Large Scale Text Mining with Highly Accurate Detection of Negatives

Jakub Zavrel[a] Remko Bonnema[a] Martijn Spitters[a] Gert Meijerink[b] Gerard Mulder[a]

[a] *Textkernel BV, Nieuwendammerkade 28 a17, 1022 AB Amsterdam*
[b] *Unisys Nederland NV, Tupolevlaan 1,1119 NW Schiphol-Rijk*

Abstract

We describe a case study of a document understanding workflow system at Kadaster. The system saves around 75% of manual processing cost in coding semantically complex information from a 15+ million document archive. The system uses a combination of a highly accurate (>99,5% precision) single class classifier for negatives, and a domain specific region of interest detection module. The recognition modules are integrated in a web based workflow system supporting large scale distributed manual coding.

1 Application background

In many organisations, very large archives of electronic or paper documents exist, whose contents are crucial to the core tasks of the organisation. In the best cases, the documents are available in a full text information retrieval system. However, the information that is essential for the organisation is in the concepts, meanings and relationships in the text rather than in the keywords. Often the concepts are complex, can be expressed in text in many different ways, and require domain expertise to be recognized by a reader. The ideal situation would be to code the semantic information as meta-data in a structured information repository. The cost associated with converting the existing document archives by manual coding, however, are huge, and are often considered prohibitive. Automatic classification methods are often not considered applicable because state-of-the art recognition accuracy (90-98% correct) may not not meet the strict information quality standards of some organizations (close to 100% correct). In this paper we discuss a text mining approach that can achieve significant cost reduction in this process, and at the same time guarantee the required information quality standards for our client.

The above situation is the case at Kadaster (The Dutch Land Registration Authority). The document archive contains 15 million documents which are contracts about the transfer of ownership of real estate. The more modern part of this archive is available as PDF documents, the older part of the archive as images scanned from microfilm. The documents can be considered noisy text because a large part of the text has been obtained by OCR from images with very poor quality. The concept that is to be coded for the whole archive is the concept of « Erfdienstbaarheid » (explained in more detail below). In a pilot project it was determined that a human coding operator needs an average of three minutes per document to determine whether it contains the concept (positive class) and if so, code which lot numbers are in that particular relationship, and identify all text fragments that specify the concept. Hence the goal of the project was to save cost on a budget of five hundred thousand to one million man hours. The legally required information quality standard is that 99.5% of the documents have to be correctly coded.

2 Easements

The concept of *easement* (in Dutch: Erfdienstbaarheid) concerns the right of the owner of one lot of land (the dominant tenement) to restrict the freedom of another's use of land (the servient tenement), or to guarantee his own use of it. It is a right (and obligation) that is treated as part of the property itself, and as such it passes from one owner of a lot to the next. It is only established or abandoned (positive class) in a notarial contract in the Kadaster archive. Kadaster codes the concept as `<document id, date, dominant tenement id, servient tenement id, text fragment>`. The specific vocabulary specifying the right and its establishment or abandonment is highly variable. Table 1 below gives a few examples of indicative phrases. It is not feasible to arrive at a complete list of all possible easements and their formulations in text without reading the whole archive. An extra complication in recognition of the concept is that all contracts about a lot, that succeed a positive class document in time, will quote the relevant fragments from the positive class document verbatim. And a final non-trivial complication is the presence of high amount of character recognition noise. Some examples are given in Table 2 below.

Erfdienstbaarheid
Vestiging Erfdienstbaarheden
Dienend erf
Lijdend erf
Recht van overpad
Recht van weg
Recht van uitweg
Recht van pad
Wordt bij deze gevestigd
Bij deze wordt ten laste van
Gevestigd het recht van erfdienstbaarheid

Table 1. Indicative phrases

lijdende erTdienstbaar-
lijdende er f d ien s tbaarheden
eredienstbuarhnden
glijdende arfdienstbaarheden,
e r f dien stbaarheden
e r f dienstbaartieden
lijden d e eri'dienstbaarheden,
orfdlenstbaarhedön^
erraienstbaiirheden ^
irfdienstbaarheden,
lijdende erxdienstbaarheuen,
Cfjdicnstbaarhedcn

Table 2. OCR noise

3 Accurate single class classifier for negatives

The approach used to detect easement descriptions in the documents is a string matching approach based on the nearest neighbor classifier. Textkernel has a very fast string matching engine, based on q-gram matching and string similarity computation, that has been optimized for millisecond retrieval in text databases with millions of records, called FuzzServer. FuzzServer is able to exhaustively measure the similarity of all fragments up to a certain string length from the document against the text fragments from the training set. This makes the approach both robust to linguistic variation and large amounts of OCR noise. To train the single class classifier, we collected instances of positive and near positive fragments from the training partition of a small manually annotated training set (4000 documents).

The main opportunity for time savings within the project is the fact that only 10-20% of the documents are of the positive class. For the positive class documents the system identifies the fragments where easements might be mentioned. The manual coding of the parcel numbers of the tenements is still needed. However, it is very difficult to differentiate between truly positive documents and false positives (e.g. Containing verbatim citations of positive documents). We have therefore focused on training a highly accurate classifier for the negative class. We call this the problem of *being sure about what is not there.* This classifier not only has features about matches of positive fragments in the training set, but also about negative matches and about the total completeness and quality of the underlying OCR results. By tuning the thresholds of this classifier we were able to achieve a precision on the negative class of > 99.5% at a recall of 55% on a held out portion of the training data. This means that close to one half of the total document archive no longer needs to be read by humans, because we can be certain that it does not contain easements.

4 The workflow system

The solution at Kadaster is set up as a pipeline system consisting of Workflow Agents. The agents log the changes they make to the pipeline state in a central document database to track the flow of documents through the system. Documents are loaded in batches into the system's spool directory, where the Pipeline Filler agent picks them up and sends them to the OCR and recognition engines. The OCR and recognition engines work in parallel on 56 CPU cores to deliver the needed throughput. The results are delivered into the spool directory for manual coding. Kadaster operators work with Textkernel's web based workflow application, called Sourcebox to process the non-negative documents. In Sourcebox, documents are grouped into small sets for operators to manually code and check. After manual coding, quality assurance is performed on a daily basis by blind evaluation of small (1% of total) system and human coded output samples to ensure an continued accuracy of 99.5% on the whole document stream.

The workflow in Sourcebox is based on a number of roles: operator, manager, QA-operator, and QA-manager. A regular operator receives only those documents that have been classified as positive by the automatic classifiers. For these documents, the system highlights the fragments with positive matches in the image of the document, allowing fast orientation in the document. The workflow system achieves a further 50% time reduction for manual coding, bringing the total reduction to approximately 75% of the man hours budget.