Making a Cold Start in Legal Recommendation: an Experiment

Boer, A.; Winkels, R.

DOI
10.3233/978-1-61499-726-9-131
Publication date
2016
Document Version
Author accepted manuscript
Published in
Legal Knowledge and Information Systems

Making a Cold Start in Legal Recommendation: an Experiment

Alexander Boer and Radboud Winkels
Leibniz Center for Law, University of Amsterdam, The Netherlands
e-mail: A.W.F.Boer@uva.nl

Abstract.
Since the OpenLaws portal is envisioned as an open environment for collaboration between legal professionals, recommendation will eventually become a collaborative filtering problem. This paper addresses the *cold start* problem for such a portal, where initial recommendations will have to be given, while collaborative filtering data is initially too sparse to produce recommendations. We implemented a hybrid recommendation approach, starting with a *latent dirichlet allocation* topic model, and progressing to collaborative filtering, and critically evaluated it. Main conclusion is that giving recommendations, even bad ones, will influence user selections.

1. Introduction

The OpenLaws portal is envisioned as an open environment for collaboration between legal professionals enriching and classifying sources of law. Recommendation by the system is therefore mainly seen as a collaborative filtering problem, sourcing the wisdom of the professional legal crowd. Since the portal is intended to cover (at least) the primary sources of law (codified law, regulations of general impact, and published court decisions) of the members of the European Union, gaining sufficient traction in each language area and jurisdiction to find a sufficiently sizable legal crowd to sort documents into folders by topic.

We implemented a hybrid recommendation approach for the portal, starting with a *latent dirichlet allocation* topic model for each corpus included in the portal, and progressing to collaborative filtering, leveraging legal professional knowledge. This paper addresses the *cold start* problem for such a portal, where initial recommendations will have to be given on corpuses that lack useful classification metadata, in languages most of us do not speak, while collaborative filtering data is initially too sparse to produce recommendations.

The technique implemented for initial recommendations has been evaluated for expert user acceptance, and has been comparatively evaluated relative to alternative solutions, in [7]. The results of those evaluations were promising, and gave us no strong reasons to worry about the quality of the recommendations produced in the expert's eyes. Comtemplating the *cold start* of the portal leads to new critical questions, however:

To what degree will the rather arbitrary initial topic model produced for a corpus influence the choices made by the users of the system? Are professional users able to resist the temptation of simply following recommendations, even if they are bad ones?

We critically evaluated a mock up cold start situation in an experiment with users, each consecutively using the same recommender system, as it acquires more information for collaborative filtering. To show the framing effect of the topic model's recommendations effects, we intentionally manipulated recommendations. Main conclusion of this single run is that giving recommendations, even bad ones, does influence user selections. The cold start problem in OpenLaws legal recommendation requires more, and methodologically more rigorous, experiment to establish the magnitude of the problem, with large amounts of users. A more reassuring secondary conclusion is that the users, given enough effort, did indeed show some tendency to pick the documents that the recommender system actually identifed as being good.

In the following sections, we first present the hybrid recommendation approach proposed for OpenLaws, and the *cold start* situation we will encounter as the OpenLaws infrastructure includes new jurisdictions and language areas. Next we present the method that we used for evaluating the approach in a mock up cold start situation, and the results of the experiment. We end with a conclusions and a discussion of the cold start problem.

2. Recommendation Approach in OpenLaws

OpenLaws is intended as an open infrastructure for legal professionals and organizations. Recommendation of documents by the system is therefore mainly seen as a collaborative filtering problem, sourcing the selection and annotation activities of the professional legal crowd using the portal. User selections and annotations will, over time, describe user-defined classes of documents. Because one first needs sufficient users creating the information to be used for the recommendation function, there appears to be no choice but including an initial mechanism for recommending documents where user data is absent or too sparse. This initial mechanism is (amongst other features) a topic model.

