2009 Sixth International Conference on Information Technology: New Generations

# Citation Analysis: An Approach for Facilitating the Understanding and the Analysis of Regulatory Compliance Documents

Mohammad Hamdaqa and Abdelwahab Hamou-Lhadj
*Department of Electrical and Computer Engineering*
*Concordia University*
*Montréal, Québec, Canada*
*{m_hamdaq, abdelw}@ece.concordia.ca*

## Abstract

*Regulated companies are required to comply with the many laws, regulations, standards, and guidelines that apply to them. The sheer volume of regulatory compliance requirements for even a small company can be considerably high, which renders the understanding of such authoritative rules a challenging task without tool support. After inspecting several regulatory documents, we noticed that they contain a significant number of citations that, if explored effectively, can reveal important information about the containing documents. In this paper, we propose a technique called citation analysis that aims at helping users to understand and analyze regulatory documents in an efficient manner. Our approach is based on the exploration of citation graphs extracted from regulatory documents. We discuss the challenges when dealing with citations. We also present an overview of a tool that can support citation analysis.*

**Keywords:** Compliance management, regulatory compliance, citation analysis, citation graphs.

## 1. Introduction

The need for companies to comply with the variety of regulations, laws, standards, and guidelines that apply to their industries has never been more important than in recent years. This increase in attention to regulatory compliance is driven by many factors including corporate scandals, the removal of trade barriers, and a shift in government, business and social views. A few examples of these regulations, standards, and guidelines include the Sarbanes-Oxley (SOX) act, the Health Insurance Portability and Accountability act (HIPPA), the quality management systems standard (ISO 9001:2000), etc. Failure to adhere to these authoritative rules may lead to negative publicity, customer dissatisfaction, loss of business, severe fines, lawsuits and legal actions [1].

While regulatory compliance obligations were created with the intention to ensure proper and ethical conduct of business, the sheer volume of authoritative rules represents a significant challenge for many organizations. For instance, a publicly traded pharmaceutical company has to comply at least with the FDA (Food and Drugs Administration) regulations, HIPPA, and the Sarbanes-Oxley (SOX) Act. To further complicate matters, as noted by Breaux et al. [2], regulatory compliance documents are written by lawyers and other experts in the field, while the end users, who are responsible to read, understand, and analyze these regulations, are auditors and technical stakeholders who do not necessarily have the same level of expertise.

There is a need for a better way of representing and organizing regulatory compliance documents in such a way that it easier to explore and analyze their content. This will facilitate such activities as extracting and prioritizing the main provisions, uncovering similarities and conflicts among inter-related regulatory compliance documents, checking for compliance, etc.

After inspecting manually many regulatory compliance documents, we have noticed that they contain a significant number of citations. These citations relate different parts of a regulatory document to other parts of the same document or a different one. We believe that the study of these relations can reveal important information about the regulatory documents such as the most important provisions by ranking provisions according to the number of times they are cited, or comparing two regulatory documents based on the shared provisions, etc.

In this paper, we propose a technique, called citation analysis, which aims at exploring citations that exist in legal documents with the ultimate objective being to facilitate the understanding and the analysis of their content. Although we focus, in this paper, on only regulations, the approach presented in this paper is equally applicable to other types of authoritative rules such as standards, guidelines, etc.

**Organization of the paper:** In the next section, we discuss the concept of citations in legal documents and their characteristics. In Section 3, we present our citation analysis approach, based on the exploration of a citation graph extracted from a regulatory document. We also discuss how citation analysis can be supported by a tool.

978-0-7695-3596-8/09 $25.00 © 2009 IEEE
DOI 10.1109/ITNG.2009.161

We present related work in Section 4, followed by a conclusion and future directions.

## 2. The Concept of Citations in Legal Documents

### 2.1. What is a Citation?

According to Merriam Webster dictionary [3], a citation means "*an act of quoting; especially : the citing of a previously settled case at law*". Although this definition is too narrow to be applied to the various types of citations (it only focuses on settled cases at law), it points out to the fact that a citation describes a relationship between two documents (or parts of these documents), where one document (the citing) refers to another document (the cited). Legal citations are citations found in legal documents, which usually connect the provisions of one document to the provisions of either the same document or a different one.

