Semantic Edge Labeling over Legal Citation Graphs

Ali Sadeghian
University of Florida
Gainesville, FL
asadeghian@ufl.edu

Laksshman Sundaram
University of Florida
Gainesville, FL
sundaram@cise.ufl.edu

Daisy Zhe Wang
University of Florida
Gainesville, FL
daisyw@cise.ufl.edu

William F. Hamilton
University of Florida
Gainesville, FL
hamiltonw@law.ufl.edu

Karl Branting
MITRE Corporation
7525 Colshire Dr
McLean, VA 22102
lbranting@mitre.org

Craig Pfeifer
MITRE Corporation
7525 Colshire Dr
McLean, VA 22102
cpfeifer@mitre.org

ABSTRACT

Citations, as in when a certain statute is being cited in another statute, differ in meaning, and we aim to annotate each edge with a semantic label that expresses this meaning or purpose. Our efforts involve defining, annotating and automatically assigning each citation edge with a specific semantic label. In this paper we define a gold set of labels that cover a vast majority of citation types that appear in the United States Code (US Code) but still specific enough to meaningfully group each citation. We proposed a Linear-Chain CRF based model to extract the useful features needed to label each citation. The extracted features were then mapped to a vector space using a word embedding technique and we used clustering methods to group the citations to their corresponding labels. This paper analyzes the content and structure of the US Code, but most of the techniques used can be easily generalized to other legal documents. It is worth mentioning that during this process we also collected a human labeled data set of the US Code that can be very useful for future research.

Keywords

US Code; Legal citation graph; Automatic citation analysis; Conditional Random Fields; K-means clustering

1. INTRODUCTION

New regulations and laws in the United States are legislated or existing ones are evolved through a complex legal cycle. This system involves numerous organizations, parties, and individuals. Individual legal rules seldom exist in isolation, but instead typically occur as components of broader statutory, regulatory, and common-law frameworks consisting of numerous interconnected rules, regulations, and rulings. The complexity of these frameworks impedes comprehension and compliance by government agencies, businesses, and citizens and makes amending legislation laborious and error-prone for regulatory agencies and legislative drafters.

Citation networks are a promising recent approach to improving the intelligibility of complex rule frameworks. In a citation network, rules are represented by nodes and citations are represented by edges [16, 20]. Citation networks can often permit a complex regulatory framework to be comprehended at a glance. Techniques for automatically representing and displaying citation networks is an active area of research.

Computer assisted and automatic systems have been and are growing rapidly in every field. The legal domain is also no exception to this trend [11, 18, 7]. Specially there has been extensive research in designing programs and intelligent software that can address the challenging and expensive task of information extraction from general text. Information extraction is of special importance from a legal perspective since almost all the information in this domain is collected in natural human language. This techniques can be utilized to aid in the automation of creating and displaying meaningful citation networks.

An important aspect of citation-network use is that, generally, only a small subgraph is relevant for any particular application or task. Indeed, visualizations of entire citation networks are generally incomprehensible ”hairballs.”

The subgraph of a citation network relevant to a particular task depends both on the attributes of the nodes (i.e., rules) and edges (i.e., citations). For example, a subgraph relevant to public health emergencies would include both nodes defining the powers and duties of agents (e.g., doctors, epidemiologists, coroners) and citations indicating the relative authority among these agents. In general, the portion of a statutory framework relevant to given task consists of the subgraph induced by nodes and edges having a semantic relationship to the task.

While nodes relevant to a given task (e.g., UAV licensing) can typically be found using information-retrieval techniques, such as term-vector or topic similarity, identification of relevant edges, is much less well understood. Various researchers have proposed different taxonomies of edges in citation graphs [9, 13, 5], but there is not yet a consensus on the most useful set of edge types. Moreover, there has been little progress in automatically applying semantic labels to citations edges, which is essential for large-scale citation network visualization and analysis tools.