A *topic model* is a type of statistical model for discovering the abstract *topics* that occur in a collection of documents, by treating the documents as sets of observations. The technique used in OpenLaws, Latent Dirichlet allocation (LDA), is such a generative statistical model that allows sets of observations to be explained by unobserved classes that explain similarities between observations. LDA was first presented as a model for topic discovery [1], but essentially the same model was also proposed independently in the study of population genetics.

If observations are words collected into documents, LDA posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics (see the right hand triangle in Fig. 1). The same latent class discovery technique can however be applied to any kind of object that can be treated as a set of observations. It has been used as well as a technique for collaborative filtering, treating users as objects, and user selections as observations [6].

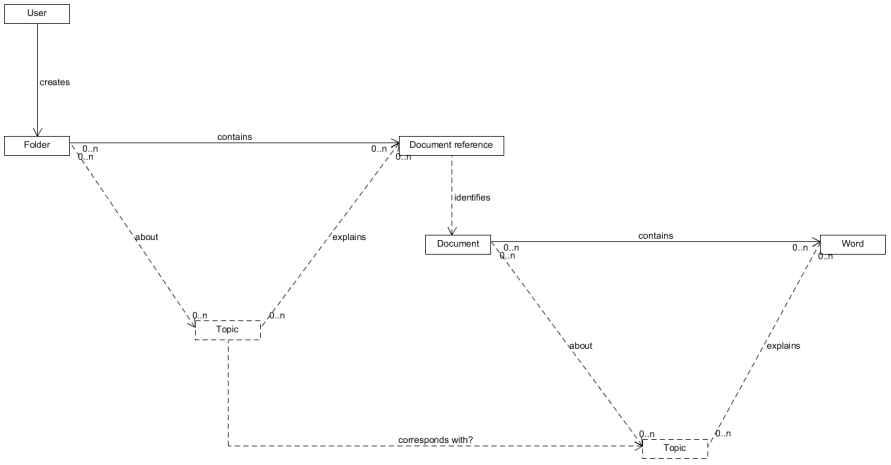

Figure 1. Two similar classification problems: latent topics in document content and in user-created folders.

For OpenLaws we envisioned a *hybrid* approach as follows:

- Selections and annotations made by a user on individual documents can be conceived of as placements of documents into folders defined by this user, at least as soon as the user attributes the same feature to more than one document (see user and folder in Fig. 1).
- LDA can be used to discover classes of folders by looking at these folders and the document references they contain as sets of observations (see left hand triangle in Fig. 1). This is an approach to collaborative filtering.
- LDA can be used to discover classes of documents by looking at the documents as sets of words as well (see the right hand triangle in Fig. 1). These classes can be added as folders containing document references. Topic models can be added as *faux users* creating folders of documents, by users using the topic modeling pipeline.

Treating LDA topic models as *faux users* creating folders with documents has the benefit of allowing multiple topic models for the same documents (some of which may be better executed or better evaluated than others) to co-exist peacefully in the portal, and permitting users to decide whether a topic model as a whole is trustworthy, as they can do with a user.

A seondary feature of the approach is that keyword searches can be treated as searches into LDA keyword indices, as a topic model produces an index of typical words that is produces for each class of documents that it identifies.

2.1. The Cold Start Problem

As collections of documents are added to the OpenLaws portal, we will have no control over whether the topic models added are good. Their utility will in any case be limited, but one can play with the parameters of the algorithm, and the

set of observations extracted from a document. In [7] we for instance applied topic modeling to references to other documents in a document only, reasoning that selecting a small subset known to be very relevant as a fingerprint might outperform using the whole text (which it did not). In the fielded system, there will be no control over whether topic modelers will perform any finetuning or evaluation.

Every time a collection is added, the system makes – as it were – a cold start (cf. [5]), behaving dumb for that collection and the users interested in it. This obviously gives reason for concern over the potentailly strong prompting effect of initial recommendations produced by a topic model, because adding such a topic model will be easy, and the topic model will account for most of the recommendations in cold start situations. This effect has been convincingly shown, for instance in [4]. First users can have a huge impact on the performance of collaborative filtering recommendation, and bad first users can cause bad recommender systems.

