japan). For inter-case law links it is a bit more difficult, but we have shown it works for the Dutch case (Winkels e.a., 2011 and so did Van Opijnen, 2014). That leaves finding citations in case law to legislation and possibly finding links in other sources of law like commentaries to both case law and legislation. In this paper we will focus on finding references to legislation in Dutch case law and how we can use these to improve access to Dutch sources of law.

2.1 Creating a Network from Dutch Case Law

The Dutch portal for case law contains a small, but growing part of all judicial decisions in the Netherlands. Case citations in these decisions are sometimes explicitly marked in metadata (e.g. the first instance case); references to legislation only the main one(s) in recent cases. The texts are available in an XML format, basically divided in paragraphs, with a few metadata elements. The most relevant metadata for our purpose are:

- The date of the decision ('Datum uitspraak')
- The field(s) of law ('Rechtsgebieden')
- The court ('Instantie').

The court decisions do not contain inline, explicit, machine readable links to cited legislation or other cases. So even when the metadata contain such references, we do not know in which

[1] Recently one has started to add some links to cited legislation in metadata.

paragraph the case or article was cited, nor how often. We resort to parsing techniques to make these citations explicit and count them. In Winkels e.a. (2014) we described doing that for a subset of all case law belonging to `immigration law‘ and that contained the actual text of the verdict (13,311 documents at the time)[2]. For locating references to legislation we use regular expressions together with a list of names and abbreviations of Dutch laws. This list also contains the official identifier of the law (the BWB-number), which can be used for resolving the reference later on. We consider high precision to be more important than high recall. Users will forgive us if we miss a reference, but be annoyed by false ones. We evaluated this procedure by checking 25 randomly selected documents by hand. These documents contained 163 references to legislation of which 141 were correctly identified (recall of 87%). There was one false positive (precision of 99%).

Resolving the references was a bit trickier, since sometimes they used anaphora, e.g. referring to ‘that law’. In such cases, the citation was resolved by using the previous law identifier if it existed, i.e. we assume that the complete law was introduced just before in the text and resolved correctly. We used the same process for resolving ambiguous title abbreviations; e.g. `WAV` is an abbreviation of ‘Wet Arbeid Vreemdelingen’, ‘Wet Ammoniak en Veehouderij’ and ‘Wet Ambulancevervoer’. Most of the time the full title is used before the abbreviation is used. We evaluated the process by checking 250 random ones by hand, giving a recall of 85% and a precision of 95%.

The final network of the 13,311 case documents has 85,639 links to legislation (on average 6.5 references per case); the links connect the ECLI identifier[3] of the case with the BWB identifier of the (part of) law it refers to. Since case decisions may refer to the same source of law, e.g. an article, more than once, we count the number of references and compute the weight of the link between the case and the article as: $W = {}^{1}/_{n}$ where n is the amount of occurrences of a certain reference and W is the weight of the edge. The lower the weight, the stronger the impact on the network is. In earlier work we used this network to suggest possibly relevant documents based on the current one a user was inspecting. When a user selects an article of Dutch law, we check whether this appears in the network. If so, we select the most ‘relevant’ nodes in its surroundings as suggestions. Relevancy of nodes in a network can be computed in different ways, for instance based on the number of incoming references, the number of outgoing, the relevance of the incoming references, etc. We have experimented with several ones, including ‘PageRank’ and ‘betweenness centrality’ (Winkels e.a., 2013, 2014).

3. Adding Other Features of Documents

In previous experiments we only exploited the web of citations between documents. Now we will explore whether we can improve suggestions for relevant documents by including other features, notably similarity measures based on the comparison of the actual text of documents. As stated in the introduction, we focus on case law in this study.

[2] Some cases only contain metadata and not the actual text of the judgment.

[3] ‘European Case Law Identifier’; see Council conclusions on ECLI at: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52011XG0429(01)

Relevancy for case law is hard to define; it is subjective, depends on the task or problem of the user searching for case law and also on the type of user. For a student or novice, well known landmark cases might be very relevant, while for a legal expert these are probably not. He or she will be more interested in less known or very recent new cases. In previous research we have concentrated on legal expert users and we still are, but we will also have a look at the preferences of novices.

