

# Concept Extraction from Legal Cases: The Use of a Statistic of Coincidence

Marie-Francine Moens  
Interdisciplinary Centre for Law & IT  
Katholieke Universiteit Leuven, Belgium  
Tiensestraat 41  
B-3000 Leuven, Belgium  
marie-  
france.moens@law.kuleuven.ac.be

Roxana Angheluta  
Interdisciplinary Centre for Law & IT  
Katholieke Universiteit Leuven, Belgium  
Tiensestraat 41  
B-3000 Leuven, Belgium  
roxana.angheluta@law.kuleuven.ac.be

## ABSTRACT

Effective retrieval of court decisions is important. Automatically identifying legal concepts in the decision texts would be very helpful. In this paper we investigate how a statistics for hypothesis testing, i.e., the likelihood ratio, can help in this task. We describe how this statistic can be used for detecting important multi-term phrases in the case texts, how it can be used to find correlated terms, and how it is a means for feature or topic signature selection in automated case categorization. The technology has been tested upon more than 600 US cases.

## Keywords

Conceptual information retrieval, concept extraction, ontology building.

## 1. INTRODUCTION

We are confronted with a large number of court decisions, which are stored in a database. Their effective retrieval is a fundamental challenge for legal information management. Besides some fixed elements, the most important content of the decisions is found in the natural language texts of the decisions. Current commercial retrieval systems either rely upon a manual indexing of the case texts or upon a full text search (i.e., every word in the text acts as a search key). The disadvantage of the former is the tremendous cost, which is a problem that only aggravates with the current logarithmic growth of the number of cases. The disadvantage of the latter is the lack of reliable retrieval results. The occurrence of a word or phrase in a text is no guarantee for the text's relevance to the search request. This is a fundamental and difficult issue that is especially apparent when retrieving legal cases. The success of future information systems is largely dependent on a good automated content analysis

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*Proceedings of the Ninth International Conference of Artificial Intelligence and Law June 24-28, 2003, Edinburgh, UK*

Copyright 2003 ACM 1-58113-000-0/00/0000 ...\$5.00.

of the unstructured information found in the natural language texts of the cases and on the automatic selection or attribution of conceptual index terms.

One facet of the content analysis regards the automated recognition of concepts in the case texts. The importance of concepts for effective case retrieval has been largely acknowledged (e.g., [2], [8], [11], [19], [22]). McCarty in 1984 [14] stated the need for a parsing of the decisions and automated assignment of legal concepts. Concept extraction involves the identification of concept-referring terms and phrases from the text and - if possible - their generalization into more abstract concepts. The concepts will then be used as index terms in retrieval [18]. In addition, concepts and their linguistic appearances in texts can constitute valuable ontological knowledge that otherwise would have to be acquired manually [cf. [4]].

In this paper we investigate how a statistic of surprise and coincidence can help in the detection of concepts in cases. The likelihood ratio [9] is an approach for hypothesis testing. It is a number that tells us how much more likely one hypothesis is than the other. We use hypothesis testing for three different tasks that assist in concept identification. The first two tasks refer to the identification of collocations in texts, i.e., detecting words that co-occur more often than by chance. More specifically, in the first task the likelihood ratio is used to detect phrases consisting of two or more words that have a specific (possibly legal) meaning due to their combination. Secondly, the likelihood ratio is used to detect words that often co-occur in the same context where context is defined as the same sentence, passage or text. In this way, we might detect terms that are semantically related. A third task that is considered in this paper is using the likelihood ratio as a feature selection metric for the automatic categorization of case texts with concepts. Equally, this task also allows detecting so-called topic signatures, i.e., terms that signal a certain topic in a decision.

The paper is organized as follows. We first describe the computation of the likelihood ratio and how it is valuable for hypothesis testing. In the next part we describe how the likelihood ratio is used in the three tasks described above. The corpus of cases used in the experiments is represented shortly. Then follows a discussion of the results and a section on related research.

