Semi-Automatic Construction of Skeleton Concept Maps from Case Judgments
Boer, A.W.F.; Sijtsma, B.

Published in:
NAiL 2014: 2nd international workshop on "Network Analysis in Law": Wednesday december 10th 2014: in conjunction with JURIX 2014: 27th International Conference on Legal Knowledge and Information Systems, Krakow, Poland

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
2ND INTERNATIONAL WORKSHOP ON “NETWORK ANALYSIS IN LAW”

WEDNESDAY DECEMBER 10TH 2014
IN CONJUNCTION WITH JURIX 2014:
27TH INTERNATIONAL CONFERENCE ON LEGAL KNOWLEDGE AND INFORMATION SYSTEMS, KRAKOW, POLAND

RADBOUD WINKELS
NICOLA LETTIERI
Semi-Automatic Construction of Skeleton Concept Maps from Case Judgments

Alexander Boer and Bas Sijtsma
Leibniz Center for Law, University of Amsterdam, The Netherlands
e-mail: A.W.F.Boer@uva.nl

Abstract. This paper proposes an approach to generating Skeleton Conceptual Maps (SCM) semi automatically from legal case documents provided by the United Kingdom's Supreme Court. SCM are incomplete knowledge representations for the purpose of scaffolding learning. The proposed system intends to provide students with a tool to pre-process text and to extract knowledge from documents in a time saving manner. A combination of natural language processing methods and proposition extraction algorithms are used to generate the output. Conclusion is that improvements are necessary to provide results that adequately support students.

1. Introduction

Conceptual maps (or concept maps) are a graphical representation of knowledge, containing a hierarchical structure of core concepts and concept relationships. The conceptual map can serve as a scaffold to structure prior and newly attained knowledge into an existing cognitive structure [14]. Since its initial formulation in 1972, much research has been dedicated towards its use as a cognitive tool in education and as an analytical tool for domain experts.

Skeleton concept maps are incomplete representations of information prepared by teachers or domain experts, containing some of the building blocks required to form a conceptual map. Skeleton concept maps serve as a foundation to start a process of meaningful learning. The skeleton concept map has proven to be an effective tool in education, promoting collaborative learning and self-assessment ([13], [15], [7]). However, building skeleton conceptual maps requires considerable time and effort. Since 2001, researchers have therefore attempted to construct skeleton concept maps automatically in real time, each dealing with parts of the problem.

We tried to generate skeleton conceptual maps semi-automatically from legal case judgments provided by the United Kingdoms Supreme Court, using a state-of-the-art approach. These documents are very large. A random selection of the available documents showed that some contain up to 130 pages. Automated extraction of skeleton conceptual maps could therefore save a substantial amount of time, either in preparing education of law students, or as a preprocessing phase in legal knowledge engineering. To evaluate performance, the results for a single judgment are compared to a gold standard concept map made by an expert in section 5.
1.1. Relation to argument mapping and concept extraction from legal text

Argument mapping, more familiar in education in law, produces graphs that are syntactically similar. [5] summarizes research on argument maps, and concludes that argument maps can significantly improve critical thinking, especially amongst students with poor argument analysis skills. In law, notable improvements in student performance in legal case analysis have been found by [12]. Various attempts have been made to automatically generate argument maps from legal text [21].

Argument maps are however less suitable for expressing what the case judgment teaches about domain conceptualization, even though this function of judgments is undisputably important. See for instance the conclusions section of [12]. Argument maps contain proposition nodes rather than concept nodes. While regulatory text can be processed automatically to extract concept definitions (cf. for instance [20]), the concept map appears more suitable for exploring less explicit domain conceptualization in case judgments. The goal of this work was to determine whether tested and tried methods in other education environments produce an acceptable SCM from a complete, lengthy case judgment.

2. Concept maps

In 1972, Joseph Novak and his team of researchers at Cornell University were studying emergence of understanding of science concepts in children. Because of the trouble Novak and his colleagues had determining whether the children understood the taught concepts, they were searching for a tool that would facilitate explicit display of cognitive structure [14], and developed the Conceptual Map. Since then, the concept map has proven itself useful in many different applications. For instance: as a teaching and learning strategy tool; as a tool for gathering, diagnosing and modeling knowledge; and as an application to guide a collaboration process [7]. It has received wide attention in the academic world, and has proven its validity in many different contexts [10].

