

Retrieval of Case Law to Provide Layman with Information about Liability: Preliminary Results of the BEST-project

Elisabeth M. Uijttenbroek, Arno R. Lodder, Michel C.A. Klein, Gwen R. Wildeboer, Wouter Van Steenbergen, Rory L.L. Sie, Paul E.M. Huygen & Frank Van Harmelen

Centre of Electronic Dispute Resolution – CEDIRE.ORG, VU University Amsterdam
<http://best-project.nl>

Abstract. This paper describes the experiments carried out in the context of the BEST-project, an interdisciplinary project with researchers from the Law faculty and the AI department of the VU University Amsterdam. The aim of the project is to provide laymen with information about their legal position in a liability case, based on retrieved case law. The process basically comes down to (1) analyzing the input of a layman in terms of a layman ontology, (2) mapping this ontology to a legal ontology, (3) retrieve relevant case law based, and finally (4) present the results in a comprehensible way to the layman. This paper describes the experiments undertaken regarding step 4, and in particular step 3.

Keywords: concept-based search, case law, information retrieval

1 Introduction

[5, 9, 12] show that popular and influential applications in most countries are case-management systems. These systems helped to reshape the organization of courts and contributed to the reduction of case loads. Still, the judiciary is faced with more cases than they can handle.

Litigation is the traditional and public dispute resolution process, but several other so-called private dispute resolution processes exist of which the most prominent ones are negotiation, mediation, and arbitration. Arbitration and litigation are adversarial procedures, in which a third decides the case. Mediation and negotiation are consensual procedures in which the disputants aim at reaching agreement, either on their own or helped by a third called the mediator or facilitator. This third does not impose a decision upon the parties, but merely guides the procedure.

A decision to either go to court or to mediate (or negotiate, arbitrate) should be based on a well-informed choice. Currently the necessary information to make such a

decision is often lacking. One of the aims of the BEST-project¹ is to provide litigants with information about the expected outcome of a court proceedings.

In literature as well as practice of Alternative Dispute Resolution the Harvard method is influential. It is based on work carried out in the setting of the so-called PON: the Project on Negotiation. This Harvard Negotiation Project introduced the concept of principled negotiation, which advocates separating the problem from the people. Fundamental to the concept of principled negotiation is the notion of *Know your best alternative to a negotiated agreement* (BATNA).

In the BEST-project we are developing a system that supports users by retrieving relevant case law on liability. In this way parties are given the opportunity to form a judgment about whether they could hold another party liable for certain caused damage or if they could be held liable themselves. Also, parties can determine how much room for negotiation is available

We develop a system for intelligent disclosure of case-law in which the retrieval is based on search terms provided by laymen. The main challenge we face is to match the different terminology used in case law and by laymen. Laymen describe cases in their own words, which differs from the vocabulary used by legal experts and in legal texts. We therefore decoupled the task of giving a meaningful description of the legal case at hand from the task of retrieving similar case law from the public available case law database www.rechtspraak.nl.

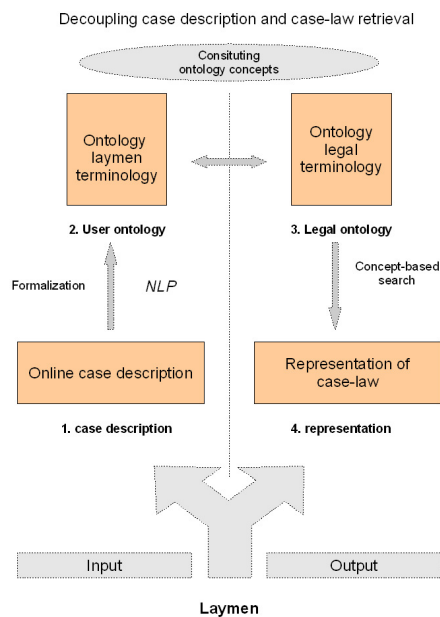


Fig. 1. Schematic overview of the BEST-project

¹ BATNA Establishment using Semantic web Technology, <http://best-project.nl>

Figure 1 shows a schematic overview of the retrieval process. First a case description is entered by the layman user. An ontology with layman concepts is used to structure the input and guides the user by entering relevant aspects of the case at hand. The laymen ontology is mapped to a second, legal ontology that is used for indexing case law. The retrieved case-law is then presented in a way comprehensible to the user and provides information relevant for his legal position.

In this contribution we first describe the thesaurus-based retrieval technique that we use and present the experimental results. Although the retrieval results were satisfactory, we considered this approach too static for relating the retrieval results to the laymen input. Therefore, in the following section 3 we propose a method to define search documents on the level of the concepts in a code section. This approach allowed us to use our thesaurus-based statistical retrieval techniques together with a visualization technique novel to the legal domain, which is described in section 4.

In section 5 a method is proposed to represent case-law to the users in a clear and comprehensible manner, based on recommender techniques.

In the last section we look ahead to our future work, that encompasses the first and second stage of the model.

2 Concept-based search

2.1 Applied technique

We need to obtain retrieval results that:

- are relevant to the case description of the laymen, and
- show the conditions necessary to establish liability.

This information should contribute to better insight by the layman about his legal position. The retrieval experiments we conducted with a statistical indexing technique are described below.

We used a thesaurus-based statistical indexing technique [13]. A thesaurus is used to create a vector representation of each document. Documents are compared by their vector representations. For searching a “query document” is created, and the vector representation of this query document is compared with the vector representations of the other documents [7].

This technique has been implemented in a commercially available software tool.² The main advantage of this technique over standard information retrieval techniques based on the *vector space model* [14], is that the indexing is guided by a thesaurus. This means that only terms relevant to a specific domain are taken into account. The indexing method works roughly as follows [11]. The indexing algorithm first detects sentences in documents and removes stop-words. After this it normalizes the remaining words, which means that nouns are reduced to the singular form and verbs to the first person singular form. In our experiments, we have used a specialized

² by Collexis BV, <http://www.collexis.nl>

normalization engine for the Dutch language. From these normalized terms or phrases, the relevant ones are then identified using a domain-specific thesaurus.