## 2. THE LIKELIHOOD RATIO FOR A BINOMIAL DISTRIBUTION

### 2.1 Definition

The way we use the likelihood ratio is by testing the hypothesis whether two terms in a corpus occur independently [9]. If this hypothesis can be rejected we have a strong indication that the occurrence of the terms is dependent and the terms are correlated. Depending upon the task (see below) these terms then occur as consequent words that form a phrase, as correlated words that often occur in the same sentence, passage or text, or as a term that often co-occurs with a certain category term.

Given two words  $w_1$  and  $w_2$  we test the following hypothesis:

$$H_1: P(w_2|w_1) = p_1 = p_2 = P(w_2|!w_1)$$

**accept**  $H_1$ :  $w_1$  and  $w_2$  occur independently

**reject**  $H_1$ : we have a strong indication that  $w_1$  and  $w_2$  are correlated.

A binomial distribution results when one has a series of trials with only two outcomes (i.e., Bernoulli trials), each trial being independent from all the others. Repeatedly tossing a (possibly unfair) coin is a prototypical example. The probability  $P_k$  that in  $n$  independent trials an event  $A$  occurs precisely  $k$  times is given by the following probability function, where  $p$  is the probability of  $A$  in a single trial:

$$P_k = \binom{n}{k} p^k (1-p)^{n-k}$$

We assume that the following two events are distributed according to a binomial distribution:

1. observing  $w_1$  and  $w_2$  occurring together ( $c_{12}$  times) when we observe  $w_1$  occurring  $c_1$  times has probability  $p_1$
2. observing  $w_2$  occurring without  $w_1$  ( $c_2 - c_{12}$  times) when we observe the other words except  $w_1$  ( $N - c_1$  times where  $N$  = total number of events) has probability  $p_2$ .

We can build the likelihood function for this model:

$$H(p_1, p_2; c_{12}, c_1, c_2 - c_{12}, N - c_1) =$$

$$\binom{c_1}{c_{12}} p_1^{c_{12}} (1-p_1)^{c_1 - c_{12}} \binom{N - c_1}{c_2 - c_{12}} p_2^{c_2 - c_{12}} (1-p_2)^{N - c_1 - c_2 + c_{12}}$$

We can define a likelihood ratio  $\lambda$  as the maximum value of the likelihood function over the subspace  $\Omega_0$  (where  $p_1 = p_2$ ) to the maximum value of the likelihood over the entire parameter space  $\Omega$  (set of all possible values for  $p_1$  and  $p_2$ ):

$$\lambda = \frac{\max_{p_1, p_2 \in \Omega_0} H(p_1, p_2; c_{12}, c_1, c_2 - c_{12}, N - c_1)}{\max_{p_1, p_2 \in \Omega} H(p_1, p_2; c_{12}, c_1, c_2 - c_{12}, N - c_1)}$$

$$\lambda = \frac{\max_p H(p, p; c_{12}, c_1, c_2 - c_{12}, N - c_1)}{\max_{p_1, p_2} H(p_1, p_2; c_{12}, c_1, c_2 - c_{12}, N - c_1)}$$

The maxima of the likelihood functions are achieved for:

$$p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1} \quad p = \frac{c_2}{N}$$

This reduces the ratio to:

$$\lambda = \frac{L(p, c_{12}, c_1) L(p, c_2 - c_{12}, N - c_1)}{L(p_1, c_{12}, c_1) L(p_2, c_2 - c_{12}, N - c_1)}$$

Where:

$$L(p, k, n) = p^k (1-p)^{n-k}$$

Taking the logarithm of the likelihood ratio gives:

$$\begin{aligned} -2 \log \lambda &= 2[\log L(p_1, c_{12}, c_1) + \log L(p_2, c_2 - c_{12}, N - c_1) \\ &\quad - \log L(p, c_{12}, c_1) - \log L(p, c_2 - c_{12}, N - c_1)] \end{aligned}$$

= value that is asymptotically chi-square distributed when  $H_1$  is true. By selecting a confidence level from the chi-square distribution table, the obtained likelihood ratio value allows us to accept or to reject the hypothesis.