San Diego ’California USA
© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

This paper first reviews the related work in Section 2. Followed by precisely describing our research problem in Section 3 and the proposed automated system to tackle this problem in Section 4. In Section 5, we describe the data set used to evaluate our system as well as the proposed gold standard label set used for labeling the citation graph. And finally we conclude the paper in Section 6 by a summary of the results and a plan for future research on this study.

2. RELATED WORK

There has been various previous research projects addressing the detection, resolution and labeling of citations in the legal domain. But to our knowledge there has not been any prior work on a systematic approach to automatically detecting and labeling of cross references with a detailed semantic label set.

In [9] M. Hamdaqa et al., lay the grounds and propose techniques for analysis of citation networks. One of their key contributions is to review methods of automatically detecting the presence of citation in legal texts. They note that even this simple sounding task alone, is not easy. Although there have been numerous standards and books devoted to proper citation, in many cases the citation text does not follow the correct format and style thus making it hard for automatic extraction of citations from legal documents. They also propose a categorization schema for citations which groups a citation as either an Assertion or an Amendment, which they elaborate in their second paper [10], we will discuss more on this later in this section.

In a more recent work [1], M adedjouma et al., study and investigate the natural language patterns used in cross reference expressions to automatically detect and link a citation to its target. One of their main contributions is in the detection of complicated cross references that are written in natural language. But, unlike us, they do not approach the task of labeling the citations and limit their work on resolving the citation links.

In [13] Maxwell et al., aim to develop a system to help software companies comply with all the regulations. They study the taxonomy of legal-cross references in the acts related to healthcare and financial information systems. They claim to be the first to identify concrete examples of conflicting compliance requirements due to cross-references in legal texts. They analyse different patterns of cross-references that occur in these case studies to obtain seven cross-reference types: *constraint, exception, definition, unrelated, incorrect, general,* and *prioritization* and use grounded theory (the discovery of theory from data [8]) to conjecture that this set of labels are generalizable to other legal domains. Their definitions of *constraint*, *exception*, *definition* and *prioritization* are very similar to our "Limitation", "Exception", "Definition", "Delegation of Authority". While their *unrelated* label does not apply to general purpose citation labeling and only points out the cross-references that are not related to laws governing software systems. Although we have a more detailed set of labels, we do not have a label that corresponds to *incorrect* since we do not look at the cited text and thus we are not able to determine if the citation is indeed correctly citing the desired section of the citee.

T.D. Breaux et al., in [5] design and propose "Frame-Based Requirements Analysis Method (FBRAM)". FBRAM is a software which helps to generate a context-free markup language. Their system facilitates the creation of a model used to systematically acquire a semi-formal representation of requirements from legal texts. The set of labels used in this work are *Exclusion, Fact, Definition, Permission, Obligation, Refrainment.* Their approach in this paper is quite different from ours, since they group/label the text and requirements in the cited text while we are interested in the bigger picture of why the statute is being cited. We must also note that FBRAM is utterly relying on a human analyst and mainly helps only if an analyst manually annotates the whole regulatory document first while we use artificial intelligence and machine learning methods to label cross-references.

In a sequel to their first paper [9], M. Hamdaqa et al. explore the relationships between the citing and the cited law in [10]. Their work is the closest approach to ours in the sense that they also offer an automated system that classifies each citation based on its semantic role in the context. They give a list of advantages in why would one want to explore the relationships among provisions created through citations from one to the other. In short: it is useful in understanding the impact of changes in a law and those depending on it; checking consistencies/conflicts between multiple regulations; eases navigation through laws and their dependencies. They also propose grouping of each edge into Assertions (Definition, Specification, Compliance) and three subtypes of Amendments. They claim that using the verb that is directly related to the citation, one can label the citation into one of the two main groups but do not talk about the possibility of grouping them to the smaller subgroups nor they give numerical evaluations of the accuracy of their approach. In contrast we label each citation into a more refined set and also provide experimental results.