A list of the relevant concepts identified in a document is called a *concept fingerprint* of that document. For each identified concept a unique concept identifier is added to the fingerprint. This concept identifier is assigned a relevance score, based on term frequency and the specificity of the term in the thesaurus (which is the depth in the hierarchy), and the lexical similarity of the term with the textual contents. A fingerprint can be seen as a vector in a high dimensional space. The dimensions of this space are formed by the concepts of the thesaurus. The weight (or value) in each dimension is the relevance score for the concept in the document. The search is performed by a matching engine in the software, which matches a search vector with the vectors of the indexed documents. The vector for the search query is calculated in a similar way as described above. The matching engine will compute the distance between the query vector and the vectors of the documents. The result of the matching engine will be a set of document vectors sorted on their distance to the query vector. This is presented to the user as a ranking on relevance of the indexed documents.

2.2 Experimental set-up

2.2.1 Data source

Although the case law database used to disclose similar cases, the public website www.rechtspraak.nl, can be accessed online, for processing purposes we have locally stored all available cases. The approximately 100.000 cases is a low number, given the over 1 million legal verdicts annually. Nonetheless, this database contains almost all digitally available newer case law (1999-) in the Netherlands. The verdicts have some meta-data attached to them, e.g. the location of the court, the date of the verdict, a unique identifier (LJN), and for around 50% of the verdicts (the newest) a summary of a few lines. Internally, the documents have no computer parsable structure, but are plain text.

2.2.2 Research questions

First of all, we wanted to know whether a concept-based search technique as described above is suitable for the retrieval of case law in which a prototypical legal case is described. To obtain relevant retrieval results - that is a prototypical legal case similar to the case described by the layman - an effective search document has to be created. We conducted different experiments to find out what the best method is to create a search document.

In our first experiment we distilled the relevant terms for a specific legal case category from Code sections. So the search documents consisted of the terminology used in the text of the Code (we call this: *code-based fingerprints*). In our second experiment, we did the same for case law as we did for the Code, so now created case-based fingerprints. The search documents in this experiment consisted of terminology used in case law. Since we did not use automated techniques we labelled this method *case-based manually created fingerprint*. Finally, we selected a number

of relevant cases and used these together as one search query (*case-based automatically generated fingerprints*).

Because terminology in the code differs from terminology in case law, we expected that the search with case-based fingerprints would be better suited to identify relevant cases. Second, we expected that the indexing process for the *generated fingerprints* would automatically distinguish the most important terms and therefore perform better than the manually created fingerprints.

2.2.3 Procedure

We started with a selection of specific legal categories for which we wanted to identify relevant case law. We chose three fairly different types of liability and a fourth one that is related to one of the other situations:

- liability for misleading advertisements;
- liability for non-subordinates;
- liability for real estate;
- liability for subordinates.

The reason for this choice is twofold. First, diverse legal cases allows us to check whether our set up is suited for distinguishing legal cases at all. Second, the broad selection prevents us from drawing conclusions based on a not representative subset of liability. The idea behind the two similar situations (liability for subordinates / nonsubordinates) is that this will learn us whether the technique is also able to distinguish different legal cases that are quite similar to each other but are based on different Code sections. For each of the four specific legal categories we did the following three things:

1. we distilled the relevant terms from the Code;
2. from the court decisions database we selected a number of relevant cases;
3. we analyzed relevant cases and made a list of important terms used.

In addition, we created a thesaurus with legal terms. The thesaurus is manually created from the terms identified in task 1 and 3 in the list above and the terms identified in the Code for other types of liability. This resulted in a thesaurus with 360 concepts. The structure of this thesaurus is imposed by the structure of the law itself, i.e. “liability” is the root concept with more specific types of liability below it, e.g. “liability for persons”, which in turn has “liability for subordinates” below it. The relevant terms are placed below the types of liability for which they hold, including some synonyms. The thesaurus is used to index the data set (i.e. creating fingerprints for each document in the repository) and the other material is used to create different search documents for which fingerprints are calculated. We then used the search documents’ fingerprints to search for relevant cases. The top of the highest ranked results were evaluated on relevance.

2.3 Results

2.3.1 Code-based fingerprints

In a first set of experiments, we evaluated the code-based fingerprints, i.e. the fingerprints with terms distilled from the sections of the code. We did not expect very good results here, as we assumed that the vocabulary used in the cases is different from the vocabulary in the code text. Nevertheless, Dutch law is built on the Code (in contrast with the Common Law tradition), so we could ignore this in our retrieval process. As can be seen in Table 1, the correctness figures for the 10 highest ranked are indeed quite low. For two fingerprints, this set did not contain any relevant result at all. In the other one, we only found 3 relevant cases, but also two cases in which the article searched for was only casually mentioned. Note that for section 6:170, we found 8 slightly relevant cases.

Code sections	Correctness
6:170 subordinates (including slightly relevant 90%)	10 %
6:171 non-subordinates	0 %
6:174 real estate (including slightly relevant 50%)	30 %
6:194 misleading advertisements	0 %
6:162 unlawful act (including slightly relevant 48%)	40 %

Table 1. Correctness figures for code-based fingerprints

2.3.2 Case-based manually created fingerprints

We did the same experiment for case-based manually created fingerprints—fingerprints based on important terms identified by the expert in a selection of the case law. This resulted in the figures printed in Table 2. For two of the three sections the results are fairly good. For one article, the results are not so good; interestingly, this is a section for which a good result was obtained for the code-based fingerprints.

Code sections	Correctness
6:171 non-subordinates (incl. slightly relevant 70%)	50 %
6:174 Real estate	10 %
6:194 Misleading advertisements (incl. slightly relevant 92%)	83 %

Table 2. Correctness figures case-based manually created fingerprints

2.3.3 Case-based automatically generated fingerprints

Thirdly, we evaluated the performance of automatically generated fingerprints—fingerprints based on the full text of a set of pre-selected relevant cases. We started with fingerprints based on 5 case descriptions for 3 different legal cases. The results vary for the different Code sections (see Table 3). The table lists the number of cases used to create the fingerprint, the number of relevant cases as fraction of the total

number of evaluated cases, and this fraction represented as a percentage. We evaluated the relevance of the first 15 returned documents, but we did not count the documents that were used to *create* the fingerprint. This explains the difference in the totals in the column with the correctness.