3. Evaluation Method

A prototype recommender system was built. In it users can search keywords and full text, or retrieve specific documents by ECLI for a specific query topic, and add these to a user-selected folder. The user is presented with recommended folders, both topic-generated and user-selected, but these are randomized, in the expectation that presentation order may dominate actual recommender scores in predicting user selections. We assumed that users, using the results of previous users, will get the answer more or less right eventually, so that we can measure progress towards a *best* user selection. We ran the experiment only once, because of limited access to test users. Test users were not rewarded for participating, and were mostly students, unfamiliar with the domain of interest (taxation). Test users were not warned about the randomization.

3.1. Data preprocessing

The dataset initially collected contained 25.095 Dutch court decision XML documents from 1994 to 2016 marked, by metadata, as being in the area of taxation. Each document is identifed by its European Case Law identifier (ECLI; e.g. ECLI-NL-CRVB-2005-AU3922). After stripping very short decisions we ended up with 10.658 useful documents. With these an LDA topic model was built in Mallet 2.0.7. As output of Mallet we get two databases:

- A distribution of keywords over identifed topics
- An allocation of documents to topics, with a score that indicates a proportion ($0 > p < 1$) of the document generated by that topic

After some testing we selected a distribution of the corpus over 150 topics, and generated folders from the output, adding documents by p score. At least 5 documents were added to each folder, but documents with $p < 0.7$ were discarded, resulting in 150 folders with 6.12 documents on average, ranging from 5 to 111

documents. Each folder is associated with the keyword list for the corresponding topic.

User-selected folders were processed to look indistuinguishable from LDA generated folders. LDA was used to generate a list of relevant keywords for the folder. Cosine similarity between folders was calculated for determining similarity between folders for recommendation.

3.2. Setup of prototype

A simple web-based prototype recommender system was developed. The system starts by presenting a page of instructions, including a description of the topic to look for, and then a page containing references to all documents, with the associated list of keywords. From this page, the user selects three relevant documents, which brings the user to a page of recommended folders, from which the user could pick new documents, resulting in a new page of recommendations, etc. Eventually the final results of the user are saved as a new folder in the database, for use by the next user.

The rank order in which recommendations were presented (presentation rank) was random. Actual recommendation scores were however recorded (recommendation rank).

3.3. Hypotheses

We were interested in the following hypotheses:

1 Users read only a limited number of documents in a folder The last selected document from a folder determines the number of documents the users have read at least.

2 Users tend to select the same documents as being relevant We differentate (i) the initial selection from the whole list, from (ii) the result after recommendations.

3 User-selected folders are favoured Documents from user-selected folders have a greater probability of being selected than documents from topic folders.

4 Presentation order predicts selection Presentation rank correlates with user selection order, both for (i) folders, and (ii) documents in folders.

5 Recommendation score predicts selection Recommendation rank correlates with user selection order of documents from a folder.

6 Presentation order dominates recommendation score Correlation of recommendation rank with probability of documents being selected from a folder will be higher than correlation of presentation rank with probability of folder being selected.

The tests that we could perform on the results of the experiment are discussed in the following section.

4. Results

A total of 28 users used the prototype, each using the database enriched with the user-selected folders of previous users. The amount of effort they appeared to have put in is considerable. Looking at the last selected document from a folder, users appear to have read at least 67.92% (average of 7.05) of documents in folders from which they selected documents. Users put in more work than we expected. Although three possible search topics were offered initially, users showed strong preference for learning about taxation of income from renting out one's house, and we removed the other two topics after some time.

Somewhat surprisingly, initial document selections from the entire list were completely unpredictable. Users selected 145 different documents, of which only 28 occurred in multiple initial selections. We did not log search phrases, but these must have differed considerably.