3.1 Reference Similarity Combined with Text Similarity

The first experiment concerns case law within the Dutch tax domain, about 6,000 documents. They were again taken from the official Dutch portal based on the metadata 'Field(s) of law' (see above). After some pre-processing in which XML tags and strange characters were removed, we used the same parser as described earlier to detect references to legislation. A small test revealed that this time overall recall of the parser was only 55%. Main reason is missing names and abbreviations of tax laws and the use of different terms in the tax domain (like 'protocol') compared to other domains of law. The effect of these omissions is multiplied if a case refers to the same item more than once.

Our hypothesis is that 'similar' cases can be identified by similarity in the legislation they cite and by similarity of the words used in the judgements. As a baseline we use the similarity in words used (text similarity): Bag-of-words combined with normalised TF/IDF weighting and cosine similarity.[4] TF/IDF corrects for frequency of terms and normalisation for the length of documents; otherwise documents with highly frequent words and long documents would be overrated by the cosine similarity measure. For any document in our set (the focus document) we can now select the *n*-most similar other documents from that set. The left part of Table 1 gives an example for the 10 most similar documents found this way for a verdict of the court of Amsterdam of 2010 (BO1378[5]) on the value of a house. The most similar case is one of the court of Alkmaar of 2008 (BC6103[6]), also on the value of a house.

Table 1: 10 Most similar documents to RBAMS-2010-BO1378 according to bag-of-words (left) or bag-of-references (right)

	ECLI-NL-RBAMS-2010-BO1378			
	Bag of Words		**Bag of References**	
1	RBALK-2008-BC6103	0.72	RBSGR-2008-BD1495	0.89
2	RBDOR-2010-BM0117	0.69	RBUTR-2010-BU4490	0.73
3	RBALK-2011-BQ0469	0.64	RBOVE-2014-951	0.69
4	RBARN-2006-AY9465	0.64	RBALK-2008-BD7537	0.69
5	RBDOR-2010-BO5257	0.62	RBALK-2007-BB9105	0.69
6	RBAMS-2011-BV6758	0.61	RBAMS-2011-BQ424	0.65
7	RBDOR-2010-BM2339	0.61	RBALK-2008-BC4175	0.64
8	GHAMS-2013-CA2684	0.61	RBALK-2012-BX0044	0.59
9	RBAMS-2011-BR6478	0.60	RBALK-2008-BD5937	0.58
10	RBHAA-2006-AZ2187	0.59	GHAMS-2001-AD8208	0.57

[4] TfidfVectorizer of SciKit-learn (Pedregosa e.a. 2011) was used with minimal document frequency set to 1 and maximum to .7.

[5] http://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBAMS:2010:BO1378

[6] http://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBALK:2008:BC6103

The same algorithms are used to calculate the reference structure similarity between two documents.[7] The right part of Table 1 gives the 10 most similar documents based on that method. Here number 1 is a verdict of the court of The Hague of 2008 (BD1495[8]) on pollution tax for commercial property.

The final step combines the bag-of-words similarity score and the bag-of-references similarity score by taking the average of the two scores. Now we can determine the *n* most similar documents for any *focus* document. We chose to only take into account the focus documents that have at least four outgoing references for now, this because of the relatively low recall of the parser. If a focus document has only one outgoing reference, the bag-of-references similarity scores will be either 0.0 or 1.0; this may be acceptable in a later stage with a superior reference parser, but at this stage these extreme scores are too uncertain.

