

Citation Analysis: An Approach for Facilitating the Analysis of Regulatory Compliance

Documents

Mohammad Hamdaqa

A Thesis

In

The Department

Of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Applied Science at
Concordia University
Montreal, Quebec, Canada

December 2009

© Mohammad Hamdaqa, 2009

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Mohammad Hamdaqa

Entitled: Citation Analysis: An Approach for Facilitating the
Analysis of Regulatory Compliance Documents

and submitted in partial fulfillment of the requirements for the degree of

Master of Electrical and Computer Engineering

Complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____Chair
_____Examiner
_____Examiner
_____Supervisor

Approved by:

_____Chair of Department or Graduate Program
Director

_____20_____

_____Dean of Faculty

ABSTRACT

Citation Analysis: An Approach for Facilitating the Analysis of Regulatory Compliance Documents
Mohammad Hamdaqa

For global software companies with a client base that covers a large number of regulated businesses, regulatory compliance represents a significant challenge. The problem is that the world of compliance is becoming increasingly complex due to an increase in the number of regulations, laws, and standards that are being introduced every year. In addition, these laws are often created based on existing laws, resulting in a complex set of interdependent provisions where changes made in one place can propagate to affect other laws. To further complicate matters, many of these laws are created by different legislative bodies resulting in overlapping and sometimes conflicting provisions. There is clearly a need to investigate techniques and tools that can alleviate organizations from the complexity of dealing with regulatory compliance. In this thesis, we present an approach that facilitates the analysis of multiple regulations by exploring the citation relationship that links them to each other. Our approach is based on building a citation graph that can be used by an analyst to navigate through the provisions of various interrelated laws to uncover overlaps and possible conflicts or to simply understand specific law documents. We also present a tool called CompDSS (Compliance Decision Support System) that supports our approach. Finally, we show the effectiveness of the presented approach by applying it to different regulations.

Acknowledgements

My special and sincere gratitude for the intellectual sparring partner my supervisor Dr. Abdelwahab Hamou-Lhadj, who I owe this work to him. During two years Dr. Hamou-Lhadj was a kind friend and mentor. His vast knowledge, sound advice, prompts feedback, continuous support, encouragement, patience, and confidence in my success was the reason I'm presenting this today. Thank you Dr. Hamou-Lhadj.

Special thanks for those who support me and help me to achieve my goals:

To my best friend and unison
the wind beneath my wings
my wife Ayesha

To the sweetest ethereal smile
dispelled my gloomy clouds
(The Hope) My Daughter Lujain

To those who witnessed my past
Who believes in my future
accepts me the way I am
and will never let me down
My parents (Adnan and Nabela)

Finally to all my friends especially my wonderful lab colleagues for their hints, tips and tricks during brainstorming. Thank you for spending the time trying to help me even when you were busy with courses or research. Thank you for the happy moments, and making this research an enjoyable experience.

Table of Contents

Table of Contents	v
List of Figures	vii
List of Tables	x
Chapter 1: Introduction	1
1.1 Motivations.....	1
1.2 Thesis Contributions	4
1.3 Organization of the Thesis	5
Chapter 2: Background	7
2.1 What is Compliance?	7
2.2 Complying with Multiple Regulations	8
2.3 Related Work.....	9
Chapter 3: Citation Analysis Framework	13
3.1 Citations in Legal Documents	13
3.2 Citation Styles	14
3.3 Citation Analysis	19
3.3.1 Definition.....	19
3.3.2 Building Citation Graphs.....	20
3.4 Analysis of Citation Graphs	27
3.4.1 Citation Depth Limiting.....	28

3.4.2 Fan-In and Fan-Out Analysis	28
3.4.3 Relatedness Analysis	31
3.5 Compliance Decision Support System (CompDSS)	39
3.5.1 CompDSS Architecture	39
3.5.2 Law Parser	39
3.5.3 Graph Builder (CGBuilder)	45
3.5.4 Analyzer.....	50
3.5.5 Visualization Engine.....	51
Chapter 4: Application.....	53
4.1 Target Regulations	53
4.2 Applying Citation Analysis.....	56
4.2.1 Building the Citation Graph	56
4.2.2 Result of Applying Citation Analysis to GLBA, HIPAA and SOX.....	56
Chapter 5: Conclusion and Future Work	65
Bibliography	68

