Automatic Document Classification in Integrated Legal Content Collections

KEES VAN NOORTWIJK, Erasmus School of Law
KOEN VAN NOORTWIJK, Utrecht University[1]

Since most legal documents are released in digital form nowadays it has become more and more important to be able to search these documents adequately. Metadata, such as the area of law a certain document is about, included with digital publications can help to find similar publications faster, however the included metadata are very often incomplete. Automatic document classification can supplement essential metadata in legal documents released from different sources, such that integrated legal content systems can offer uniform options for the selection of relevant documents.

[1] Kees van Noortwijk is associate professor of Law and Technology at Erasmus School of Law, Rotterdam, The Netherlands. Koen van Noortwijk is graduate student in Artificial Intelligence at Utrecht University, The Netherlands.

1 Automatic classification of documents in heterogeneous content sets

Let us discuss a basic implementation of a search engine in order to understand document classification. A simple search engine might work in the following way: By counting the number of times each specific word is used per webpage in combination with a few other factors (which pages link to the respective page, what is the subject of those linked pages, etc.), all pages can be indexed. Then, when a search query is executed by a user, the search engine returns the indexed pages that rank the highest on the query. Although already very useful, such an engine can be improved easily by allowing users to add requirements to the returned pages. For example, when searching for videos of cats, you would only want the search engine to return pages that include actual videos of cats, and not pages that mention the words "cat" and "video". This behavior can be achieved by classifying all pages on the type of content they contain (pages could for example include videos, text and audio files).

Classification of text documents is in essence an elaborate version of the previously discussed procedure, where we are not looking to classify documents based on their type of content, but on the properties of the text they contain. In the previous example, our classes would have been types of content (text, video or audio), whereas now, classes of text properties could for example include a finite set of subjects (for instance if we know all documents to be about either cats, computers or law). By maintaining a set of positive examples for each class (the training set), we can start classifying new documents by assessing their similarity to each class on the assumption that all classes have a certain uniqueness that separates them from each other. Different algorithms exist that are designed to do this classifying automatically. Choosing between these algorithms is generally done based on the characteristics of the input data, since each algorithm responds differently to different circumstances, such as the size of the training set, the number of features per document and the independence of the features. Simpler models are usually preferred over more complex ones, since they provide greater insight into how classification comes about. However when trained sufficiently, complex models can generally outperform simpler models. Given the nature of an application in the legal field, we expect documents to show many great similarities within classes. Specifically since legal documents concerning the same area of law tend to use the same domain specific language, it is to be expected that a relatively simple model, such as a naïve Bayes, tree ensemble or k-nearest neighbors model will be sufficiently able to differentiate between classes.

A few similar applications already exist in very different fields. Perhaps surprisingly, a comparable application can be found in spam filtering, where the classifier is only provided with two classes, namely: spam, and not spam [1]. The classification is based on the fact that the content of any spam message shows high similarity to the content of other spam messages, such as the use of capital letters, or the use of specific sales related vocabulary. Other interesting applications of this classification technique include sentiment analysis and genre classification [2] [3]. In the first, documents are analyzed based on the general sentiment in the use of words, such that each document can be classified as being written in a, for example, happy, sad or angry tone. Secondly, genre classification automates distinguishing genres in documents such as library books, or song lyrics.

Based on the results of previously mentioned applications (accuracy of about 97% for spam filters and 66% for genre classifiers [2] [3]), it is to be expected that a classifier trained to classify legal documents can achieve a classification accuracy in a similar range as the genre classifier. A few challenges can be identified beforehand. Two of these are most noteworthy: incomplete training data and non-linear separability of classes. Firstly, the training data may turn out to be insufficiently homogeneous. It is imaginable to create a perfectly separated training set, but still end up with a classifier performing poorly, due to the fact that the data in each class in the training set are insufficiently similar. In this case, it is next to impossible for any classifier to create a well performing model. Possible countermeasures include leaving bad examples out of the training set, and adding more training data. Secondly, it is highly likely that certain words will be highly predictive for multiple classes when classifying legal documents, which can make it hard for a classifier to differentiate between these classes. The implementation of classification algorithms that are specifically designed to be able to handle non-linearly separated input data can be a possible solution, as well as the application of preprocessing steps where

the importance of certain features is modified, such that they play a lesser role when a scenario as described occurs.