Legal citations play an important role in enforcing the legitimacy of the arguments and propositions contained in legal documents, and hence enabling lawmakers to legitimize their actions. They maintain this legitimacy while at the same time minimize the space needed to write legal documents [4]. In addition, legal citations can be used to guide the reader through which authorities to check and in which order. They also provide information about source authority of the cited document such as the name of the authority (e.g., CFR), the date on which the cited document was created, etc. Finally, like any other type of references, legal citations ensure that the original owner of the ideas, thoughts, or words of the cited documents, is given credit.

We can distinguish between two types of citations: internal and external citations. Internal citations refer to parts of the same document, whereas external citations link two different documents together.

There are various types of legal documents including cases, statutes, and administrative regulations. Cases are based on a judicial decision that is reported in the countries that use common law systems. Statutes are enacted by legislature, whereas administrative agencies adopt and amend administrative laws under the authority granted to them by statutes [5]. Our focus in this research is on statutes and administrative laws (that we refer to commonly as regulations).

### 2.2. Citation Styles in Regulations

Properly citing legal documents has always been a difficult and tedious task [4]. To help with this process, a number of citation manuals that contain a set of comprehensive rules have emerged, among which the most popular ones in the U.S. are perhaps the Bluebook [6] and the ALWD (Association of Legal Writing Directors) [7] manuals. In Canada, the most common citation manual is the Canadian Guide to Uniform Legal Citation published by McGill Law Journal [8].

Figure 1 shows the common style for citations used in U.S. federal regulations. It contains two citations, the first refers to the Code of Federal Regulations (CFR), whereas the second one is taken from the Sarbanes-Oxley Act and refers to the U.S. Code.

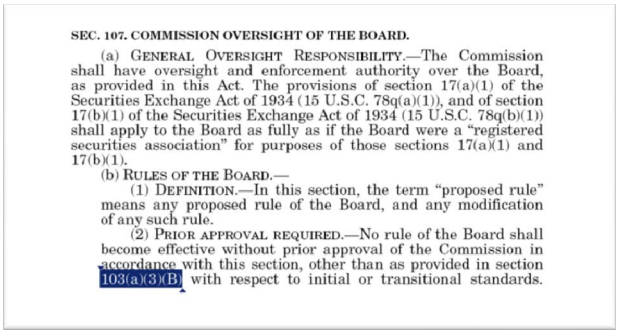
1 C.F.R. § 20.

15 U.S.C. § 78u(d)(3)(B)(iii)(II)(aa).

**Figure 1. Example of two citations using a standardize style**

A citation consists of three main components. The first component is the volume or title number. The number 1 in the first citation of Figure 1 indicates that the citation refer to Title 1 of the Code of Federal Regulations (CFR) which is "GENERAL PROVISIONS". A volume or title number is followed by the name of the regulation (the second component of the citation), which is usually abbreviated and written in a normal font without any specific formatting (e.g., bold, italics or underline). In Figure 1, C.F.R stands for the Code of Federal Regulations, whereas U.S.C., in the second citation stands for the United State Code. The third component of a citation consists of the section number, which is usually preceded by the section sign §. The first citation of Figure 1 refers to section number 20 of the United State Code. Additional information may be contained in a section number including subsections, paragraphs, subparagraphs, clauses, and sub-clauses Each part is specified using a standard format that follows the standard paragraphing hierarchy [12]. For example, a subsection is denoted by a lower case alphabet put between round brackets or parentheses. The second citation of Figure 1 refers to Section 78u of the United State Code, subsection d, paragph 3, subpargrpah B, etc. Finally, it is worth mentioning that a citation must always end with a period.

Knowing the standard citation format used in legal documents can help build tools that extract automatically the citations and the relationships among them. However, after studying many regulatory documents, we have realized that the standards have not always been followed. For example, in the SOX act, the section sign (§) is not used except in a few situations, where amendments to other acts have been references. Another example is the one in Figure 2 where the reference to the Security Exchange Act does not follow any standard.