In our experiments we used the likelihood ratio for detecting useful term correlations which are discussed below. The approach is equally suitable for detecting many other types of associations (including citation correlations) in the case texts.

To make the computations as efficient as possible, an inverted file index of the selected words in the texts is first constructed. The index contains a list of the corpus words and for each term the position, sentence and/or the document identification code in which the term occurs. These data are then used for the likelihood computation. The computational complexity of building the inverted file is of the order of  $O(n)$  where  $n$  is the number of words selected. For computing the pairwise correlations of  $n$  terms is of the order  $O(n \times n/2)$  and by following a strict order in their computations, access to the right entry in the inverted index is immediate in order to extract the values needed for the likelihood computation.

### 2.2 Detection of conceptual phrases

The detection of conceptual phrases composed of two or more terms is very useful. It allows learning a vocabulary of legal conceptual terms from text corpora. Even, if we have access to ontological knowledge containing legal conceptual terms, new concepts may always come into existence or old ones might lose their value. Currently in academic disciplines such as computer and information science, and in diverse industries and fields - including the law field - there is a large interest in ontology applications and design [10]. An ontology in this context is defined as a "formal explicit specification of a shared conceptualization", where a conceptualization refers to an abstract model of how people think about things in the world or in a specific domain, and an explicit specification means that the concepts and relationships of the abstract model are given explicit terms and definitions. The conceptual phrases that are extracted from the decisions can be used - possibly after a manual screening - in ontology building.

The hypothesis is tested whether the co-occurrence of two successive words in the text are independent. If this hypothesis is rejected, we have a strong indication that the two

words form a collocational phrase. The idea is to expand the collocation to three-word components, when each set of two composing components form a collocational phrase.

We look in our experiments for conceptual patterns of two or three terms consisting of two or three consequent nouns, two or three nouns separated by prepositions, one or two nouns preceded by an adjective and two nouns separated by a preposition one of them being preceded by an adjective. The nouns can also be proper names. The word classes or part-of-speech (POS) classes of the words are detected with the LTCHUNK software developed by Andrei Mikheev of the University of Edinburgh, UK.

### 2.3 Detection of correlated terms

The detection of correlated terms is useful in many information retrieval tasks (e.g., for query expansion, automatic thesaurus construction, for providing contextual patterns in word sense disambiguation tasks, summarization and ontology building [1], p. 123 ff., [17]).

We rely here on the assumption that semantically related terms tend to co-occur in texts. The context for co-occurrence can be defined as the same sentence, the same paragraph, the same text, or a window of a fixed number of preceding and following words in the texts. In our experiments, the hypothesis is tested whether the co-occurrence of two words in the sentences of the texts are independent. If this hypothesis is rejected, we have a strong indication that the two words are correlated. This allows building for each word a set of related terms.

The words that we use in the experiments are nouns. They are detected with the LTCHUNK software mentioned above.

In this experiment we restrict ourselves to simple term occurrences, but many kinds of hypotheses of term independence can be tested that are useful for ontology building (e.g., the detection of important head modifier relationships which could be used for detecting subsumption relationships between entities).

### 2.4 Feature and topic signature selection

There is a lot of interest for techniques that learn the linguistic manifestations of a concept in a text. The patterns that are acquired through these algorithms can then be used for predicting concepts corresponding to words or phrases in previously unseen texts. Usually, these patterns consist of lexical items (words or phrases); sometimes the algorithm also takes into account the syntactic word classes or the syntactic (e.g., subject, object) or semantic roles (e.g., 'speaker', 'addressee') of the terms. Concept learning for legal cases is important if one wants to categorize them according to the broad domains to which they belong (e.g., 'taxation', 'insurance', 'bankruptcy', 'real property') or according to the legal issues that are discussed (e.g., 'master and servants', 'negligence') or if one wants to classify facts into factors (e.g., 'duty of reasonable medical care').

