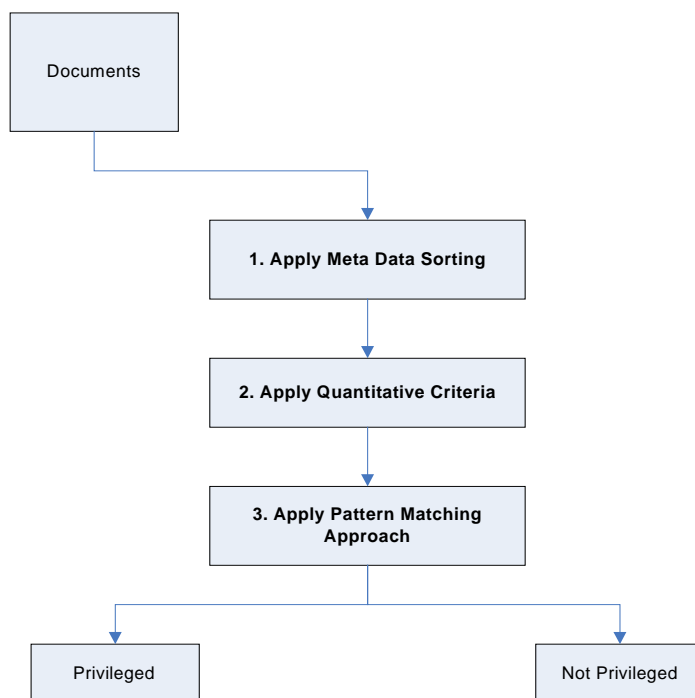


Auto Classification of Legal Documents

Torsten Volk, Steve Parker, Douglas Wolfire
January 2006

Goal: Utilize software to automatically classify a large quantity of documents as "Privileged" or "Not Privileged" based on certain objective and quasi-objective Criteria.

Classification Process: The classification process is typically divided into 3 stages:



- 1. Meta Data Sorting:** Basing the document classification on metadata serves as an effective pre-classification criterion. In other words, all documents carrying a specific meta property (such as "letter has been sent by attorney"), are moved into a preliminary category.
- 2. Sorting by clearly defined quantitative Criteria:** These criteria can be based on the occurrence of exactly defined keywords in a specific database field or part of a document. When analyzing structured documents, like letters, it often plays a role where on the paper a specific keyword is located. Applying these quantitative criteria helps further refine the preliminary categorization achieved in step 1.

3. **Pattern Matching**¹: Pattern matching follows the principles of Latent Semantic Analysis (LSA). LSA is based on probabilistic modeling and is not related with Artificial Intelligence (AI). In other words, the software **cannot** directly understand the content of a document. What it does is analyze term constellations (frequency of nouns, verbs, noun phrases etc.) in documents already categorized within the individual sentences and paragraphs and how they compare to term constellation occurring in a) the overall document, b) similar documents and c) in documents belonging to a different category. On this basis LSA applications can reliably calculate the likelihood of unknown documents belonging to a certain category, like privileged or not privileged.

There are 3 Phases of applying LSA based (pattern Matching) software:

The Learning Phase

Documents already categorized according to a company classification scheme undergo analysis which associates a semantic descriptor to them (frequency of nouns, verbs, noun phrases, etc.).

These documents are used as a basis for learning, and enable the software to create the categorization model using an algorithm that combines the various semantic descriptors assigned to the same category. A minimum number of documents per category, between 25 and 50 depending on the document size, is required to guarantee the model quality.

The Categorization Phase

New documents will be sorted according to the above established categorization model. Its semantic descriptor is compared to that of previously categorized documents. One or more categories are proposed for each document, with a confidence indicator.

Evaluation of automatic categorization

The categorization template is assessed as follows: 90% of documents already categorized are used for learning. The remaining 10%, which were not used in the learning phase, are used as a test corpus. The quality (precision and recall) can thus be measured for each category. Another method consists of validating the categorization model by using test sets.

¹ See our whitepaper on the Pattern Matching Approach (Latent Semantic Analysis)

Limitations of Software based Classification Approaches

Software cannot understand document content or determine the purpose of documents. Therefore, classification errors are unavoidable. The share of these errors can be as minimal as <1% depending on:

1. **The clarity of Categorization Criteria:** Before the classification effort, all criteria needs to be clearly defined. It is vital that there are no overlaps between sorting criteria for different categories. In other words, if “letter has been sent by attorney” is the only evaluation criteria and we can reliably determine whether this has been the case, the classification outcome will be 100% correct. If we have to rely on less clear cut criteria, like the content of the text in the document’s “body”, we need to be certain to provide the software with as many example documents embodying the respective category, as possible.
2. **Availability and reliability of Metadata and Quantitative Criteria:** The more the classification procedure relies on Metadata and quantitative criteria, the less effort will be required for achieving reliable results. Metadata and Quantitative Criteria (e.g. the word unicorn appearing at the top left of a document) are especially important when dealing with documents that are too short for the application of the pattern matching approach as a final sorting criteria.
3. **Number of correctly pre-classified documents:** Generally, the more correctly pre-classified documents are available for the software to be used as blueprint, the better the classification outcome will be.

Advantages of Software based Classification Approaches

1. **Consistent Results:** Once the software has been setup to reliably sort documents according to metadata, further quantitative data and text patterns, consistently high quality results can be expected in future.
2. **Fully Automated Real Time Processing:** As there is no human intervention required, document sorting can happen in real time.

The implementation of a classification engine can be divided into the following 3 steps:

1. **Requirements Analysis:** Due to the significant investment affiliated with the purchase and configuration of an Auto Classification Solution, a detailed requirements analysis will be crucial. It will exactly define the type and structure of documents to be classified, the quantity of documents, the sorting criteria, the requirements in terms of accuracy of outcome etc.

2. **Software Selection:** On the basis of a clear cut requirements profile, a suitable software will be selected.
3. **Software Installation, Configuration and Training:** Installing, configuring and training the respective software package can take from 2 weeks to 6 months, depending on the requirements in terms of complexity of the categorization scheme, structure of the target documents and accuracy requirements.

Cost: Depending on the complexity of the classification criteria, the quantity of categories required and the heterogeneity of the document structure, the implementation of an Auto Classification Solution could cost \$50-\$500k.

Conclusion: Fully automated classification of vast numbers of semi-structured and unstructured documents can be done to a high degree of accuracy. In cases where 100% accuracy is required, auto-classification tools can be used for document pre-sorting and so still save many staff hours.