Code section	Corectness
6:170 subordinates	70 %
6:171 non-subordinates	0 %
6 :174 real estate	37 %
6 :194 misleading advertisements	77 %
6:162 unlawful act	32 %

Table 3. Corectness figures of automatically generated case-based fingerprints

A hypothetical explanation for the diverse results is that the sets of documents from which the fingerprint are generated are too small. To check this, we generated a fingerprint from a larger set of documents (20 cases) for the worst performing legal case, i.e. “real estate”. Because the total number of cases in the data set for real estate are between 40 and 100, it is not realistic to base fingerprints on much more than 20 documents. As can be seen in the table, the results were still not good: only 2 out of the 10 cases highest ranked were relevant.

Finally, we have generated a fingerprint from a very large set of documents. We used the general “unlawful act” section 6:162 for this, as this is the only section for which we had enough cases (around 500) to do this experiment. The fingerprint for section 6:162 is based on 249 cases. To our surprise, the results were still disappointing: only 8 from the 30 highest ranked cases were relevant and not yet used to create the fingerprint. Even when the cases about the related section 6:174 were considered as relevant, we only count 13 cases. Moreover, the first case which was not part of the fingerprint appeared to be irrelevant.

2.4 Additional experiments

While looking for an explanation for the results of the previous experiments, especially the under-performance of the fingerprint for “liability for real estate”, we considered that the wide variety of the factual situations underlying a specific legal category probably blurred the legal similarity. For example, “liability for real estate” copes with all kinds of real estate, including roads, and accidents with all kinds of vehicles because of shortcomings in the road. However, we also found out that there are *typical phrases* that are used to prove a specific type of liability. Therefore, we extended the *case-based manually created fingerprints* with such phrases. We distinguished the different argumentation lines used to prove something and typical phrases used in the judges’ argumentation. On average, we added around eight phrases per legal category. Translated examples of such phrases for “real estate” are:

- “causing danger for persons or objects”,
- “owner of a property”, and
- “requirements that in a given situation”.

We have added these phrases also to the thesaurus and re-indexed the complete repository. The results of this experiment are listed in table 4. The figures indicate that there are more relevant cases returned than in previous experiments. What is also interesting, but not visible in the figures, is that the ordering seems to be better than in previous experiments: the relevant and irrelevant cases were less intermixed than before.

In this experiment we also counted the number of relevant cases that did not explicitly mention the section number. These are interesting cases, because they can be found by relevant wording only, and not because the section number is mentioned. As can be seen, there are at least some cases that are relevant, but do not literally contain the section number.

Code section	Correctness
6:170 subordinates	88 %
6:171 subordinates (including slightly different 53%)	47 %
6:174 Real estate (including slightly different 87%)	80 %
6:194 Misleading advertisements	80 %

Table 4. Correctness figures for case-based manually created fingerprints after adding legal phrases

Finally, we redid the last experiments with no other concepts in the thesaurus, i.e. we reduced the thesaurus to the four different types of liability and their relevant phrases. This resulted in a thesaurus of 25 concepts expressed in 50 terms (i.e., 25 synonyms). When using this thesaurus to index the complete data set, around 7000 documents (out of 68.000) were ignored because none of the terms in the documents were similar to terms in the thesaurus. The remaining documents were indexed with only 1.08 terms on average. This suggests that sensible results are unlikely, because it almost means that for each document a single keyword is attached. The correctness figures for this method were still quite high, 70% for the first 10 hits in for “sub-ordinates liability”. However, all of them literally contained the section number. As we have seen in one of the previous experiments, there are also relevant cases in which the article number is not literally mentioned.

2.5 Discussion

Several observations can be made.

First, we noted that only for one legal category (“liability for misleading advertisements”, Section 6:194) the results for the *automatic case-based generated fingerprints* were notably better than the *code-based fingerprints*. A possible explanation is that the specific code text uses very abstract formulations, which have only a few terms in common with actual cases. We noted that the code specifies a non-exhaustive list of possible misleading statements (“about the contents”, “about the amount”, etc.). The terms used in this list of typical misleading statements will not frequently occur, as they describe statements at an abstract level. These abstract terms

are different from the concrete terms that are used in case law. Thus, even although case law contains the term 'misleading advertisement' very often, the resulting fingerprint will be quite different. The automatically generated fingerprints from the cases do contain the concrete terms from the cases, of course.

A second interesting observation is that when using code-based search, we found for some of the legal categories (e.g., sections 6:162, 6:170 and 6:194) many *indirectly relevant* cases, i.e. cases in which the article was only casually mentioned. This finding can possibly be explained by the *interpretive* character of the legal concepts mentioned in the code for these articles. When such concepts are not precisely defined the legislator intentionally left room for interpretation by judges. Legal reasoning that involves interpretation, is a manifestation of the application of a vague concept. An example of such a vague concept is 'the reasonable man' or 'an act or omission violating a rule of unwritten law pertaining to proper social conduct'. In situations where vague concepts are used case law determines the meaning of these concepts. Court decisions often refer to a concept with an interpretive character, which causes a lot of indirectly relevant retrieval results. Therefore, a high number of *indirectly relevant* cases would be a sign of code text that is characterized by interpretive concepts.

Another observation, which is not directly visible in the figures, is that the analysis of the results showed that the type of cases returned for the *automatic case-based fingerprints* and the *code-based fingerprints* are very different for sections 6:171, 6:174, and 6:162, although the percentages of correctness are comparable. *Code-based fingerprints* resulted in cases that literally contained some non-interpretable concepts in code sections, while *case-based fingerprints* resulted in cases that define the meaning of interpretive concepts in the code. This suggest that *code-based fingerprints* are useful for finding *non-interpretive concepts*, e.g. concepts that have a precise meaning in the law, while *case-based fingerprints* are more useful to find *interpretive concepts*. This is in agreement with the intuition that the meaning of interpretive concepts is defined by case law.