List of Figures

Figure 2-1 Overlap between regulations.....	9
Figure 3-1 Classification of Citation Styles.....	15
Figure 3-2 Example of two citations using the Bluebook standard.....	16
Figure 3-3 Example of a provision taken from the SOX act showing a reference (highlighted) that does not comply with the standard.....	18
Figure 3-4 Example illustrates citations ambiguity found in SOX.....	18
Figure 3-5 The regular expressions used to extract provisions according to the paragraph hierarchy structure	22
Figure 3-6 Automata for the regular expression that captures the standardized citation styles	23
Figure 3-7 a) Example of a provision taken from SOX before tagging b) Example a tagged provision.....	25
Figure 3-8 Example of the matching process used in extracting the relations	26
Figure 3-9 Example of two provisions taken from SOX	27
Figure 3-10 A citation graph extracted from the provisions described in Figure 4-9.....	27
Figure 3-11 A Citation Graph extracted from SOX, shows (15 U.S.C 78) in the middle, and the large number of SOX provision that refer to this Act.....	30
Figure 3-12 Explanation of overlaps at the provision level.....	32
Figure 3-13 citation graph that shows a conflict between DMCA and TEACHA	36

Figure 3-14 Overall Architecture of CompDSS components that support citation analysis	39
Figure 3-15 A snapshot of one screen from the CompDSS tool, representing the paragraph hierarchy builder (one phase in the LParser)	41
Figure 3-16 A procedure to detect the start and the end of a provision then converting the whole provision to one line.....	1
Figure 3-17 An algorithm to create provision identifier	43
Figure 3-18 A UML class diagram of the CompDSS data model.	44
Figure 3-19 The steps to determine the relationships between the Act provisions and the citations embedded in these provisions.....	46
Figure 3-20 Classification of POS taggers based on (Van Guilder 1995) classification..	47
Figure 3-21 A screen snapshot of the CompDSS tool that shows the extracted provision of the tagging process	49
Figure 3-22 A citation graph that shows the relationship between SOX, GLBA and HIPAA	52
Figure 4-1 A graph showing SOX and HIPAA provisions amending the same provision of the 29 U.S.C 1132 law	59
Figure 4-2 Provisions taken from SOX & HIPAA to illustrate the first case of overlap and possible conflicts.....	60
Figure 4-3 An example of SOX and HIPAA provisions amending the same provision of the 29 U.S.C 1132 law but different sections	62
Figure 4-4 Provisions taken from SOX and GLBA to illustrate a case of overlap that can lead to a conflict unless careful monitoring of the changes is performed	63

Figure 4-5 Example of SOX and GLBA provisions amending the same provision “15

U.S.C 78q” 64

List of Tables

Table 3.1 Analysis Patterns	33
Table 3.2 Some of the relations between sensets.....	49
Table 4.1 Common regulations between SOX, GLBA, and HIPAA.....	56
Table 4.2 Common regulation sections between SOX, GLBA, and HIPAA.....	57
Table 4.3 Statistical information about the three regulations.....	57
Table 4.4 Fan-out analysis.....	58
Table 4.5 Fan-in analysis.....	58

Chapter 1: Introduction

“Regulations grow at the same rate as weeds.”

Norman R. Augustine

1.1 Motivations

For many regulated organizations, regardless of geography and industry sector, regulatory compliance has become an integral part of their business landscape. Recently, there has been a noticeable increase of attention to regulatory compliance, which is driven by many factors, among which perhaps the most important one is the number of large U.S corporations such as Enron, WorldCom, The Red Cross, etc. that have been found guilty of wrongdoing, frauds, and corruption. Other drivers include market globalisation and the removal of trade barriers, the ever reliance on Information Technology (IT) and the necessity to protect customers privacy information and the need for secure systems, and a higher need for business continuity and assurance (*Ernst & Young, 2006*). As a result, more and more regulations and laws are being introduced every year putting further constraints on the way companies are operated, managed, controlled, and governed. The consequences of not complying with these laws can be devastating and may include substantial fines, financial losses, lawsuits, customer dissatisfaction, and loss of reputation and market confidence (*Hamdaqa & Hamou-Lhadj, 2009*) (*Hamou-Lhadj & Hamou-Lhadj, 2009*).