Most of the existing techniques for concept learning involve algorithms of supervised learning. Example texts are manually annotated with semantic tags and these are used to train a classifier. The following pattern classification techniques are often used: discrimination techniques (e.g., support vector machines); Bayesian classifiers (e.g., naive Bayes algorithm); and the induction of decision rules and trees (e.g., the C4.5 algorithm).

The main difficulty of training text classifiers on legal

cases is that there are a large number of features - e.g., the words in the text - that are entirely unrelated to the desired classification scheme, and that often there are only very few training examples for each category. This problem is particularly striking when one considers the results of classifying the entire text of a case based upon all its words except for stopwords [5], [21]. Learning classification patterns from case passages that are tagged with factor categories [6] improves the results, but such an approach requires the passages to be manually annotated and this quickly becomes a huge task, especially when a large number of factors is to be considered. Anyway, even promising classification techniques will have to cope with the fact that there are probably only a limited number of examples to train from. When confronted with a large number of features and few training examples, an initial feature selection that eliminates noisy and irrelevant features before training is useful [3], p.7 ff.

The likelihood ratio can be used to identify the correlation between a word and the category term. The approach is useful as a first feature selection step when training a text classifier. Besides, the terms that were detected as being correlated with a concept or topic can function as topic signatures. Topic signatures signal the presence of a complex concept, i.e. a concept that consists of several related components in fixed relationships [12]. A restaurant visit, for example, involves at least the components menu, eat, pay and possibly waiter. Topic signatures are very useful for the automatic construction of ontological knowledge (for instance, for identifying the terms that signal a certain concept in a text).

The hypothesis is tested whether the occurrence of a word in the texts and the category term are independent. If this hypothesis is rejected, we have a strong indication that the word and the category term are associated. Consequently, the word is a good feature to use in the training of the text classifier or is a good topic signature.

In our experiment we use the nouns of the case texts (detected with the LTPOS tagger). The category codes are Westlaw code numbers attached to the case. In the three experiments we reject the hypothesis of independence when it has only a confidence of 0.001 as found in the chi-square distribution table with one degree of freedom.

## 3. DISCUSSION OF THE RESULTS

### 3.1 Training corpus

Our training corpus consists of 634 US decisions of the Supreme Court and the Court of Appeals, and appeal cases of the Superior Court. The decisions are preceded by headnotes provided by Westlaw. They cite the key number and categories that are manually assigned to the decisions. These categories belong to a fixed classification scheme. The assignment of each category is followed by a small explanation. For detecting conceptual phrases and correlated terms, we did not use the headnotes.

The texts of the decisions in Text Only format usually range from about 10 KB to 350 KB in size.

The cases first cite some structural information, such as the parties, the dates of the different procedures, the previous history of the case (account of original case and appeals) and name of the court. Then follow the description of the facts, and the opinion of the parties and of the judge. The cases end with a disposition or a conclusion. The cases cite a

**Table 1: Examples of conceptual phrases**

Conceptual phrases	
adult court	Appellate Department
adversary proceeding	appellate rights
anterior cervical fusion	Bachelor of Science
antisocial adolescents	Bankruptcy Code
appellate court	Bankruptcy Court
appellate courts	Barbie Dolls

complete account of the claims of the plaintiff, affirmations, reversals, dissents, evidence, conclusions, opinion, reference to statute law and references to other cases. Sometimes an explanation of a vague term is given. There is often a lot of detail: examples are given to illustrate concepts. Large case texts have sections that bear titles.

Facts regarding the defendant and plaintiff are often explicitly mentioned: they are introduced by terms that refer to the defendant or plaintiff. The facts might contain very unusual details, like "a green cloth hanging over the license plate". The facts and the opinion can be of any situation in society. So, the vocabulary is very diverse, ranging, for instance, from house robberies to the technical details of the procedures of DNA tests. Often you have the facts followed by the point of views of the different parties, which are elaborated in a real discussion between the parties. Sometimes the facts are grouped under a heading. The factual circumstances are often told in a narrative style. They are then described by short sentences that narrate a sequence of events. There are also accounts of situations regarding the jury, for instance, when the jury viewed the crime scenes. Sometimes there is a piece of opinion on a certain topic. The account of the opinion can be quite long. The cases are also characterized by a large number of literal quotations from texts of other cases. In addition, reference to statute law is very important in these cases.