Formative Evaluation

A small group of experts was asked to evaluate the system. They were asked to first read a focus document, randomly selected from a prepared database[9] and shown on an evaluation website that was created for this purpose. Subsequently they were asked to read six recommended documents and rank them on relevancy to the focus document. The most relevant one is ranked first (1), the least relevant last (6). The six documents were three with the highest similarity scores for the baseline implementation (bag-of-words only) and three documents with the highest similarity scores for the bag-of-words combined with the bag-of-references. They were also asked to give an overall score for the relevance of a suggestion on a scale from 1-10 with 1 representing 'not relevant' and 10 'very relevant'. This was done to assess the overall quality of suggestions. Even very bad suggestions may be ranked after all. Figure 1 gives an example screen from the evaluation site. Our intuition was that it is easier to rank suggestions for relevance than to assess the overall relevance of a suggestion.

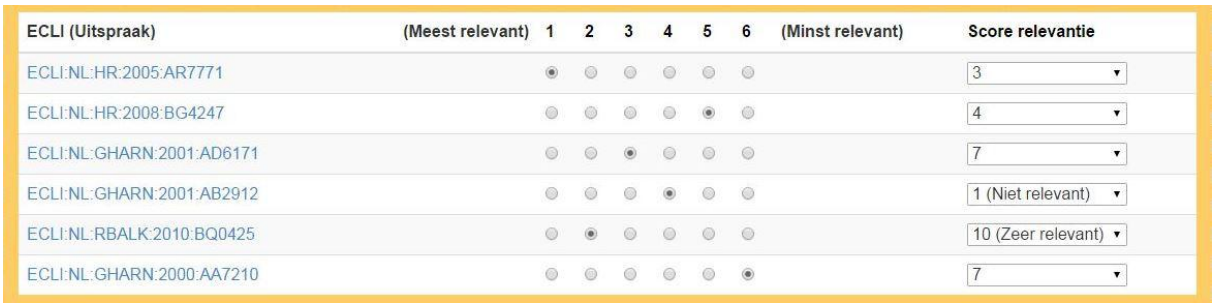

Figure 1: Screenshot of the evaluation site (in Dutch)

This method of evaluation is rather subjective if a small amount of test subjects is used, but if can serve to determine if the research is heading in the right direction and find possible bugs and errors (Shani and Gunawardana, 2011).

[7] No maximum document frequency this time.

[8] http://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBSGR:2008:BD1495

[9] Because we are interested in whether adding the bag-of-references to the similarity scores improves performance, documents that had recommendations that occurred in both sets were removed.

Four experts evaluated 18 cases; 15 unique ones and 3 the same for all experts. Results are presented in Table 2. It is clear that adding the bag-of-references worsens performance. This may partly be due to the fact that only 55% of references were found.

Table 2: Results expert evaluation

	Average rank	*Average score*
Bag-of-words only (baseline)	3.06	5.58
Bag-of-words with bag-of-references	3.94	4.46

It may also be that similarity in references reflects something else than similarity in text. To further investigate this aspect, we examined to what extend the bag-of-words approach suggests the same documents as the bag-of-references approach. We analysed the top ten recommendations for 1,000 focus documents for both approaches. We ignored recommendations with a similarity score of zero. Of the remaining 9,528 recommendations, 1,135 were recommended by both algorithms (about 12%).

Again, the amount of overlap may turn out to be higher with a better performing parser. Another explanation is that similarity in references reflects a more abstract commonality than our evaluators could spot or found useful.

3.2 Network Analysis Combined with Topic Modelling

A more advanced approach of comparing similarity of texts than the bag-of-words approach discussed above is that of *topic modelling*. A topic model represents a document, a court judgement in this case, as a mixture of topics. A topic is a set of words or phrases. Perhaps the most common topic model currently in use is Latent Dirichlet Allocation (LDA) from Blei et al. (2003). One downside of this approach is that it treats documents as bags of words, i.e. ignores word order and therefor word phrases like 'European Union' or 'our Minister'. Several extensions of LDA have been proposed, one of which is Turbo Topics (Blei and Lafferty, 2009) which starts like LDA, but subsequently significant words that are preceding or succeeding topic-words are searched in the texts and added to the topic model. In this experiment we use the open source implementation of Turbo Topics in MALLET.[10]

To be able to compare the results with suggestions purely based on network analysis, we decided to use the same domain as in the earlier work described above, namely immigration law (Winkels e.a., 2014). After some pre-processing and selection of those cases with actual content, we worked with a set of nearly 13,500 cases.