SEC. 107. COMMISSION OVERSIGHT OF THE BOARD.

(a) GENERAL OVERSIGHT RESPONSIBILITY.—The Commission shall have oversight and enforcement authority over the Board, as provided in this Act. The provisions of section 17(a)(1) of the Securities Exchange Act of 1934 (15 U.S.C. 78q(a)(1)), and of section 17(b)(1) of the Securities Exchange Act of 1934 (15 U.S.C. 78q(b)(1)) shall apply to the Board as fully as if the Board were a "registered securities association" for purposes of those sections 17(a)(1) and 17(b)(1).

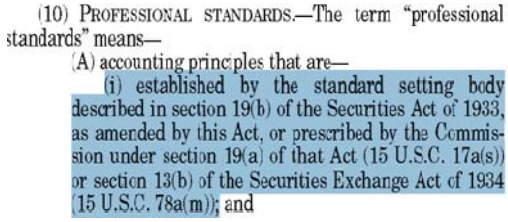
(b) RULES OF THE BOARD.—

(1) DEFINITION.—In this section, the term "proposed rule" means any proposed rule of the Board, and any modification of any such rule.

(2) PRIOR APPROVAL REQUIRED.—No rule of the Board shall become effective without prior approval of the Commission in accordance with this section, other than as provided in section 103(a)(3)(B), with respect to initial or transitional standards.

**Figure 2. Extracted of the SOX act showing a reference (highlighted) that does not comply with the standard**

In addition to this, there are situations where standardized styles are not sufficient in order to extract citations. For example, Figure 3 shows a provision taken from SOX, the usage of the terms "this act" and "that act" make understand which document is being cited difficult. In addition, this example, the "this act" is considered as an implicit internal reference, since it does explicitly indicate the exact location of the amendment that is referred to by the text.



(10) PROFESSIONAL STANDARDS.—The term "professional standards" means—

(A) accounting principles that are—

(i) established by the standard setting body described in section 19(b) of the Securities Act of 1933, as amended by this Act, or prescribed by the Commission under section 19(a) of that Act (15 U.S.C. 17a(s)) or section 13(b) of the Securities Exchange Act of 1934 (15 U.S.C. 78a(m)); and

**Figure 3. Example of a parallel citation found in SOX**

Another source of ambiguity when dealing with citations is the use of parallel citations, which occur when a document contains two citations, one after another, that refer to the exact same information found in two other different documents. Parallel citations usually appear in the context of court cases, where the first cited document is the original and official source of the information, whereas the second one refers in many situations to a document archived by a private company such as the Westlaw company in the U.S., that specialized in reporting on court cases [9]. The objective of parallel citation is to enforce the existence of the cited provision.

Parallel citations have also been used extensively in other types of regulatory documents (e.g., statutes and administrative laws). In most cases, however, the second cited document refers to a legal repository of codes that correspond to the existing laws. For example, In the U.S., each law is codified and its code is added to a code repository known as the U.S. Code [10], which is a legal document that contains codes of all U.S laws. The example in Figure 3 shows a parallel citation, where both cited documents, the "Securities Exchange Act of 1934 section 13(b)" and the "15 U.S.C. 78a(m)" of the United States Codes, contain the exact same information.

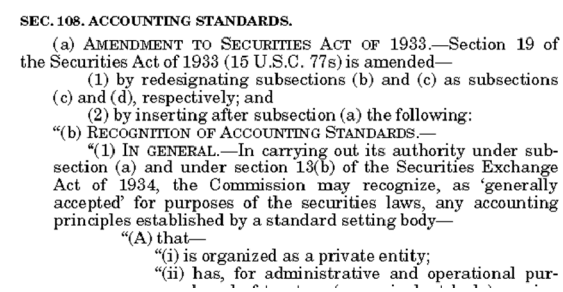
## 3. Citation Analysis

The objective of citation analysis is to facilitate the understanding and analysis of inter-related regulatory compliance documents by exploration the citations that connect provisions to other provisions within the same document or between different documents. Citation analysis uses a citation graph extracted from a regulatory document.