The concept map, as formalised by Novak and his team, is a structured diagram containing concepts connected by linking phrases. Its a hierarchical tree-like structure, which is often concentrated around a focus question. At the top the superordinate concept is displayed, and the meaning of a concept is in part determined by the concepts directly related to it [15]. Concepts are indicated by a label inside a box, often consisting of a noun phrase. Linking phrases are the arcs connecting associated concepts, in most cases consisting of a verb phrase. A linking phrase is inherently bi-directional: a concept linking from concept A to concept B with the linking phrase contains, also links concept B to concept A with the linking phrase contained by. Concepts connected by a linking phrase are referred to by [15] as propositions (see figure 1). The concept map proposition is syntactically equivalent to for instance the triple in RDF, and basic propositions in many other representation languages.

2.1. Constructing a concept map

In [15] Novak discusses how to actually construct a conceptual map. As has been explained, concept maps target a focus question. Next a “parking lot” of concepts related to
Figure 1. Breakdown of a proposition

Figure 2. An example of a skeleton map as defined by [13]

the focus question is made. This serves as a pre-processing step to get a clear idea of the domain. The concepts are then ordered by importance. Next the concepts are placed on the map one by one, while simultaneously identifying the links between them. Finally, when all links have been drawn, they are given names as to represent the relation between them.

[15] discuss providing a "skeleton concept map" to learners as to scaffold the learning process. A skeleton concept map is an incomplete representation of the domain, a concept map that is halfway through its construction. This approach attempts to provide a foundation for students to construct their own concept maps. One such skeleton map as suggested by [15] is displayed in figure 2. The use of skeleton concept maps in teaching is part of what Novak dubbed "the new model of education" [13]. It has proven to be an effective, but labour-intensive, tool for collaborative learning, increasing long-term meaningful learning amongst its users [13,1,7].

3. Research on generating concept maps

Various research groups have taken on the problem of generating conceptual maps in a (semi-)automatic way. Approaches differ in their purpose, the data source, the methods used, the resulting output, and in the evaluation of results.
3.1. Data Sources and Methods used in the generation of conceptual maps

The data sources used for the concept map construction in the reviewed research was in most cases dependent on the purpose of the research. For educational purposes, often data sources related to education were used. Broadly, we can identify structured and unstructured data sources. Unstructured text is considered to be regular text that is not pre-annotated either computationally or manually. The data sources in approaches using unstructured texts were text books of a specific educational domain [17]; texts by domain experts [9]; and non-specific unstructured text ([16,2]). Structured data sources were pre-annotated. Among this category were texts from a large corpus with semantic structures manually annotated [8] and statistical data with annotations of related concepts [18].

Processes involving linguistic analysis are most common. In some cases, external assets such as WordNet or NomBank are used to compliment the performance of the system by finding morphological variations of words ([16,9,17]). The approach is to build propositions from sentences and apply rules to filter results that are unlikely to make sense or add value. Statistical approaches either make use of a quantitative data source or quantify the contents of the data source first. In the quantitative data source, a relation is established between statistics and concepts, and further analyzed to extract regularities [18]. Quantification of the data source includes counting the appearances of concepts in text [2]. Although the results of the approach never represent a conceptual map as formalised by [15], one can consider the results to be a halfway construction of one.

The machine learning techniques that are applied involve requesting feedback from users to increase the performance of the system. These techniques however are not used in the automatic construction of conceptual maps alone, but used to complement performance of linguistic processes ([17,16]). The output of linguistic analysis combines with machine learning is of significantly better quality than the other approaches.

3.2. Results and evaluations

The results of the used methods can be categorized into three classes: a fully connected graph; a connected graph without named relations; or merely proposition triples. The highest quality output is achieved by a combination of linguistic and machine learning processes: it generates a fully connected and labeled graph. The results are not sorted hierarchically, however, as intended by [15]. [8] also claim to produce connected graphs, but they generate a lot of “small” graphs containing only a few connected concepts. [18,17] produce connected concepts without describing the relations between these concepts. Approaches producing proposition triples only are meant to provide information for readers to interpret as a text pre-processing step [9]. The fully connected graph is made manually after analysis of the propositions.