2 Combining legal sources in integrated collections

The amount of legal information available digitally has grown tremendously in the past two decades. Legislation, case law reports as well as legal comments and other legal literature can be retrieved from online databanks, which tend to replace traditional 'paper' library collections in whole or in part [4]. Publishers play an important role here. They usually provide these digital legal resources on a subscription basis, by means of their own retrieval systems. Governments and the judiciary also publish important legal content, for instance legislation texts and case law reports. Furthermore, most law firms have their own 'internal' document collections, which also can be searched. All in all, it is not uncommon for a lawyer to make use of up to ten, or even more, separate content collections, each of them with its own specific retrieval functionalities. For a single task, it might be necessary to consult a number of these, repeating the same query over and over again in different retrieval systems.

As that situation is far from optimal, we now see an increasing number of initiatives to combine online legal content collections and make them available through one single retrieval interface. This is sometimes referred to as 'content integration' or 'content aggregation' and it can be a very effective way to simplify information gathering and processing for legal professionals. For basic full text searching and for browsing the available documents in such an integrated system, usually no specific additions or adjustments to the content are necessary. But when collections grow in size, supplementary selection mechanisms, such as the addition of drill-down options to choose certain subsets of a previously retrieved set of documents, might become necessary to enable users to pinpoint the required content. Categories for drill-down are usually based on metadata, added to each document. Common ones are for instance the type of publication (article in a journal, official government publication, (chapter in a) book, report, etc.), the source name (name of the journal, book or collection a document is part of) and maybe the year of publication. Such drill-down categories can usually be attributed to each document in the collection in a uniform way, even if the documents have totally different origins. But other drill-down categories, potentially very useful for lawyers, such as for instance the area of law a document is on, are often more difficult to apply. The reason is that not all documents contain metadata relating to these categories. It is not uncommon that up to 40% of all documents in an integrated collection contain no useable metadata to establish the area of law they relate to. In that case, such documents would not be selected if the user of an integrated retrieval system would choose to activate a particular area-of-law drill-down category, even if that area of law would definitely be applicable to them. The problem is often even more prominent in documents from private, 'internal' documents collections law firms maintain themselves, should such documents be part of the integrated collection that is queried. It is not uncommon that these internal document collections (for instance from 'know how' systems) only contain very limited metadata.

In order to avoid such problems, automatic classification of documents is a powerful addition to content collections in which not every document contains compatible metadata. Automatic classifiers have been around for quite some time now, but they are infamous for the need to be trained properly before they can be applied. Training a classifier for a particular category of documents could for instance involve the manual selection of a set of documents known to belong to that category, after which the classifier is able to select other documents for the same category by looking for certain 'features' obtained from the training documents. Although that approach might work, a major drawback is the effort needed for selecting the training documents, and possibly also to update the training set whenever the content collection changes (new types of documents added).

Integrating content collections and making these available to law firms now provides us with a powerful option to overcome these problems. For in such integrated collections, the probability is high that at least a certain number of documents contain metadata that attribute them to certain classes (for instance, a class representing a particular area of law). By using all documents containing the necessary metadata for a certain category as training documents for the automatic classifier of that category, and next applying that classifier to all documents lacking such metadata, manual training as well as manual updating of the training material can

be effectively eliminated. When this is supplemented with a simple user feedback feature, which enables a lawyer using the system to indicate incorrect automatic classifications, the result would be that the entire content collection available to a certain law firm can be queried much more effectively, precisely and efficiently than would otherwise be possible.

REFERENCES

[1] S. Youn and D. McLeod. 2007. *A Comparative Study for Email Classification*. In Elleithy K. (eds) Advances and Innovations in Systems, Computing Sciences and Software Engineering. Springer, Dordrecht, The Netherlands

[2] A. Pak, & P. Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREc*, Vol. 10, 2010.

[3] B. Kessler, G. Number and H. Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 32-38.

[4] Gorham, Ursual, and Paul T. Jaeger. The Law School Library or the Library at the Law School: How Lessons from Other Types of Libraries Can Inform the Evolution of the Academic Law Library in the Digital Age. *Law Libr. J.* 109 (2017): 51.