### 3.2 Detection of conceptual phrases

The list of extracted conceptual phrases (examples are in table 1) show many useful concepts including proper name phrases and compound terms that correspond to some conventional concept. They refer to what is commonly referred to as a collocation in the linguistic literature and are often characterized by limited compositionality, i.e., the meaning of the expression cannot be completely predicted by the meaning of the parts because there is an element of meaning added to the combination.

### 3.3 Detection of correlated terms

Most of the term associations found are valuable (examples are in table 2). Many terms are only associated with only one other term of the corpus. This is probably due to the fact that our training corpus is not very large. In our experiments we investigated the co-occurrence of nouns, however one could equally investigate the association of different syntactic word classes such as the association of a verb with nouns.

### 3.4 Feature and topic signature selection

We could find very interesting category features or topic signatures (examples are in table 3). Most of them seem very valuable as privileged features for training a text classifier.

An intrinsic evaluation, i.e., comparing the results with

**Table 2: Examples of term correlations**

Term	Correlated terms
claw	beak, spur, teeth
disputes	explanations, investigations, invocation, arson
instrumentalities	knives, razors, pocket

**Table 3: Examples of features**

Category	Topic signatures
Carrier	car, derailment, passenger, passengers,railroad, trains
Criminal Law	charge, prison, process, plea, statute
Seamen	block, blocks, deck, hospital, rope, shipmate, shipowner, vessel

concepts, term correlations and terms that signal legal concepts that are manually drafted, requires such evaluation data, which we do not have at the moment. We plan in the near future to test the results extrinsically, i.e., testing whether the technologies improve the results of some other task (i.e., using term correlations in defining important concepts or using the technology for feature selection before training a text classifier).

## 4. RELATED RESEARCH

The detection of term relationships has been studied in information retrieval research for the automatic construction of thesauri [1], p. 123 ff. Different statistics can be used for this task including term clustering methods, the point-wise mutual information statistic [13], p. 495 ff., chi-square statistic and latent semantic indexing [7]. When training text classifiers, the following statistics are commonly used for feature selection: information gain [15], p. 57 ff., mutual information statistic [13], p. 66 ff. and the chi-square statistic [16], p. 213 ff. In the legal field neural networks have been used to detect term relationships [20]. Compared to the techniques of term clustering, latent semantic indexing and neural networks, the computation of the likelihood ratio is computationally not very complex. It is thought that the likelihood ratio is in general more appropriate than the chi-square test for discovering collocational phrases. We lack however studies that test and compare the success rate of the different techniques for detecting legal concepts and their relationships. One advantage of the likelihood ratio is that it has a clear intuitive interpretation, because the ratio gives you how much more likely one hypothesis is than the other. In addition, the metric especially appropriate for sparse data processing [9].

In the legal domain the automatic extraction of concepts from legal texts and the detection of a set of terms that signal concepts has, for instance, been studied by [5], [6], [21]. Because of its great importance in legal information retrieval and ontology building, additional research is certainly needed to make advances in the current strategies, which was one of the goals of our experiments.

## 5. CONCLUSION

Conceptual retrieval has proven its usefulness in information retrieval. For the retrieval of court decisions, the concepts refer to the facts, factors and legal issues discussed in a case. In this paper we have investigated how the likelihood ratio statistic applied on part-of-speech tagged case texts can assist in the automatic identification of concepts. The statistic was used both in supervised and unsupervised learning tasks. From a corpus of cases candidate conceptual terms, semantically related terms, and feature terms or signatures that signal a certain concept are automatically learned.