Generally, the researchers that performed an evaluation attempted to identify either the measure of “coverage” of their output on the initial data source, or rated the quality of the output on a certain scale. Authors differ in the way they measure coverage. For instance, [9] takes a set of pre-annotated documents, and measure the precision and recall only for the concepts that were extracted. Linking phrases or proposition structure are not taking into consideration. [17] on the other hand, take a hybrid approach by using a gold-standard concept map created by a domain expert, and then having human judges
rate both concept maps on a scale that represents coverage. [16] measure the precision score. [8] do not measure the coverage. Instead, they have domain experts rate the output on a three-point scale of comprehensibility. Finally, [18] use statistical measurements to rate the redundancy and circularity of the output.

The many different approaches to evaluating the output demonstrates the difficulty of rating concept maps properly. Indeed, [19] states that concept maps do not lend themselves to a single dominant method of evaluation.

### 3.3. Proposition extraction algorithm

Conceptual maps consist of a set of linked propositions, putting proposition extraction at the center of the problem. As has been briefly mentioned in the literature review above, natural language processing approaches make use of triple extraction algorithms. These algorithms consist of specific grammars, a rule-set that define a proposition. For instance, consider the following sentence: *Bas eats pizza on Sundays*. A very basic grammar may define a proposition as a sequence (noun - verb - noun). Running this rule-set over the example sentence, the proposition: *(Bas - eats - pizza)* would be extracted.

We adopted the more advanced algorithm of [6]. [6] criticize algorithms defined by others in the field of information extraction by stating that the resulting linking phrases are often incoherent and omit critical information. This is caused by algorithms that prioritize the extraction of noun phrases as concepts. The algorithm prioritizes and selects linking phrases based on a syntactic and a lexical constraint. A linking phrase should be either:

1. a verb;
2. a verb followed by nouns, adjectives or adverbs, but ending in a preposition; or
3. multiple adjacent sequences of the above.

However, the syntactic constraint can match linking phrases that are very long, and therefore become overly specific. To combat this, a lexical constraint is used. When the noun phrase to the left of the linking phrase is an anaphor, the algorithm attempts to perform anaphora resolution as in [16]. When a word such as “which” is found, it is skipped as a concept, and the noun phrase to the left of the anaphor is selected. [6] analyzed a corpus of 500 million sentences, comparing the algorithm to a set of state-of-the-art proposition extraction tools on its precision and recall scores, concluding that the algorithm outperformed the other tools substantially.

### 4. Method

For our experiments, legal documents provided by the United Kingdom’s Supreme Court were used [3]. The Supreme Court provides both the full length judgement, some containing up to 130 pages, and a press summary. A single document was used for the construction of a conceptual map, but we expect some form of generalization of performance across the corpus.

A linguistic analysis approach based on [6] was used for the extraction of conceptual maps. The availability of free high-quality natural language processing libraries makes this approach very accessible.
4.1. Input data

The proposed methodology is tested on a Judgement document from the United Kingdoms Supreme Court [3]. The document was randomly selected from the first page of the Supreme Court website. It describes the judgement made on the 21st of May 2014. It contains twenty five pages: the first four consist of an introduction of the case, followed by six pages of information related to the appeal. Then, three pages describe the decision of the Upper Tribunal. The remainder of the document describes the issues that the Supreme Court has to consider, and the arguments that led them to their final judgement.

Summarized, the case deals with the use of expert evidence in asylum appeals. In asylum appeals, a crucial issue is whether an appellant is honest about where they come from. The appellants claimed to come from a particular region of Somalia where they were at risk of persecution. However, based on linguistic evidence, the Secretary of State dismissed those claims, stating that their speaking was linked to Kenya instead. The linguistic evidence was in the form of linguistic analysis reports, provided by a Swedish commercial organisation called “Sprakab”. Sprakab’s methods involve listening to audio recordings of speech, followed by interviews with the speaker that are processed by anonymous linguists and analysts.

The Supreme Court was to consider five issues: if the immigration judges were to attribute weight to the reports generated by Sprakab; if the witnesses from the Sprakab organisations are to be granted anonymity; if there are any particular rules governing expert evidence offered by organisations instead of individuals; to what extent the evidence not in a form specified by Practice Directions can be accepted; and to what extent the Upper Tribunal can give guidance to the weight that is given to the expert evidence reports.

The appeal is unanimously dismissed by the Supreme Court. It decided that the Sprakab expert evidence reports, if the process is properly checked, were to be accepted. The weight given to these reports should be examined critically in any future cases. The expert witnesses anonymity is justified in this particular case, but it remains to be determined in any future cases.