3. PROBLEM STATEMENT

As described in the previous sections, dealing with citations in the legal documents is an important task. In this paper we propose a system that can label each cross-reference according to a predefined set of labels. For the purposes of this paper we only discuss the US Code and its underlying citation graph, but in general our approach can be modified to apply to any other legal citation graph.

A citation graph refers to a graphical representation of all the cross-references in a document to other documents or parts of itself. Nodes, or vertices, in a citation graph are representing the section that is being cited or is citing another section. Edges in this graph are directed and if part of statute A is citing a part in statute B, there is an edge from A to B.

In this paper we introduce an automated semantic labeling of edges in a citation graph. We label/group the edges into a set of predefined labels that classify each edge based on their reason for being cited. For example, in:

> *subsection (a) of this section shall not apply to that portion of the employee's accrued benefit to which the requirements of* <u>*section 409(h) of title 26*</u> *apply*

The cited statute, *section 409(h) of title 26*, imposes a limitation to where the obligations of the citing text would apply.

In the next section we will provide a descriptive summary of each part of the overall system.

4. THE AUTOMATED SYSTEM

As we stated in the previous sections, the main focus of this work is to build a system that can automatically label the edges in a citation graph with a predefined set of labels, each of which represents a possible relationship between the citing provision and the cited provision, that is, the purpose for the citation. The first step towards this goal is to be able to automatically detect the presence and span of each citation in the document. We will next describe our citation extraction method.

4.1 Extracting the Citation Text

The first step towards building this system is to able to identify a citation. Cross-references in the legal domain mostly follow standards and predefined templates. The Bluebook [3] or the newer Citation Manual from US Association of Legal Writing Directors (ALWD) [17] are among the manuals that contain rules for proper citing of legal texts. But as previously mentioned these rules are not always followed.

To extract the citations from a document (e.g., the US Code), we used a complex regex pattern-matching schema that attempts to locate and identify a variety of known formats for citations. The result is the extraction of a number of known corpora types, which then go through an additional processing schema developed to split each extraction - which can potentially include multiple references to the same or different corpora, such as ”26 USC sections 1, 2, and 3 . . . ” or ”28 USC 121 and 10 CFR” - into individual elements and then re-combine them according to basic citation rules, so that it would produce the following: ”26 USC 1”, ”26 USC 2”, ”26 USC 3”, ”28 USC 121” and ”10 CFR” as 5 separate references.

4.2 Feature Extraction

A key idea in this method is our novel feature selection. We find a section of the text related to the citation, the *predicate*, and use this as the main feature in our classification. The *predicate* of a citation to be that portion of the text immediately preceding the citation that expresses the citation’s meaning.

During the annotation process along with collecting a labeled set of citations we also asked each annotator to tag the span of the corresponding ”predicate”, which we will talk about in more details in section 5.3. For the purposes of this work, we define the *predicate* as:

1. The full span of words, that

2. Directly expresses the relationship of the cited provision to something in the current section, and

3. That would make sense if applied to any other provision, i.e., contains nothing specific to the subject matter of the particular section (e.g., funds, exemption), and

4. That expresses as much of the semantics (meaning and purpose) of the relationship as possible without violating 1-3.

For example, in:

> *. . . all provisions excluded from this chapter under Section 42 U.S.C 1879 . . .*

the word *under* is not the full possible span that still satisfies (2)-(4), thus violating criterion (1). The phrase *provisions excluded from this chapter under* includes *provisions*, which is not a relationship but is instead the thing that the citation applies to, violating criteria (2) and (3). However, *excluded from this chapter under* satisfies all 4 criteria.

To automatically extract the *predicate* we designed and trained a linear-chain Conditional Random Field (CRF) on our collected annotated data. The correlated sequential structure of the words in a predicate can be well captured with this type of graphical models, which our experimental results in section 5.4 demonstrate too.