### 3.1. Citation Graphs

A citation graph is directed non-ordered graph G=(V, E) where:

- V = Represents a set of vertices (or nodes) which consist of the citing and the cited provisions.
- E = Represents a set of edges. An edge between Node A and Node B exists if A has a citation to B. It should be noted that the citation graph includes both internal and external citations.



SEC. 108. ACCOUNTING STANDARDS.

(a) AMENDMENT TO SECURITIES ACT OF 1933.—Section 19 of the Securities Act of 1933 (15 U.S.C. 77s) is amended—

(1) by redesignating subsections (b) and (c) as subsections (c) and (d), respectively; and

(2) by inserting after subsection (a) the following:

"(b) RECOGNITION OF ACCOUNTING STANDARDS.—

"(1) IN GENERAL.—In carrying out its authority under subsection (a) and under section 13(b) of the Securities Exchange Act of 1934, the Commission may recognize, as 'generally accepted' for purposes of the securities laws, any accounting principles established by a standard setting body—

"(A) that—

"(i) is organized as a private entity;

"(ii) has, for administrative and operational pur-

**Figure 4. Example of another provision taken from SOX**

After manually inspecting many regulations, we found that the relations that relate the citing provisions to the cited ones can be grouped in two categories: Assertions and Amendments. A citation is considered as an assertion if it refers to a provision that is used to support the writer's point of view, through examples, definitions, or any other additional information. An assertion relation can be further divided into the following subtypes:

- Definition: This is the case where the cited provision defines the citing provision.
- Specification: This is the case where the cited provision provides more information about the citing provision.

- Compliance: In this case, this relation indicates a cited provision that complies with the citing one.

A citation is considered as an amendment if the citing provision amends the cited provision (or part of it). Amendments can be divided into subtypes including:

- Amendment by insertion: The citing provision adds more details or complete parts to the cited provision.
- Amendment by deletion: The citing provision deletes parts of the cited provision.
- Amendment by striking: This relation is used to attract the readers' attention by crossing the information about a cited provision that is not longer valid and inserting new parts or details.
- Amendment by redesignation: This occurs when he cited provision changes the name. The new name is then reflected in the citing provision.

Figure 5 shows a citation graph that is extracted manually from the paragraph of Figure 4 taken from the SOX act.

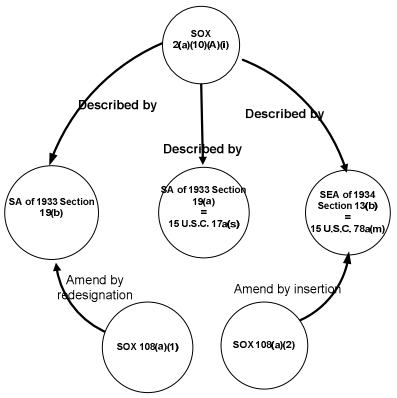



**Figure 5. A citation graph extracted from the provision described in Figure 4**

## 3.2. Using Citation Graphs in a Tool

In this subsection, we discuss the architecture of a tool that we are developing and which leverages citation analysis for the understanding and analysis of regulatory documents. Figure 6 shows the main components of the proposed tool.

At a high-level, the tool takes a regulatory document (in PDF or HTML), converts it into a structured format, generates a citation graph from it, which can be further analyzed by the users.

Most regulatory documents are represented in PDF or HTML formats in an unstructured manner. These formats can be hard to process as shown by Mehrdad and Lethbridge [11]. There is a need to transform the regulatory document into a structured format. We use XML as the data carrier of the structured version of the document. Our preliminary results performing these transformations on sample regulatory documents showed that the process can be automated at a great extent. This is due to the fact that most regulations use the standard paragraphing hierarchy style [12]. In addition, during the transformation, we annotate each provision of the input document with its full name using the standard citation format. For example, each provision in the structured version of the SOX act will be preceded by the volume number, abbreviation that refers to the act (SOX in our case), the title number, and section number.