After all cases were represented as mixtures of topics, we selected the *n* best suggestions for each case by calculating the similarity between the topic mixtures. We calculate the sum of squared errors between a specified case and all other cases and convert this to a similarity value ranging from 0% to 100% overlap. Table 3 gives an example for three suggestions for case ECLI:NL:RBSGR:2009:BH7787 (a case of 2009 of the court of The Hague about a refugee from Turkey who was a courier for PKK) with similarity measures. It obviously has a

[10] MAchine Learning for LanguagE Toolkit, McCallum (2002).

similarity measure of 100% with itself. Next comes a case from the same court of 2004 with a similarity in topics of more than 99% (also about a member of the PKK), etc.

Table 3: Example output of 3 suggestions for case ECLI:NL:RBSGR:2009:BH7787

Case	Similarity measure
ECLI:NL:RBSGR:2009:BH7787	100.00
ECLI:NL:RBSGR:2004:AQ5970	99.28
ECLI:NL:RBDHA:2015:4915	99.01
ECLI:NL:RVS:2004:AQ5615	98.67

Formative Evaluation

For this project, the evaluators consisted of three novices, two legal experts with experience in the immigration law, and two legal experts without experience in this field. For the results the four legal experts were pooled together, since there was no difference in their evaluations. The evaluators were given five randomly selected cases for which they ranked three suggestions from best to worst, and for which they stated whether any of the three suggestions is good enough for a recommender system. One of the three suggestions was most similar according to the method using a topic model, another was the most relevant based on the references to legislation according to Winkels et al. (2014) and one of the suggestions was based on a combination of the two methods. For the combination of the two methods, a list of the top 200 suggestions according to the topic model was obtained (for each randomly selected case), and also a list of the top ten suggestions according to the references to legislation. The first suggestion from the top ten list that appeared in the other list of the top 200 suggestions, was chosen as the suggestion based on both methods. The evaluators were unaware of which suggestion was obtained by which method.

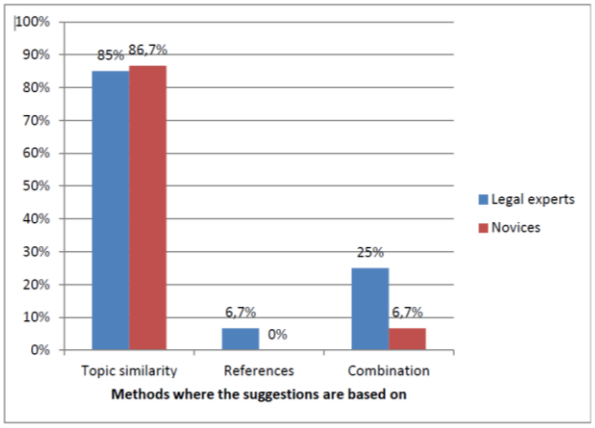

Figure 2: Percentage of evaluators that thought a suggestion was good enough for recommendation

Both legal experts and novices showed significant preference for the suggestions based on topic similarity. The legal experts wanted to see 85% of the suggestions based on topic similarity in a recommender system, for the novices this was 87% (see Figure 2). The legal experts ranked the suggestions based on topic similarity as best suggestion 80% of the time,

while the novices always ranked the suggestion based on topic similarity as best suggestion. This indicates that suggestions based on topic similarity can give useful suggestions within Dutch case law.

4. Conclusions

We described results of research that is part of the OpenLaws.eu project. Ultimate aim is to deliver a platform for using, sharing and enriching big open legal data. Besides offering users the opportunity to collect, organize and annotate legal data, we also want to offer automatic suggestions based on analysis of existing data. Based on analysis of the network of both legislation and case law, we offer users suggestions for interesting material based on their current focus document.