4.3 Classification

One of our main contributions is the automatic process of labeling citations in a legal citation graph. To achieve this goal we utilize an unsupervised learning algorithm to cluster the citations based on a simple word embedding of the extracted *predicates.*

More precisely we first train a shallow two layered neural network on a large corpus of English text extracted from wikipedia and fine tuned it by another round of training on the whole corpus of US Code. This approach is a well known method for representing words as vectors in a high dimensional space of real numbers first introduced by Tomas Mikolov et al. in [15]. We then use these vectors as the underlying representation of words in the predicate and cluster them using k-means. Subsequently each citation is labeled based on the cluster representing it. More detailed explanation and experimental evaluations are presented in Section 5.

4.4 Complete System

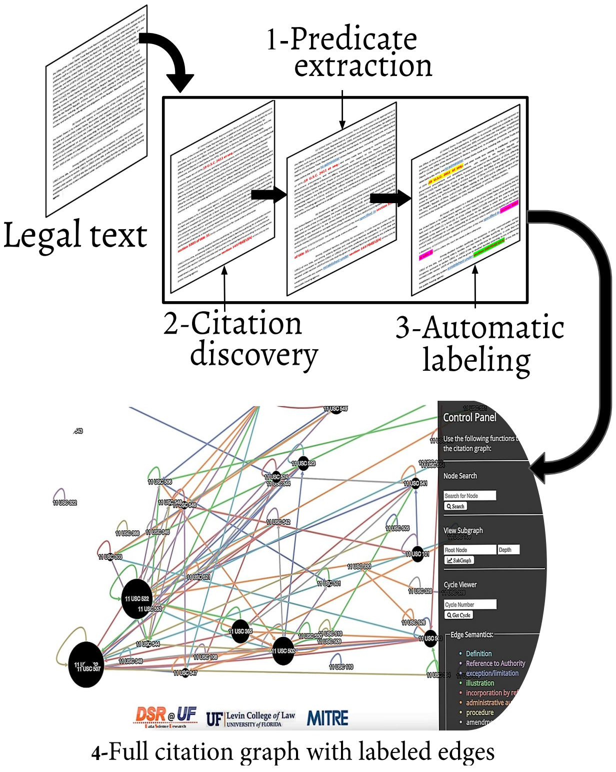

Figure 1: Over view of the end-to-end system.

In summery the complete system enables automatic labeling of the citations in a legal document. After the legal document is given to the system input, in step one it detects all the citations present in the document using the methods described in Section 4.1. It then automatically extracts what we call the predicate which contains information about the type of the at hand citation, this step was described in Sections 4.2 and 5.4. In the next step it utilizes machine learning techniques described in Sections 4.3 and 5.5 to assign to the citation, an appropriate label. The final labled graph is then illustraded using our graphical user interface where each edge type is colored according to its type. Figure 1 shows a diagram of the complete system.

5. EVALUATION

In this section we evaluate the proposed gold standard label set used to capture the purpose of the citations, Annotated Dataset, CRF model and final Clustering Algorithm. We first briefly acquaint the reader with the dataset used, i.e. the US Code.

5.1 The Dataset

The dataset used to demonstrate the use of our system is the US code, which is a consolidation of the general and permanent laws of the United States. There are in total over 130,000 different citations in the US Code. The collection of US Code used was taken from the online repository of Legal Information Institute of Cornell Law School [6]. There are over 29000 distinct citations to statutes in the US Code, Code of Federal Regulations and other sources. These laws cite 26417 distinct US codes with the US law code ”42 USC 1395x” being cited the highest.

Next we introduce the set of labels used in the semantic labeling of the citations.

5.2 Designing the Golden Labels

We inspected many random provisions found in the US Code and proposed a primary set of labels that could capture the relations found there in. This labels along with a set of unlabeled citations from the US Code was then annotated by a group of expert human annotators.