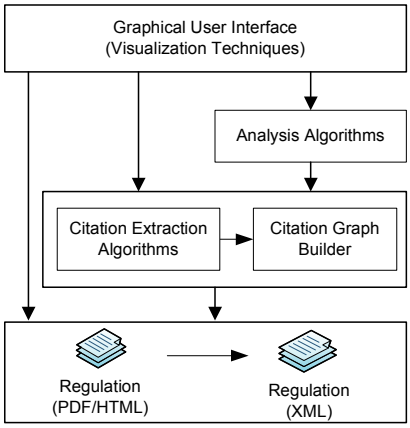



**Figure 6. An overview of the architecture of a citation analysis tool**

Once the regulatory document is saved in such a way that is easy to process, we proceed with extracting the citations and building the citation graph. As previously mentioned, automatic extraction of citations can be challenging due to many factors including the non-compliance to existing citation style standards, the existing of implicit and parallel citations, the various types of relations between the citing and cited provisions (i.e., assertion and amendment), etc. We are currently in the process of investigating how a combination of regular expressions and text mining techniques can be used to build automatic citation extraction algorithms.

The aim of the analysis component is to enable the user (e.g., an auditor, a compliance manager, etc.) to extract meaningful information from the regulatory documents through the analysis of its corresponding citation graph including:

- Searching for specific provisions.
- Ranking and prioritizing provisions based on the number of times they are cited.

- Comparing regulatory documents to uncover similarities and conflicts based on the commonly cited provisions. For this purpose, the tool needs to be able to process citation graphs extracted from various regulatory documents.
- Slicing the graph to focus on only a subset of citations (e.g., display only the internal citations).
- Navigating regulatory documents in a more efficient manner.

A citation graph might be considerably complex depending on the complexity of the regulatory document from which the graph is constructed. A usable tool should support various visualization techniques (such as the use of color codes, highlights, icons, etc.) to allow the user to efficiently browse large citation graph.

## 4. Related Work

The field of compliance support is a broad area that combines law, natural language processing, information retrieval, decision support, and visualization. Despite the fact that there is extensive work done on each of these fields, there have not been a lot of studies that combines all these fields together to tackle the compliance support problem.

R.L. Jacobson proposed a proprietary tool, called the Interactive Legal Citation Checker, for extracting citations from legal documents to help authors of legal documents to write citations that comply with a citation style [13]. The tool, however, only extracts the citations and does not detect the various relations among citations (i.e., assertion and amendments). We intend, however, to investigate how we can enhance the author's tool to support the extraction of the citation graph presented in this paper. Our work is also different from Jacobson's work in such a way that we focus on the understanding and analysis of legal document and not only on checking citation formats.

Zhang et al. proposed a prototype tool, called the Semantics-Based Legal Citation Network [14], which is a visualization tool that focuses on case law citations. Although, the authors' approach classified the citations into relations (background and forward chaining citations), the goal of the citation network is different from our citation graph. Their work aimed to simulate the way attorneys study law documents, and propose a "general attorney behavior model" [15], which can help them achieve this task in a more efficient manner. Our work differs from the authors study in such a way that we use citation analysis to improve the understanding and analysis of legal documents by users who are not necessarily lawyers. We also focus on using citation analysis to detect similarities and conflict among various citations.

Another related work consists of the REGNET project, led by members of the Engineering Informatics Group from Stanford University. The project aims at creating an information infrastructure that supports U.S. federal and state regulations, and focuses on environmental regulations and related documents [16, 17]. One of the main outcomes of the project is a set of techniques for locating and comparing related regulations based on information retrieval techniques, feature matching, etc. [17, 18, 19]. Although the work of the REGNET approach does not focus on citation analysis, we think that their techniques complement the ones we presented in this paper and we intend to investigate their work in the future.

## 5. Conclusion

In this paper, we proposed a citation analysis technique that can facilitate the understanding and analysis of legal documents. Our approach is based on building a citation graph from a regulatory document. We anticipate that the graph can help authoritative rule users to detect the rank and prioritize the important provisions of a document, understand the relationship between various documents, compare different documents and detect similarities and conflicts, etc.