4.2. Architecture of the system

The construction of a conceptual map will proceed in a series of steps, as shown in figure 3. The steps are explained below.

NLP parsing: sentence boundary detection, tokenization, part-of-speech tagging and chunking with the help of Apache OpenNLP.

Extract nouns: Noun phrases tagged in the previous step will be extracted and saved in a separate file. The consequence of this approach is that any concept not represented by a noun, or concepts incorrectly missed by the POS-tagger, will not be extracted.

Match morphological variations: Several variations of the same noun may be extracted. To deal with this, one can perform the task of stemming and lemmatization. The lexical database that is used for this step is WordNet [11].

Rank nouns: We generate conceptual maps from a single data file. Term frequency in the input data is therefore deemed appropriate for ranking.
Proposition extraction: Uses the algorithm proposed by [6]. [6] made their extraction algorithm fully available for others to use. Other authors omitted details related to the extraction algorithm or did not specify the algorithm at all.

Proposition filtering: Two filters will be applied to narrow down the list of available propositions. The first filter that is carried out is the repetition filter, as proposed by [17]. The second filter is a safety net in case the NLP methods fail. Whenever more than half of an item in the proposition is not regular text, it is discarded.

Select propositions containing top ranked nouns: A concept map often consists of approximately thirty different concepts [15]. Therefore, only propositions containing at least one of the thirty most appearing nouns identified in step 4 are be selected. The result of this step is a list of propositions in hierarchical order.

Combine propositions into skeleton concept map: Propositions are manually selected and combined to form the skeleton concept map. Items taken from the list of most important concepts generated in step four will be put in the concept “parking lot”, where they can be used by a learner to connect them to the rest of the graph.

4.3. The resulting output

Because the proposed method is capable of returning propositions with named relations, the output of the system is a set of propositions ordered by their importance. These will be combined into a skeleton concept map by hand.

4.4. Evaluating the output

A commonly used technique for evaluating various aspects of the concept map is using a gold-standard approach. To evaluate the results of the proposed approach, a method similar to those of [9] is used. A set of instructions to construct a conceptual map is provided to a domain expert. These instructions are taken directly from [15]. The expert is then given the same data source as the data used in this research, and will proceed to attempt to construct a concept map.

All propositions generated by the domain expert are extracted and compared to the propositions that were automatically constructed. Using the gold-standard propositions, a measure of precision and recall can be calculated. All distinct concepts identified by the domain expert are used as “correct” concepts. The precision score is calculated by the ra-
tio of correct concepts that were extracted compared to the total amount of concepts that were extracted. The recall score will be calculated by the ratio of the correctly extracted concepts compared to the total “correct” concepts identified by the domain expert. These two measures will serve as a way to rate the coverage of the used method on the input data.

A lawyer competent in the area was given a short introduction on the theoretical basis of conceptual maps. The resulting concept map can be found in figure 4. Note that all propositions in this map come from the press summary of the judgment. The lawyer expressed the opinion that judgements were not appropriate for conceptual maps, and, looking at the bottom of figure 4, apparently felt indeed constrained in not being able to link to propositions, as in an argument map.

5. Results

5.1. Evaluation of the language processing tasks

The extraction of concepts proceeds according to the natural language processing pipeline. From the POS tagged text, all nouns are extracted. This means that any errors made in the identification of nouns by the tagger will propagate into the steps that follow. To evaluate the performance of the tagger, the Press Summary document of the same judgement ([4]) is analyzed. Firstly, all nouns in the text were annotated by hand. Then, the NLP pipeline was run over the document and compared with the manually notated data.

Out of 960 tokens, the tagger identified 263 nouns, missing only 3 nouns by mistaking them for verbs or adjectives. It did not incorrectly tag any tokens as a noun that were in fact of a different part of speech.
5.2. Evaluation of concept ranking

To rank concepts, nouns were extracted from the full judgement text and lemmatized with the lexical database. All identical lemmas were summed to count the frequency of appearance. In total, 2930 nouns were extracted from the judgement. Out of these nouns, 674 distinct lemmas were identified. 518 of these occurred less than five times.

The conceptual map theory mentioned that a limit of approximately thirty concepts should be sufficient to create a concept map of a domain. Examining the thirty most frequently occurring nouns, it becomes evident that important concepts from the judgement are not included. The unaltered top thirty lemmas by frequency are used in the proposition selection for building the concept map.