After analysing the results, we merged the labels that were too close to be separated and caused confusion; also expanded the labels by adding new labels found to be necessary. Integrating the feed back we got from the first round of annotations we updated the labels. We believe that the purpose of each citation can be effectively captured with this set of 9 labels.

- **Legal Basis**: A relationship between a program/entity and the statute that is its legal basis.
- **Authority**: A relationship under which one party is permitted to direct another party to perform an action.
- **Definition**: A citation that directly defines the subject, brings a definition for a term used in the rule.
- **Example or illustrations**: A citation to a rule that is used to introduce something chosen as a typical case or is defining the subject by illustrating/describing it.
- **Exception**: A link between a rule and a set of circumstances where that rule doesn’t apply.
- **Criterion**: A link from a conclusion to the ”standard/criterion”, but not how (not the procedure), of reaching that.
- **Limitation**: A relationship between a description and a restriction on that.
- **Procedure**: A link from an activity to a description of how that activity should be performed
- **Amended by/Amendment to**: A relationship between two versions of the rule.

As we discuss in Section 5.3, the final round of annotations by the human experts confirmed the validity of this labels. The result is a label set long enough to cover almost all of the citations and also short enough for practical use.

5.3 Annotation Process

To apply machine learning paradigms for labelling citations and also test the coverage of the gold standard label set described in the earlier section, we need to have a set of data manually annotated. Manual annotations lead to semantic problems similar to ones discussed in [5]. A crowd sourced option like Amazon Mechanical Turk, as mentioned in [2] can be a good medium for the manual annotation process. The problem with crowd sourcing is the absence of critical legal expertise with the annotators, which impacta their judging abilities for a domain specific task. The manual annotators could experience problems like Logical Ambiguity, which would need legal expertise to be resolved. To mitigate these issues associated, the manual annotator group for the project comprised of 7 Graduate law students, with the guidance of a team of legal experts.

The experiment was designed to run in two stages. The first stage comprised of a set of 200 randomly selcted citations that were distributed to the annotators. The first set of annotations helped us expand and modify the gold standard to include for the citations that were deemed to be included in newer labels by the manual annotators. The second round then generated the training samples for our machine learning paradigms to label the citation. The second round also validated the gold standard as the manual annotators did not find a need to expand the labels set to accommodate for any citation. The second round of annotations produced 394 labeled citations that served as the basis for the the Clustering algorithms explained in the following sections.

Out of the 394 citations that were manually annotated, only one was found to need a new label not in our label set. This confirms that our label set covers the citations in the US Code very well.

5.4 Predicate Extraction

To find the proper predicate in the context of the citing provision, we used a linear-chain CRF. Conditional Random Fields are probabilistic graphical models that very well capture the sequential property present in words in natural language [12]. A detailed description of CRFs can be found in [12, 19].

To create the features, we manually looked at samples of the raw text and considered the properties of the predicate.

First, the predicate is almost always preceding the citation. Second, the predicates usually have a certain part of speech (POS) role. For example:

- Preposition-Verb-Preposition *The term "commodity broker" means futures ... or commodity options dealer, as defined in section 348d.*, or
- Preposition *Notwithstanding subsections (a), (b), and (c) of this section and paragraph (2) of this subsection, the court shall disallow any claim for reimbursement or ...*, or
- Preposition-Noun-Preposition *Without regard to the the Richard B. Russell National School Lunch Act (4 U.S.C. 1751 et seq.)*, or etc.

Third, specific words such as *under, defined, amended*, tend to appear more in predicate span.

Trying to keep the features as simple as possible and keeping these properties in mind, we defined the following set of features for each token. Also, we replace the whole span of the target citation text, for which we intend to find the predicate, with a unique character sequence not present in the rest of the corpus, i.e, `C1CITE`. This will make it easier for the CRF to recognize the the citation and work with it as single word. To mark the span of each predicate we used the standard Begin/In/Out (BIO) encoding for tagging the predicate chunks and the other tokens.