Another interesting finding is that manually created fingerprints in general perform better than automatically generated fingerprints (except for one of our examples, i.e. "real estate"). This is contrary to what we expected. This might have to do with the large number of real world situations in which some legal concepts can be relevant. To describe these situations different (ambiguous) terms can be used. It is therefore more difficult to distinguish them only by looking at the terms used. This is in particular a problem for the automatic method, as it uses the number of occurrences of the terms as the measure to calculate the relevance. When manually creating fingerprints the most irrelevant terms are probably left out.

Finally, we have seen that by adding typical legal phrases the results improve. There are more relevant cases returned and the distinction between relevant and irrelevant seems to be crisper. However, the phrases alone are not sufficient. It seems that the phrases help to eliminate irrelevant cases in the top of the ranking (improve precision), but that additional concepts in the thesaurus are required for finding relevant documents that do not contain the literal article number (improve recall). A hypothesis is that the phrases are especially helpful for retrieving the concepts that need additional interpretation, i.e. the vague concepts.

3 Search documents

3.1 Concept-based search documents: technique enabling visualization in a later stage

In the experiments described above we created search documents for each section of the code. The conditions to establish liability can be found in the relevant code section. To provide laymen with relevant information about his legal position, it is necessary to make at least clear which conditions need to be fulfilled to establish liability. For this reason we conducted a following series of experiments. We created search documents for each condition necessary to establish a specific type of liability. For example, in Dutch tort law liability based on the general section 6:162 BW can only be established if the following conditions are fulfilled:

- the presence of an unlawful act (that is: an infringement of a right, a violation of a statutory duty, and an act or omission violating proper social conduct);
- damage;
- a causal relation between the act and the damage;
- accountability.

For each of these conditions search documents were created. We did this for 15 different sections of Dutch tort law. Tort law doctrine has been used to determine the necessary conditions. However, doctrine was not always decisive. For the retrieval of case law also other factors should be taken into account, such as the relevance of a concept in the light of the case law to be retrieved or the different contexts in which the same concept is used. The following criteria have been used to divide a section into legal concepts:

- a. The legal concept should have a certain level of broadness to make it applicable to a large category of case law;
- b. The legal concept should be precise enough to be relevant in a particular factual context;
- c. Tort law doctrine is the leading guideline;
- d. Coherence between the different concepts distinguished.

3.2 Open textured and clear concepts

Legal reasoning is indeterminate due to its open, procedural nature [8]. Bench-Capon & Sergot [1] share the view that indeterminacy of law is a consequence of open texture. They define an open textured term as one whose extension or use cannot be determined in advance of its application. This means that the application of an open textured concept in code sections cannot be derived from the code itself. Open texture

is the main reason to treat the legal domain as a specific domain of retrieval. We used the following indicators [cf. 15] to determine the open textured character of a concept.

1. *Ambiguity* - A term is ambiguous if there are more definitions for one concept. Dutch Tort Law terminology is characterized by ambiguity. For example, the term ‘accountability’ could relate to the establishment of liability but it is also used to determine the amount of compensation that has to be paid.
2. *Granularity* - The degree to which a concept is abstract in its nature. Such as “amount” or “duration”.
3. *Discretionary statutes* - Only the framework for discretionary room can be given, but discretion can be described in the form of a “shopping list”. For example in section 6:194 different circumstances under which an advertisement will be judged misleading are enumerated.
4. *Jurisprudence* - Judges often give an interpretation of relevant, vague concepts. An example from section 6:162 is ‘an act or omission violating proper social conduct’.
5. *Socio-political environment* - A changed socio-political environment could indicate that a certain term is subject to interpretation. In section 6:175 regarding the liability for waste products it is determined that a product will under any circumstances qualify as a waste product if a legally binding decision said so. New waste products come and others disappear, and the legally binding decision can be adapted to the newly identified (dangerous) waste products.
6. *Completeness of knowledge*- The last indicator of open texture is the completeness of knowledge in a specific domain or field. If there are two or even more definitions for the same term, classification ambiguity comes into play. To obtain relevant retrieval results an ambiguous concept should be characterized as an open textured concept and treated as such. The term “work” is an example of a term that leads to classification ambiguities. Work can relate to labor law issues but also to the object of copyright infringements (the created work). Search documents need to be defined in such a way that retrieval results are restricted to the right interpretation of a specific term.

Clear concepts do not have to be interpreted. An example of a clear concept is an act violating a statutory duty. All the possible violations can be found in the Dutch code. In case law the reason for unlawfulness of the act, such as acting in conflict with the obligation to identify, or the relevant section, can be mentioned.

3.3 Creation of search documents

Distinguishing between clear concepts and open textured concepts is relevant for retrieval, for it indicates a difference in the way natural language is used [6]. If code text is used literally for the creation of a search document, the retrieval results will be poor for open textured concepts because these concepts are interpreted or complemented by the judge.

Case-based fingerprints are search documents created for open textured concepts. These fingerprints are based on the terminology used in case law. The open texture necessitates that concepts are interpreted. Although it is not possible to determine the

full scope of interpretation in advance it is possible to give an estimation about the room left for interpretation. Court decisions were manually analyzed to distil relevant terms for an open textured legal concept. For clear concepts code-based search documents were created. In case of clear concepts the code text alone suffices to obtain relevant retrieval results. The following two decisions have to be made for each search documents:

A. Code-based or case-based - The search document should be either based on code text or on case law terminology;

B. Level of abstraction - The search document should be abstract enough to retrieve as much relevant court decision as possible. Different legal categories are distinguished in case law for the concept “an act or omission violating an unwritten law pertaining to proper social conduct”. These include situations of sports & play, negligence, creation of danger, etc. The search document therefore has to comprise all these categories. However, it is not necessary to define every sports & play situation there is. It is unnecessary to comprise terms as “tennis”, “football”, etc..

3.4 Experimental set-up

3.4.1 Data sources

For these experiments we also used the case law database of the public website www.rechtspraak.nl. See for more information section 2.2.1.

3.4.2 Procedure

For each of the legal concepts a search document is built as described in the previous section. The retrieval software is used to query the database for documents (case law) similar to the search document, which results in a ranking of all documents. Subsequently, the 30 most similar documents for each of the search documents are analyzed manually on their relevance. To determine the relevance of retrieval results for code-based or case-based fingerprints we set the following criteria. The retrieved court decisions should interpret or mention the norms relevant to the legal concept for which the fingerprint had been created. The legal concept can be mentioned literally, but a description of the relevant concept is also sufficient. The concept has to be mentioned at least indirectly.