We have shown that it works quite well to automatically find and resolve references to legislation in Dutch case law, at least in the immigration domain. The parser was perhaps a bit over fitted for that domain, since it performed less well in the tax domain. It can easily be improved, but we will have to check whether this has repercussions for the immigration domain and how it performs in other legal fields.

The network of references can be used to provide users of the legislative portal with relevant judicial decisions given their current focus and moreover, suggest additional relevant legislative sources.

When compared with suggestions of case law based on similarity of the actual text of the judgements, whether seen as just ‘bags of words’ or as a mixture of ‘topics’, users seem to prefer those over the suggestions based on network analysis or similarity in reference structures. We still have the intuition that similarity in reference structure indicates some common feature of cases, but perhaps it is too abstract for users. In our first experiment we treated the references as an unordered set; perhaps we should retain the order and try again.

In the future we can also use other features like:

- The hierarchical position of the law cited, e.g. whether the referred law is a European directive or treaty, or a governmental decree;
- Document structure level of the reference. A lower document structure level (e.g. article or clause instead of a chapter) suggests a more specific reference, which could indicate a different role;
- The date of a case, preferring more recent case for expert users.

Finally, it would be very useful for future research if we had a labelled training set that we could use for testing different approaches. Evaluation by users is a time consuming affair. Experts needed an average of 12 minutes to read and score one example in our small evaluations.

Acknowledgements. Part of this research is co-funded by the Civil Justice Programme of the European Union in the OpenLaws.eu project under grant JUST/2013/JCIV/AG/4562. We would like to thank our students: Erwin van den Berg and Wolf Vos who performed the experiments.

5. References

Blei, D.M., Ng, A.Y., Jordan, M.I., Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 2003, pp. 993-1022.

David Blei and John D. Lafferty. Visualizing topics with multi-word expressions. *arXiv, 0907*, 2009.

Maat, E. de, Winkels, R., and Engers, T. van (2006). Automated detection of reference structures in law. In T. van Engers (ed), *JURIX 2006*, pp. 41-50. IOS Press, Amsterdam.

McCallum, A.K. (2002). *Mallet: A machine learning for language toolkit*. http://mallet.cs.umass.edu

Newman, M. (2010). *Networks: An Introduction*. Oxford, England: Oxford University Press.

Opijnen, M. van (2014). *Op en in het web. Hoe de toegankelijkheid van rechterlijke uitspraken kan worden verbeterd*. PhD Thesis in Dutch, University of Amsterdam.

Palmirani, M., Brighi, R. and Massini, M. (2003). Automated extraction of normative references in legal texts. In: *9th International Conference on AI and Law*, pp. 105–106, ACM, New York.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*:2825-2830.

Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In: *Recommender systems handbook*, pp. 257-297. Springer, Berlin.

Tran, O.T., Le Nguyen, M. and Shimazu, A. (2013). Reference resolution in legal texts. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pp. 101-110, ACM, New York.

Wass, C., Dini, P., Eiser, T., Heistracher, Th., Lampoltshammer, Th., Marcon, G., Sageder, C., Tsiavos, P. and Winkels, R. (2013). OpenLaws.eu. In: *Proceedings of the 16th International Legal Informatics Symposium IRIS 2013*, Salzburg, Austria (2013).

Winkels, R.G.F., de Ruyter, J. & Kroese, H. (2011). Determining Authority of Dutch Case Law. In K. Atkinson (ed). *JURIX 2011*, pp. 103-112. IOS Press, Amsterdam.

Winkels, R.G.F., Boer, A. and Plantevin, I. (2013). Creating Context Networks in Dutch Legislation. In K. Ashley (ed*). JURIX 2013*, pp. 155-164. IOS Press, Amsterdam.

Winkels, R.G.F., Boer, A., Vredebregt, B. & Someren, A. van (2014). Towards a Legal Recommender System. In R. Hoekstra (ed). *JURIX 2014*. Volume 271 of Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, pp. 169-178. Best paper award.