Exact word features We used the exact lowercase token of each word and its neighboring words (before and after it) as three features for each token. We must note that this and other multi-valued categorical variables were binarized for use in the model.

Is digit feature We used a boolean feature to determine if a token is a digit or not.

Part of speech features Based on the lexical tags produced by NLTK [4], each word and its neighboring words were assigned with their corresponding POS tags. In addition to that we used the first two and the last two letters of the tag as additional features for the word and its neighbors. This helps when NLTK produces refined POS, for example NNP and NN might have to be treated the same in detecting the predicates.

Distance to citation features We used 5 boolean features determining the relative position of the word to the target citation. $f_1 = 1$ if the word appears after the citation. $f_2 = 1$ if there are no tokens between the word and the citation. $f_3 = 1$ if there is exactly one token between the word and citation. $f_4 = 1$ if there are more than two words in between. $f_5 = 1$ if there are more than four words in between.

Miscellaneous features Other features used were to determine if the word was at the beginning of a sentence, end of a sentence or if the token is a punctuation.

To evaluate the performance of this model, we applied the system to a dataset of 1000 citations and their corresponding predicates[1]. We performed a 10-fold cross validation and presented the performance results in Table 1.

Table 1: Predicate extraction performance

	Prec.	*Recall*	*F1*	*support*
B_PRD*	0.91	0.84	0.875	100
I_PRD	0.93	0.88	0.898	119
O	0.99	0.99	0.998	6518

*Following BIO encoding the begining of a predicate is tagged with B_PRD, any other word in the predicate span is tagged with I_PRD and any word that is not a part of the predicate is tagged with O.

5.5 Clustering Accuracy

As mentioned before, after extracting each citation's predicate we used word2vec [15, 14] to represent each word in the predicate as a vector in a 300 dimensional space. To further simplify the clustering, we correspond each predicate with the average of the vectors representing each of the words in that predicate. Although this averaging results in a loss of information, but due to the properties in the embedding method we used most the meaning in the predicate is still preserved.

To cluster the data we use k-means classification and cluster the whole US Code using 15 cluster centers. Note that since we have a relatively large number of labels and there is no guarantee that each form exactly one cluster in the projected space. For this reasons, we use more cluster centers to capture the spread as much as possible. This might slightly over-fit or even decrease the accuracy, but its effects are negligible compared to the relatively large dataset and number of labels.

To evaluate the performance of our clustering algorithm, we use the annotated dataset obtained from the human expert annotators. Each cluster is labeled according to the label of the closest point to the center of the cluster. We present the classification accuracy and the confusion matrix in Table 2.

Among the 394 citations that were manually annotated 1 needed a label that was not in the gold standard, we denote it with New Label Necessary (NL) in the table.

Table 2: Empirical Confusion Matrix

	Am.	Au.	Cr.	De.	Il.	Ex.	Le.	Li.	NL.	Pr.
Am.	25	0	1	0	0	2	1	0	0	0
Au.	0	3	6	0	0	0	13	0	0	0
Cr.	0	0	47	0	0	4	29	1	0	0
De.	0	0	4	44	0	1	3	0	0	0
Il.	0	0	0	0	0	0	0	1	0	0
Ex.	0	0	6	0	0	22	6	4	0	0
Le.	2	0	9	2	0	7	98	1	0	0
Li.	0	0	8	0	0	2	3	9	0	0
NL.	0	0	1	0	0	0	0	0	0	0
Pr.	0	0	10	0	0	6	13	0	0	0

6. CONCLUSION AND FUTURE WORK

We presented an automated system that automatically determines the purpose behind a citation. This enables lawyers

[1] This dataset was also obtained during the annotation process, but lacked a semantic label for the citations.