3.4.3 Results

In table 5 an overview is given of the retrieval results for 5 essential concepts. The results showed a relevance of approximately 70% (see Table 5). The relevance score for concepts created for sections of the code that are not applied regularly were below average, while the relevance score for concepts of often applied sections were above average. Obviously, less court decisions are available in the database for the sections that are less regularly invoked.

Section	Concept	Type	Relevance
6:162 BW	Act or omission violating an unwritten rule pertaining of proper social conduct	Case-based	100%
6:170 BW	Say over subordinates	Code-based	69%
6:174 BW	Danger for persons and objects	Code-based	90%
6:174 BW	Realized danger	Code-based	58%
6:174 BW	Requirements under certain conditions	Case-based	100%

Table 5. Relevance scores for individual concept queries

4 Visualizing overlap between concepts

4.1 Motivation

Each search document of the conceptual retrieval technique just elaborates upon a single concept. The retrieval software calculates a similarity value between the search document and all documents in the database. This results in a ranking of all case law according to its similarity with the search document. We assume that a similarity above some threshold value implies relevance of these retrieved cases for the concept queried for. The threshold value is pragmatically chosen such that it provides a good balance between precision and recall for all query concepts. Because each code section is split into several concepts and hence search documents, an intuitive assumption is the following:

The relevance of a retrieved case for a specific code section increases with the number of concepts of that code section for which this case is relevant. Therefore, the intersection between the sets of retrieved cases for concepts of the same code section are probably the most relevant cases.

4.2 Procedure

For the visualization of the clustering of cases we use the clustermap viewer from Aduna2³. This software creates Venn-like diagrams of objects and show if they belong to one or more sets. It allows for dynamically adding and removing of set specification, which can be helpful to see the effect of using different sets on the grouping of the objects.

Each object, in our implementation a court decision, is represented as a sphere. All retrieved cases that contain a specific concept are clustered and visualized as

³ <http://www.aduna.biz>

amoebalike shapes (blob shapes). If an object belongs to multiple clusters (which means that a court decision is relevant for more than one concept), the blob shapes overlap and the object is displayed in the overlap. The software can be configured in such a way that the darkness of the areas reflects the amounts of overlap. Therefore, one can immediately see which objects are in the highest number of clusters. For each object links to e.g. webpages can be added. In our implementation, we created direct links to the online version of the verdicts. This link points directly to the verdict at the website of rechtspraak.nl. Thus, our local database is only used to calculate the similarity between the cases and the search documents, but is not used to display the case to the user. An interface has been written that connects the Collexis search software to the Aduna clustermap viewer. This interface allows formulating queries for sets of concepts. We use this interface to specify sets of legal concepts that together represent a section of the code.

4.3 Visualization experiments

We did some experiments with different combinations of the concepts for which we defined search documents. We chose the sets of concepts in such a way that we were able to visualize overlap between the cases for concepts that together establish a certain kind of liability.

We defined 28 combinations of legal concepts for 15 Code sections. We obtained approximately 900 different court decisions for the different combinations of legal concepts. The retrieved court decisions were sometimes partly overlapping for different combinations of legal concepts. Searches for some concepts resulted in a relative small number of cases (e.g. around 5), others in a much higher number (around 200).

For the evaluation of the results we set the following criteria. Court decisions are relevant if they deal with the type of liability, for which we created a specific set of clustered concepts, resembling the conditions that need to be met to establish liability based on a specific section of the code. For example, for the code section about “wrongful acts” a set of legal concepts is created, comprising the concepts “causality”, “damage” and “wrongful act”. The court decisions showed in the overlap between these concepts, handle about wrongful acts, and contain all three constituting concepts.

4.4 Examples

In this paragraph the search for cases about specific code sections is illustrated with three examples.

4.4.1 Real estate

The first example (see Figure 2) shows the cluster map for concepts that constitute liability for real estate” (section 6:174). The concepts we considered are “real estate”(fp25), “possessor of real estate” (fp2), “danger for persons and objects”

(fp12), “requirements under certain conditions” (fp6) and “realization of danger” (fp35). There are quite some cases in which “danger for objects and persons” (fp12) and “requirements under certain conditions” (fp6) play a role. There is also a reasonable number of cases in which both concepts are present. The picture shows that there is only one case in which all concepts are important. Inspection learned us that we obtain more relevant court decisions if the concepts “damage” and “possessor of real estate” were not included.

The relevance score is 100% if these concepts are excluded. These concepts are too broad (“damage”) respectively to precise (“possessor of real estate”) in formulation.

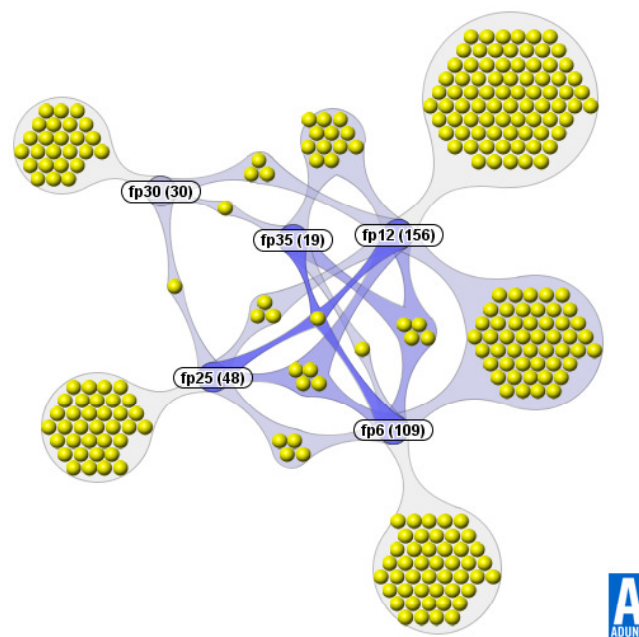


Fig. 2. The visualization of the grouping of relevant cases (yellow spheres) by the essential concepts of “liability for real estate”.