Final selections showed more overlap, with 90 documents being selected regularly, from a total 197. If we assume that the answer of the final user will be the best answer, we can calculate recall of this final answer by earlier users. Over time, we can see whether recall gradually increases. This is only the case from user 19 to final user 28. This casts doubt on the assumption that we can treat the final selection as a gold standard answer, but in any case suggests that in practice user selections were indeed hard to find among the topic folders.

Of the documents selected by users, 55.66% was copied from a user-selected folder (average 4.8 documents per folder), and 7.4% from a topic model folder (average 1.5 per folder). Although user-selected folders appear to perform much better, an independent t-test fails to find a significant difference. A considerable number of documents selected by the users was not copied from a recommendation, and these user selections do not produce useful data for testing our hypotheses.

Correlation between presentation rank and user-selection (0.186) within folders was considerably lower than correlation between recommendation rank and user-selection (0.223, or 0.353 if discarding users that copied no recommendations). Hypothesis 6 is unlikely to be true, but the scarce available data shows no significance of the difference. The recommendation function does appear to predict which documents are considered relevant by the users.

Correlation between presentation rank of folders and user-selection (0.555) was quite high. Of course, it is quite natural to start reading a the top when you get your recommendation, but the randomized recommendation could put fairly irrelevant results at the top spot, and this should be obvious, because users see the topic keywords.

5. Conclusions and Discussion

In advance, we did not know for sure how many documents there are in the corpus that are relevant to the search query. Apparently, there are at least 197. Users are unlikely to even try to produce a list of this length, and were guided by the experiment into considering a dozen selections as sufficient, making a lot of the

data collected of little use for the issue we are investigating. To do this experiment right, we need tighter control over the search topic and the desirable answer, while at the same time not manipulating average information distance (as defined in [3]). If we for instance embed 20 documents about taxation of income from renting out homes into a database of completely unrelated law, we expect better results, but then we are not solving the actual recommendation problem. It is clear, that to do this kind of experiment right in a corpus with a low average information distance, we need more users, multiple runs, and larger user selections. Evaluating randomized recommendations is decidedly unrewarding for test users, however. Overall, we find a confirmation of observation by [2] that content-based classification approaches do not perform well in legal settings. Although the recommendation score did predict user selections overall, users mostly avoided the folders made using the topic models in favour of user-selected folders.

We moreover do conclude that there is a risk involved in fielding the OpenLaws portal before solving the content-based recommendation problem adequately because of the prompting effect, which did show itself as a correlation between presentation order of *folders* and selections (0.555). Test users were not warned about the randomization, but did manage to select relevant (and recommended) documents from irrelevant folders. The magnitude of this problem cannot be established on the basis of this experiment, however, and to conduct a more realistic experiment would require considerable resources, as one should do multiple runs, with more users per run, using manipulated recommendations that must feel discouraging. Rather than a [4] *retrospective* study of a bad cold start on existing data, we need targeted cold start experiments on collections with realistic average information distance (as opposed to movies, books, etc.) for our field.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] G. Governatori. Exploiting properties of legislative texts to improve classification accuracy. In *Legal Knowledge and Information Systems: JURIX 2009, the Twenty-second Annual Conference*, volume 205, page 136. IOS Press, 2009.

[3] M. Li, X. Chen, X. Li, B. Ma, and P. M. Vitányi. The similarity metric. *IEEE transactions on Information Theory*, 50(12):3250–3264, 2004.

[4] P. Massa and P. Avesani. Trust-aware recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 17–24. ACM, 2007.

[5] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260. ACM, 2002.

[6] J. Wilson, S. Chaudhury, and B. Lall. Improving collaborative filtering based recommenders using topic modelling. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*, pages 340–346. IEEE Computer Society, 2014.

[7] R. Winkels, A. Boer, B. Vredebregt, and A. van SOMEREN. Towards a legal recommender system. In *Legal Knowledge and Information Systems: JURIX 2014, the Twenty-seventh Annual Conference*, pages 169–178. IOS Press, 2014.