5.3. Evaluation of the proposition extraction algorithm

The proposition extraction algorithm was run over the full judgement text and extracted a total of 501 propositions. The extracted propositions are put through a series of filters. Firstly, syntactically unsound and repetitive propositions were removed. Secondly, the concept ranking is applied to the remaining 473 propositions. All propositions that do not contain at least one of the top ranked lemmas in either one of its concepts is removed. This reduces the list of relevant propositions to 246.

Extracted propositions seem incoherent and incomplete for sentences that have a non-contiguous structure. One instance of a proposition that was affected is (Sprakab reports - should limit - themselves). The remaining sentence contained valuable information on what Sprakab reports should limit themselves to, but it is left out because of the way the algorithm deals with the sentence structure.

5.4. Gold-standard comparison

To evaluate the performance of the used method, the similarities between the gold-standard and the output of the used methods can be examined on three levels. The first level to examine is the concept level: were the concepts identified by the domain expert also identified by the methods used in this thesis, and were these concepts present in the extracted propositions? The second level is that of linking phrases between concepts: were the links identified by the domain expert also identified by the extraction methods used in this approach? The evaluation will consider if a linking phrase was extracted that has at least one of the two connected concepts in common. The third level is that of propositions: were any propositions identified by the domain expert also fully extracted by the methods used in this thesis?

The gold standard identified nine distinct concepts. In the top thirty list of lemmas, seven of the lemmas cover at least in part. The two concepts remaining that were not covered by the top thirty list, are covered by the top 43. Some form of leniency was applied in the scoring of the precision and recall measures. Even if the wording of the concepts is not identical, if the extracted proposition seemingly refers to the same thing, it is counted as an extracted concept. Obviously, the precision of proposition extraction is low (0.111) with only nine distinct concepts identified by the domain expert in a lengthy input text. The recall score of the proposition extraction algorithm with regards to concepts is high (0.889), which can be attributed to the low amount of “correct” concepts,
the high amount of propositions, and the leniency used in the process of comparison. Had no leniency been applied only four out of the nine concepts would have matched literally (0.444).

Out of nine distinct linking phrases, three links somewhat match on a conceptual level. The gold standard for instance identified the proposition: (Upper Tribunal - gives - general guidance). The proposition extraction algorithm generated: (another case - gave guidance on - the use of such reports). The algorithm does not perform very well on the second level of comparison with the gold standard, and not at all on the third level.

5.5. Generated skeleton concept map

Fortunately, this does not mean that the extracted propositions are useless. Comparing the generated propositions with the gold standard propositions, it’s possible to interpret references to the same concepts and links. An attempt was made to construct a skeleton concept map using the set of generated propositions and the list of highly ranked concepts.

The skeleton concept map displayed in 5 was constructed from a combination of generated propositions and the list of highest ranked concepts. Two propositions were slightly modified to be able to connect them together.

6. Conclusions

Although the skeleton concept map shown in figure 5 may potentially have some value, it could not have been created had the input data not already been analyzed. Therefore, we conclude that the used approach is not yet sufficient to reliably create skeleton concept maps in a time saving manner.

We believe that the golden standard evaluation is too demanding, however. Precision is negatively affected by 1) the expert rephrasing propositions, making it hard to find matches, and 2) the use of a concept map to reconstruct arguments by the expert, introducing modeling choices that the algorithm can, and should, not reliably replicate. For future evaluation of performance, a passive grading approach, in which an experienced teacher scores the SCM, and blindly compares with student-constructed maps, is preferable.
6.1. Suggestions for future research

To tackle the problem of too much information for the SCM, a better performance of proposition filtering will have to be accomplished. A tool appropriate to filter propositions is a likelihood ratio filter [17].

To ensure all essential information is covered by propositions, the proposition extraction algorithm will have to be reconsidered. The algorithm chosen in this thesis prioritizes the extraction of linking phrases. An additional algorithm could be added, which could prioritize the extraction of candidates for concept nodes.

Another step that can be taken to ensure the desired information is extracted is to add a priori “knowledge” to the system. Before the system runs, a user could take a glance at the title and introduction of the document, and select a set of terms deemed important. An algorithm could then prioritize the specified terms in constructing propositions.

References