## 6. ACKNOWLEDGMENTS

Our thanks to the IWT (Vlaams Instituut voor de bevordering van het Wetenschappelijk-Technologisch onderzoek in de industrie) for sponsoring this research (Grant Nr. ADV/000135/KUL). We thank Dr. John Barker for making available the texts of the court decisions. We are also grateful to Steven De Bruin, who helped with the implementation of a part of the software.

## 7. REFERENCES

- [1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, Harlow, UK, 1999.
- [2] J. Bing. Performance of legal text retrieval systems: The curse of boole. *Law Library Journal*, 79:187–202, 1987.
- [3] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, UK, 1995.
- [4] J. Breuker, A. Elhag, E. Petkov, and R. Winkels. T. Bench-Capon, A. Daskalopulu, and R. Winkels (Eds.), *Legal Knowledge and Information Systems*, chapter Ontologies for legal information serving and knowledge management, pages 73–82. IOS Press, Amsterdam, 2002.
- [5] S. Brninghaus and K. Ashley. Finding factors: Learning to classify case opinions under abstract fact categories. In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law*, pages 123–131. ACM, New York, 1997.
- [6] S. Brninghaus and K. Ashley. Toward adding knowledge to learning algorithms for indexing legal cases. In *Proceedings of the Seventh International Conference on Artificial Intelligence and Law*, pages 7–17. ACM, New York, 1999.
- [7] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41 (6):391–407, 1990.
- [8] J. Dick. Representation of legal text for conceptual retrieval. In *Proceedings of the Second International Conference on Artificial Intelligence and Law*, pages 244–253. ACM, New York, 1989.
- [9] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74, 1993.
- [10] M. Gruninger and J. Lee. Ontology applications and design. *Communications of the ACM*, 45 (2):39–41, 2002.
- [11] C. Hafner. An information retrieval system based on a computer model of legal knowledge. *UMI Research Press, Ann Arbor, MI*, 1981.
- [12] C.-Y. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the COLING Conference*, 2000. <http://www.isi.edu/natural-language/people/hovy/publications.html>.
- [13] C. Manning and H. Schtze. *Foundations of Statistical Natural Language Processing*. MIT Press Cambridge, MA, 1999.
- [14] L. McCarty. Intelligent legal information systems: problems and prospects. In C. Campbell (Ed.), *Data Processing and the Law*. Sweet & Maxwell, London, pages 125–151, 1984.
- [15] T. Mitchell. *Machine Learning*. McGraw-Hill, Boston, MA, 1997.
- [16] M.-F. Moens. Automatic indexing and abstracting of document texts. (*The Kluwer International Series on Information Retrieval 6*). *Kluwer Academic Publishers, Boston, MA.*, 2000.
- [17] M.-F. Moens, R. Angheluta, and R. De Busser. In W. Abramowicz (Ed.), *Knowledge Based Information Retrieval and Filtering*, chapter Summarization of texts found on the World Wide Web. Kluwer Academic Publishers, Boston (in press), 2003.
- [18] M.-F. Moens and R. De Busser. First steps in building a model for the retrieval of court decisions. *International Journal of Human-Computer Studies*, 57, 5:429–446, 2002.
- [19] E. Rissland, S. D., and M. Friedman. Bankxxx: Supporting legal arguments through heuristic retrieval. *Artificial Intelligence and Law*, 4 (1):1–71, 1996.
- [20] D. Rose and R. Belew. A connectionist and symbolic hybrid for improving legal research. *International Journal of Man-Machine Studies*, 35 (1):1–33, 1991.
- [21] P. Thompson. Automatic categorization of case law. In *Proceedings of the 8th International Conference on Artificial Intelligence and Law*, pages 73–82. ACM, New York, 2001.
- [22] R. Winkels, D. Bosscher, A. Boer, and R. Hoekstra. Extended conceptual retrieval. In *Legal Knowledge and Information Systems: Jurix 2000: The Thirteenth Annual Conference*, pages 85–97. IOS Press, Amsterdam, 2000.