4.4.2 Liability for subordinates

A second example (see Figure 3) illustrates the clustering of cases for “liability for subordinates” (Section 6:172 Civil Code). The essential concepts are “fault of a subordinate” (fp9), “probability of a fault” (fp17), “say over subordinates” (fp36) and “damage to others” (fp28). In this example, it is immediately clear that there is no overlap between the documents returned for “damage to others” and the other returned documents. It also shows that there are eight cases for which three of the essential concepts are relevant. Those are included in the “darkest” part of the diagram. It is also interesting to see that it almost doesn’t happen that a “fault of a

subordinate” is relevant in a case without “say over subordinates” being relevant. Based on this we could hypothesis that the requirement “fault of a subordinate” is not very important when retrieving case law, as all that these documents are already retrieved when searching for “say over subordinates”.

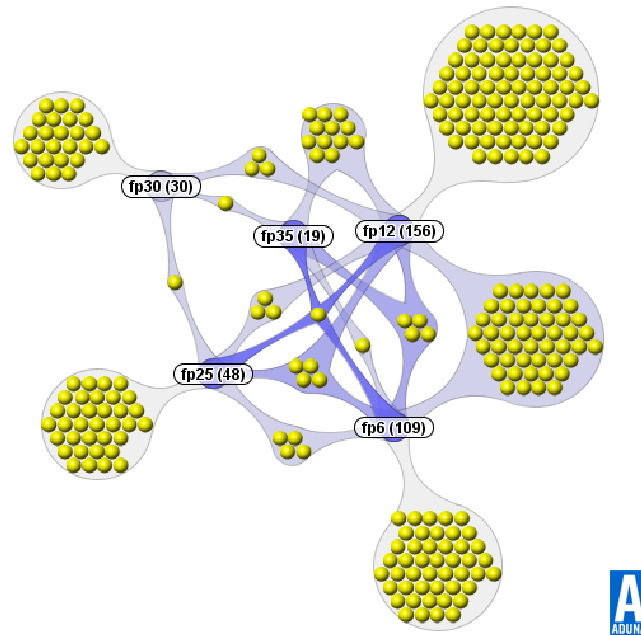


Fig. 3. The visualization of the grouping of relevant cases (yellow spheres) by the essential concepts of “liability for subordinates”.

4.5 Results

The results (see overview in Table 6) showed relatively high scores on precision for the sections that are often applied to establish liability, such as the general tort law section. Poorer results in overlap were found for sections that are not often applied, such as the liability for representatives. The results showed an average of 60% relevance. The results for often applied sections show results up to 100%, while sections of the code that are rarely applied resulted in a relevance score of less than 40%.

Section	Essential concepts	Relevance score
6:162 BW unlawful act	Causality, damage and an act or omission violating unwritten law pertaining to proper social conduct	72% if the concept “damage” is not included. This concept proved to be irrelevant
6:170 BW subordinates	Fault of a subordinate, probability of fault, say over subordinates and damage to third	100% if the concept “damage to third” is not included. This concept proved to be irrelevant
6:171 BW non-sub-ordinates	Fault during work activities, non-subordinate, damage	38% if the concept “damage” is not included. This concept proved to be irrelevant
6:174 BW real estate	danger for persons or objects, requirements under certain conditions, realized danger (damage and possessor of real estate were irrelevant and not included in the evaluation of the results)	100% if the concepts “damage” and “possessor of real estate” were not included. These legal concepts proved to be irrelevant

Table 6. Overview of the results

The precision in general is good for some of the concepts. These results were in most cases better than the straightforward approach as described in section 2. We hypothesize that the poor results for the clusters of concepts that resembled less applied sections of the code is possibly also due to the fact that www.rechtspraak.nl exists since 1999 and that few court decisions about certain types of liability are available. To validate the recall, we used standard court decisions that contain the basic interpretation and argumentation for a liability section of the code. All these court decisions were from before the launching of rechtspraak.nl (1999), and therefore added to our database. We hypothesized that these basic court decisions would be displayed by the clustermap viewer. Poor results for the recall were obtained. The court decisions relevant for a specific category of liability were not displayed by the clustermap viewer. These poor results on the recall could possibly be explained by the use of different terminology in older court decisions that is not used in the fingerprints. Another possible explanation for the poor recall results is the limited manually composed thesaurus or the limited use of terminology for the manually created fingerprints.

The clustering results show that some concepts can be omitted. An example of a redundant concept for the retrieval of case law is “damage”. The redundancy of this concept could be explained through the neutral character of the terminology related to the concept of damage. The concepts that combine possession and an object, for example “owner of real estate” seem to be too detailed and exclude a lot of relevant

court decisions. If we observed the clustering results of concepts that only relate to the object, such as “real estate”, the results for a set of concepts improved tremendously. Only 28 court decisions for the concepts “possessor of real estate” were retrieved, from which only one was part of an overlap, while for the concept of “real estate” 38 decisions were obtained, from which 9 were part of an overlap.

5 Presenting relevant court decision

5.1 Motivation

Besides retrieval of relevant case law, the comprehensible presentation of the retrieval results is an important part of a successful system to provide laymen with information about their legal position. We assume laymen will have a problem reading the verdicts and understanding the different legal concepts, i.e. the conditions to establish liability. To present an understandable explanation of the relevant verdicts, we take two steps. First, we localize in the verdicts the legal concepts that are relevant for the user’s case. With techniques from recommender systems we then decide which paragraphs are relevant for which concepts and we present the user the verdict based on these relevant paragraphs and apply also other recommender techniques. We also carried out a small user satisfaction research to find out whether the proposed presentation is indeed useful to prospective users.

5.2 Technical implementation

Recommender Systems are usually divided into two approaches: Collaborative Filtering (CF) and Content Based Filtering (CBF). In Collaborative Filtering, the preferences of communities of similar users are used to decide on recommendations for the current user [4]. With Content Based Filtering the content of certain items is processed and based thereon a decision is made about whether the user will probably be interested in the item or not, based on some predefined user characteristics and a history of interest in earlier items [3].