We also discussed what constitute citations, their characteristics, and the challenges with automatic extraction of citations. Finally, we discussed how citation analysis can be supported by a usable tool.

Immediate future directions consist of building citation extraction algorithms that takes into account the implicit and parallel citations as well as the relations that exist between the citations. The next step is to experiment with citation graphs in order to understand how they can help users understand regulations by browsing the graph. For this purpose, we need to develop a tool that supports citation analysis. The tool usability needs to be studied to cope with overly complex citation graphs.

## 6. References

[1] A. Hamou-Lhadj and A-K. Hamou-Lhadj, "Towards a Compliance Support Framework for Global Software Companies", *In Proc. of the Software Engineering Conference*, 2007.

[2] T.D. Breaux and A. I. Antón, "Mining rule semantics to understand legislative compliance", *In Proc. of the ACM Workshop on Privacy in the electronic Society,* 2005, 2005, pp. 51-54.

[3] "citation." *Merriam-Webster Online Dictionary*. 2008. http://www.merriam-webster.com, [Last access, November 25th, 2008].

[4] Citation Formats Committee of the American Association of Law Libraries, *AALL Universal Citation Guide Version 2.1*, AALL Citation Formats Committee, Dec 30 2004.

[5] OAL, "FAQs: What is the difference between a regulation and a statute?", *Office of Administrative Law official website*, http://www.oal.ca.gov/faqs.htm#7 , [Last access, November 25th, 2008].

[6] Harvard Law Review, *The Bluebook: A Uniform System of Citation,* Grand Central Publishing, July 1996.

[7] D. Dickerson and the Association of Legal Writing Directors, *ALWD Citation Manual*, Aspen Publishers, December 2005.

[8] McGill Law Journal, *The Canadian Guide to Uniform Legal Citation*, Carswell Legal Publications, September 2002.

[9] K. Castetter, *Locating the Law 4th Edition HOW TO READ A CITATION*, Southern California Association of Law Libraries, pp. 4.1, 2001.

[10] Law Reform Commission, *Statute Law Restatement*, Law Reform Commission, July 2008.

[11] M. Nojoumian and T.C. Lethbridge, "Extracting Document Structure to Facilitate a Knowledge Base Creation for The UML Superstructure Specification*", In Proc. of the ITNG conference*, 2007, pp. 393-400.

[12] Department of Justice - Canada, "Paragraphing", *Department of Justice - Canada official website*, http://www.justice.gc.ca/eng/dept-min/pub/legis/n26.html, [Last access, November 25th, 2008].

[13] R. L. Jacobson, "Interactive legal citation checker", U.S. Patent 7028259, April 11, 2006.

[14] P. Zhang, L. Koppaka, "Semantics-based legal citation network", *Proc. of the 11th international conference on Artificial intelligence and law*, 2007, pp. 123 – 130.

[15] S. A. Sutton, "The role of attorney mental models of law in case relevance determinations: an exploratory analysis", Journal of the American Society for Information Science, v.45 n.3, April 1994, pp.186-200.

[16] G.T. Lau, S. Kerrigan, H. Wang, K.H. Law and G. Wiederhold, An information infrastructure for government regulation analysis and compliance assistance, Proc. 5th Conf. on Digital Government Research, Seattle, 2004, pp. 1-2.

[17] G.T. Lau, H. Wang, and K.H. Law, Locating related regulations using a comparative analysis approach, Proc. 7th Conf. on Digital Government Research, 2006, pp. 229 - 238.

[18] G.T. Lau, K.H. Law, and G. Wiederhold, A relatedness analysis of government regulations using domain knowledge and structural organization, International Journal of Information Retrieval 9(6), 2006, pp. 657 – 680.

[19] S.L. Kerrigan and K.H. Law, A regulation-centric, logic-based compliance assistance framework, International Journal of Computing in Civil Engineering, 19(1), 2005, pp. 1-15.