Table 3: Table 2 key

Am	Amendment	Ex	Exception
Au	Authority	Le	Legal Basis
Cr	Criterion	Li	Limitation
De	Definition	NL	New Label Necessary
Il	Example or Illustration	Pr	Procedure

and policymakers better analyse the relation between different laws or users to find the necessary regulations much easier.

Our system has three main part. We first automatically extract the citations from the document, then find an informative expression from the text related to that citation which we call it as the predicate. Using Natural Language Processing (NLP) and machine learning techniques we then label the citation into one of the predefined set of citation types.

Our contributions in this paper are three-fold. We propose a gold standard label set that almost all the citations in the legal domain (specially laws and regulations) can be categorized according it and verified its coverage in manual experiment by a group of experts. We also produced a dataset of 394 annotated citations from the US code that can be used for future research on this topic. Finally we built a fully automated system for semantic labeling of the edges over a legal citation graph.

In future work we plan to have a more in depth analysis of the results from annotation process and the accuracy of a human expert. We further plan to use advanced machine learning techniques to increase the accuracy of our system by using the whole context related to the citation.

7. ACKNOWLEDGEMENTS

The authors would like to thank Vironica I Brown, Roman Diveev, Max Goldstein, Eva L Lauer, Nicholas W Long, Paul J Punzone and Joseph M Ragukonis for their contributions in the annotation process. We would also like to thank Benjamin Grider for his help in designing the graphical user interface for our system.

This work is partially supported by UF CISE Data Science Research Lab, UF Law School and ICAIR program.

8. REFERENCES

[1] M. Adedjouma, M. Sabetzadeh, and L. C. Briand. Automated detection and resolution of legal cross references: Approach and a study of luxembourg's legislation. In *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*, pages 63–72. IEEE, 2014.

[2] O. Alonso and S. Mizzaro. Using crowdsourcing for trec relevance assessment. *Information Processing & Management*, 48(6):1053–1066, 2012.

[3] H. L. R. Association. *The bluebook: A uniform system of citation.* Harvard Law Review Association, 1996.

[4] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python.* " O'Reilly Media, Inc.", 2009.

[5] T. D. Breaux and A. I. Antón. A systematic method for acquiring regulatory requirements: A frame-based approach. *RHAS-6), Delhi, India*, 2007.

[6] Cornell Law School. U.S. Code.

[7] F. Galgani and A. Hoffmann. Lexa: Towards automatic legal citation classification. In *AI 2010: Advances in Artificial Intelligence*, pages 445–454. Springer, 2010.

[8] B. Glaser and A. Strauss. The discovery grounded theory: strategies for qualitative inquiry. *Aldin, Chicago*, 1967.

[9] M. Hamdaqa and A. Hamou-Lhadj. Citation analysis: an approach for facilitating the understanding and the analysis of regulatory compliance documents. In *Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on*, pages 278–283. IEEE, 2009.

[10] M. Hamdaqa and A. Hamou-Lhadj. An approach based on citation analysis to support effective handling of regulatory compliance. *Future Generation Computer Systems*, 27(4):395–410, 2011.

[11] W. G. Harrington. Brief history of computer-assisted legal research, a. *Law. Libr. J.*, 77:543, 1984.

[12] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[13] J. C. Maxwell, A. I. Antón, P. Swire, M. Riaz, and C. M. McCraw. A legal cross-references taxonomy for reasoning about compliance requirements. *Requirements Engineering*, 17(2):99–115, 2012.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[16] T. Neale. Citation analysis of canadian case law. *J. Open Access L.*, 1:1, 2013.

[17] A. of Legal Writing Directors & Darby Dickerson. Alwd citation manual: A professional system of citation, 2000.

[18] H. L. Roitblat, A. Kershaw, and P. Oot. Document categorization in legal electronic discovery: computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.

[19] C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, pages 93–128, 2006.

[20] P. Zhang and L. Koppaka. Semantics-based legal citation network. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 123–130. ACM, 2007.