Since we want to process the court decisions on content to explain their relevance Content Based Filtering might be helpful. The content of a paragraph decides whether the user will be interested in that paragraph or not. Of course, this is not based on the preferences of the user, but on the relevance of the legal content.

Another interesting prospect is that Recommender Systems sometimes provide a *reason* for the recommendation. An ‘explanation mechanism’ tries to explain why the program believes that the user will be interested in the prospective item [10]. We investigated whether the techniques used to establish the reason for recommendation are also feasible for explaining to the user why those specific verdicts are presented to him. However, in Recommender Systems the search for recommendable items is tied to the reasoning about why a certain recommendation was made while in our

research, the search is conducted separately. Only afterwards we aim to re-establish the reasons behind the selection of the final set of verdicts. Also, history information about earlier recommendations is not available. As follows, the technique can only be applied on the content of the court decisions under scrutiny at the moment.

To test the effectiveness of the explanation system a small satisfactory research is conducted. Our basic assumption is that the explanation should convince the user that the presented verdicts are relevant for his own case. This relevance can exist in more in-depth information about the similarities and dissimilarities between his own case and the court decision represented by the system. Explanation systems that concentrate on this aspect are Keyword Style Explanation and Influence Style Explanation [2]. In Keyword Style Explanation the user is given a table explaining which words in his profile and in the content of the item had the most influence on the rank of the item. This can possibly be applied in our project to the occurrence of fingerprint terms. In Influence Style Explanation, the system tells the user how their interactions with the recommender system influenced the recommendation. In our project it might be possible to use this with the original description of the user case.

For the application of these techniques, we need to localize the legal concepts in the verdict, since they determine whether the content is relevant. This localization is described in the next section.

Since experiments showed us that it is impossible to localize the legal concepts that are extracted from the user's case in a direct manner (e.g. by keyword search), we decided to use the fingerprints from the search part of the project for localization. In GATE (General Architecture for Text Engineering) we first tokenized the relevant verdicts, then stemmed them, used a gazetteer to annotate words and phrases belonging to a concept, based on their fingerprint and finally used a transducer to be able to visualize the concepts belonging to the various annotations. We used the Snowball stemmer, a flexible gazetteer in combination with the OFAI gazetteer and the JAPE transducer. The fingerprints of each concept were provided to GATE in lists of all corresponding terms and phrases, also in stemmed version. In the next section we will describe how we processed these annotations with Recommender techniques discussed earlier to arrive at the final presentation of the verdict to the user.

5.3 Results

With the annotation of the terms from the fingerprints, we can now determine which paragraphs are relevant for which legal concepts. We used Keyword Style Explanation and designed a number of rules that state how many times a term, or multiple terms from the same fingerprint, must occur in a paragraph to deem that paragraph relevant for the particular concept that corresponds with that fingerprint. When a paragraph is relevant, we highlight it entirely (so the highlighting of the separate terms disappears) and provide the paragraph with a comment that explains the legal concept for which the paragraph is relevant. All legal concepts found in the verdict (corresponding to those extracted from the user case in another part of the program) are in general wording explained at the top of the verdict.

Besides this Keyword Style Explanation, we also used Influence Style Explanation. Certain terms or concepts were used to link the verdict to the user case. If for example the verdict was about a 'traffic accident', then the user would be notified whether this is a similarity or difference with respect to their case. This linking was done for multiple concepts in order to help the user apply certain aspects from the verdict to the user case.

5.4 User satisfaction research

As we were interested in the usefulness of this representation technique for court decisions a small scale user satisfaction research was conducted. The research group consisted of 21 participants and was divided into three groups of seven. Each group received a fictitious, but realistic, description of a case, a general explanation of the research, 4 verdicts and three different types of questions. The difference between the groups was the extra information given with the court decisions. Group 1 just received the verdicts, without any explanation. For Group 2 the court decisions were processed according to the Keyword Style, as explained in the previous section. Group 3 got the verdicts processed with Keyword Style *and* Influence Style Explanation.

Three different types of questions were formulated. The first category consisted of 'subjective' questions: propositions with an answering scale from 1 (I don't agree at all) to 5 (I agree completely). These were designed to measure the confidence the users have in the program and extent to which they feel the program is useful to obtain information about their legal position. The second category of 'objective' questions are in exam style. Those questions were designed to test the knowledge of the user about the provided case, the content of the legal concepts, and information about their legal position based on what they learned from the presented court decisions. The third and last category of questions had an open character in which the users could express what they liked about the program, what they missed and anything else they wanted to share. Some personal information was obtained to account for differences in age, education and legal knowledge.

Our overall hypothesis was that the groups would perform in increasing order. We hypothesized that group 1 would have the lowest scores for the subjective questions (meaning the highest confidence in and satisfaction with the program) and perform worst on the objective questions compared to the other groups. For group 2 these scores would improve, while group 3 would perform best on the subjective questions as well as on the objective questions. This hypothesis is based on the expectation that the extra information provided to group 2 and group 3 will contribute to an improved understanding of the presented court decisions relevant to gain more information about their legal position. The extra information provided to group 2 and 3 can help to enhance confidence in information provided by an online information system and also improve knowledge about their legal position. We hypothesize that the participants of group 1 need more time to complete the whole survey, since they will have to read the verdict on their own to find out what is relevant, whereas the other groups have the relevant paragraphs highlighted already. Out of the 21 surveys sent, we got 15 back; 5 in each group coincidentally.

Respondent	1	2	3	4	5	Average
Group 1	100 min	55 min	60 min	100 min	50 min	60,8 min
Group 2	60 min	40 min	55 min	45 min	60 min	43,3 min
Group 3	30 min	35 min	20 min	40 min	35 min	26,7 min

Table 7. Time needed to complete the survey

Question	1	2	3	4	5	6	Average
Group 1	2,4	2,8	2,6	3	3,8	4	2,9
Group 2	3	3,2	3	3,8	3,2	3,4	3,3
Group 3	3,4	3,8	3,4	4	3,8	3,6	3,7

Table 8. Average scores on the subject questions (scale 1-5)

In relation to the objective questions, answers were given in free text, which makes it impossible to analyze them with average numbers. However, interesting differences between the groups were observed. None of the respondents in group 1 mentioned the legal concepts ‘damage’, ‘causality’ and ‘an act or omission violating unwritten law pertaining to proper social conduct’, where most of those in group 2 and 3 did. Further, all respondents believed that a judge would grant the victim full compensation of the medical expenses for his foot. The majority of those in group 1 and 2 believed that the judge would not grant expenses made because of the depression. Reason given for this belief was that the victim had had depressions before, so the causal relationship could not be established in their eyes. In group 3 there were remarkably more respondents believing that the depression-related expenses *would* be granted. Answers to the question whether a compensation for not being able to play sports anymore would be granted were rather varying. Main reason for this was the need for more detailed information about sports history of the victim and alternative career prospects. Almost none of the respondents believe that a judge will grant all claimed damages: from the reactions it seems they just assume that a judge will never give you exactly what you ask for. Finally, group 3 is more reluctant to accept the offer in respect of a settlement than the other groups (4 of the 6 would not accept the offer, whereas in groups 1 and 2 only 2 of the 6 would not accept the offer).

The responses to the open questions might even have been the most useful for our research, the participants considered the task very difficult. However, apart from group 1, the average score was above ‘neutral’ towards the positive side of the scale. This indicates that they *did* learn something from the program (as could also be seen with the open questions), although they thought it was too difficult for them.

⁴ This question was about the extra information provided. Group 1 did not get any extra information, hence this question wasn’t relevant to that group.

Taking all the results together, we think we can be cautiously optimistic. The participants of group 3 were positive about their gained understanding of their case, and most of them did answer the objective questions in the way we envisioned beforehand. However, the verdicts are still very hard to read because of the legal jargon.

6 Future work

In our future work we will concentrate upon stage 1 and 2 of the system as described in section 1. We will collect case descriptions entered by laymen to analyse the terminology they use to describe legal liability cases. We already launched a website, staiknijnrecht.nl (freely translated: Am I legally right?), and will analyse the input we collect from this site. This will help us in developing a layman ontology. Right now we are beginning to develop the legal ontology, based on the analysis of the legal domain already undertaken, and the search concepts as described in section 3. This legal ontology is used to index case law.

In the end both ontologies are mapped to enable the retrieval of case law based upon a case description given by the laymen in his own wording. Only then we will know how successful the combination of the two parts of the project described in this contribution, viz. retrieval of case law and presenting the results, turns out. This will not be an simple enterprise, but the insights we gained so far makes us feel confident towards the future.

References

1. T.J.M. Bench-Capon and M.J. Sergot. Towards a rulebased presentation of open texture in law. In Walter, C. (ed.) *Computer power and legal language*: 39-61, New York, Qorum Books, 1988.
2. Bilgic, M. (2004). Explanation for Recommender Systems: Satisfaction vs. Promotion. Computer Sciences Austin, University of Texas. Undergraduate Honors: 27.
3. Bing, J. (1987). Designing text retrieval systems for "conceptual searching". International Conference on Artificial Intelligence and Law, Boston.
4. O'Donovan, J. and B. Smyth (2005). Trust in recommender systems. *International Conference on Intelligent User Interfaces*, San Diego, ACM Press.
5. Fabri, M. and F. Contini (eds.)(2001), *Justice and technology in Europe: How ICT is changing the judicial business*, The Hague, Kluwer Law International, 2001.
6. Christiaan Fluit, Frank van Harmelen, Marta Sabou Ontology-based Information Visualization: Towards Semantic Web Applications. In *Visualising the Semantic Web* (2nd edition), 2005, Springer Verlag
7. M.C.A. Klein, W. van Steenbergen, E.M. Uijtttenbroek, A.R. Lodder, F. van Harmelen (2006), Thesaurus-based retrieval of case-law. *Proceedings JURIX 2006*, p. 61-70.
8. A.R. Lodder, 'Law, Logic, Rhetoric: a Procedural Model of Legal Argumentation' (chapter 26), in: S. Rahman & J. Symons (eds.), *Logic, Epistemology, and the Unity of Science* (Logic, Epistemology, and the Unity of Science Series Vol. 1), Kluwer Academic Publishers, 2004.

9. Lodder, A.R., A. Oskamp & A.H.J Schmidt (eds.)(2001), *IT support of the Judiciary in Europe (ITeR deel 43)*, Den Haag: SDU 2001.
10. McSherry, D. (2005). "Explanation in Recommender Systems." *Artificial Intelligence Review* 24(2): 179-197
11. E.M. van Mulligen, C. van der Eijk, J.A. Kors, B.J. Schijvenaars, and B. Mons. Research for research: tools for knowledge discovery and visualization. In *Proceedings of the AMIA Symposium*, pages 835–839, 2002.
12. Oskamp, A., A.R. Lodder & M. Apistola (eds.)(2004), *IT support of the judiciary in Australia, Singapore, Venezuela, Norway, The Netherlands and Italy*, Cambridge University Press, TMC Asser Press (IT & Law series no. 4).
13. Gerard Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989
14. Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA, 1987.
15. Andrew Stranieri and John Zeleznikow. Knowledge Discovery from Legal Databases, Law and Philosophy Library, volume 69, Springer, 2005.
16. Stuckenschmidt, H., van Harmelen, F., de Waard, A., Scerri, T., Bhogal, R., van Buel, J., Crowlesmith, I., Fluit, Ch., Kampman, A., Broekstra, J., van Mulligen, E. Exploring Large Document Repositories with RDF Technology: The DOPE Project, *IEEE Intelligent Expert*, Vol. 19, No. 3, pp. 34-40.
17. Elisabeth M. Uijtttenbroek, Michel C.A. Klein, Arno R. Lodder, Frank van Harmelen & Paul Huygen, Semantic Case Law Retrieval – Findings and Challenges, *Proceedings SW4Law workshop 2007*.
18. Gwen R. Wildeboer, Michel C.A. Klein & Elisabeth M. Uijtttenbroek (2007), Explaining the Relevance of Court Decisions to Laymen, A.R. Lodder & L. Mommers (eds.) *Proceedings of JURIX 2007*, Amsterdam, Berlin, etc.: IOS Press, p. 129-138.