However, complying with the law and regulations is a challenging task for many companies. First, there are just too many laws and regulations to comply with. Consider, for example, a North-American public firm, working in the health sector. This organization should defiantly comply with the Health Insurance Portability and Accountability Act (HIPAA) if it operates in the U.S., or its equivalent, the Personal Health Information Act (PHIA) if it operates in Canada. Being a publicly traded company, U.S. laws require that this company comply with the Sarbanes Oxley Act (SOX) and perhaps the Gramm-Leach-Bliley Act (GLBA), which are regulations that govern the proper reporting of financial statements. This company must also comply with various security laws and standards to protect sensitive information such as patients' personal records. Example of these authoritative rules include the ISO 27000 standards, the Privacy, Cryptography and Security Gridlines issued by the Organization for Economic Cooperation and Development (OECD), the Privacy Act in the U.S., or the Personal Information Protection and Electronic Documents Act (PIPEDA), issued in Canada. If the activity of the company affects Homeland security, other acts must be taken into consideration such as the Federal Information Security Management Act (FISMA), the 2007 Protect America Act (PAA), and so on.

In addition, it has been shown that many existing regulations tend to overlap in their content while many others contain conflicting provisions (*Kerrigan, Lau, Zhou, Wiederhold, & Law, 2001*). This is partly attributable to the fact that laws are often developed by different organizations with little effort for consistency or convergence with similar legal requirements. In fact, conflicts may also exist even within the same regulations that undergo frequent changes and modifications (*Barnett, 2004*). To make

things more complex, regulations are rarely created from scratch, they are often based on other existing laws, creating a network of inter-related provisions where changes in one place may propagate to affect many other laws, which hinders the ability for organizations to keep their procedures and policies inline with the regulations and laws they represent. The problem is further complicated by the fact that laws and regulations are written by lawyers, but expected to be used by managers, auditors, and technical stakeholders with little knowledge of law and law jargon (*Breaux & Antón, 2005*).

Clearly, there is a need to investigate ways to help regulated companies manage a large number of possibly overlapping or conflicting regulatory compliance requirements. This translates into the need to represent and organize regulatory compliance documents in such a way that it is easier to explore and analyze their content so as to help analysts extract and prioritize the main provisions, uncover similarities and conflicts among inter-related regulatory compliance documents, check for compliance, etc. These are the objectives of our research, with a particular focus on helping global software companies cope with the increasing demand from regulated businesses for software solutions that satisfy regulations, laws and standards that apply to their industry sector. This is because many laws have a direct impact on the way software systems, used by regulated businesses, are developed and maintained. For example, a data records management tool, used by a health institution which is required to comply with HIPAA, must support many data security features such as a password-based protection mechanism, different levels of data access control, and frequent backups and data reliability techniques. Our research pertains to a newly introduced field of study that is referred to as Software Compliance (*Hamou-Lhadj, Software Compliance Research Group, 2009*) which is defined as a

software engineering field that is concerned with investigating techniques and tools to develop, test, and maintain software systems where compliance is a built-in quality attribute.

After manually inspecting 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 within 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 in a law document 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 thesis, we, therefore, 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.

1.2 Thesis Contributions

The main contributions of this thesis are as follows:

- Introduce the concept of citation analysis as a technique for facilitating the understanding and analysis of multiple regulatory compliance documents, with a particular emphasis on detecting overlaps and possible conflicts among regulations.

- Develop a set of algorithms for building citations graphs extracted from multiple regulations.
- Develop a tool, called CompDSS (Compliance Decision Support System), which allows manipulating citation graphs to perform various analyses on regulatory documents.
- Apply the techniques described in this thesis to many North-American laws to show the effectiveness of the proposed approach.

1.3 Organization of the Thesis

The rest of the thesis is structured as follows:

In Chapter 2, we present the background needed to understand the objectives of our work. The chapter provides background knowledge on compliance and compliance management. The chapter also discusses related work and how they differ from the work proposed in this thesis.

In Chapter 3, we present the citation analysis framework. To achieve this, we first discuss the concepts of citation and citation analysis. We then present a detailed approach for extracting citation graphs from multiple documents. The chapter ends with a description of our tool that supports citation analysis, CompDSS.

In Chapter 4, we validate our approach by applying it to a set of U.S. regulations, we use CompDSS to extract citation graphs from various regulations and measure various

aspects of these graphs. After that we analyse the graphs to detect overlaps and conflicts among the targeted regulations.

We conclude the thesis in Chapter 5 with a summary of the main contributions, some future directions, and a concluding remark.

Chapter 2: Background

“Studies have shown that regulatory burden and compliance is one of the most significant costs to a small bank, substantially more than anything other than your payroll.”

Bill Hutmacher

2.1 What is Compliance?

Compliance is the adherence to government laws, regulations, acts, standards, best practices and guidelines, and organizational bylaws and doctrines (*Cougias, Halpern, Herold, Koop, 2007*). There are various types of regulatory documents, which are commonly referred to as authoritative rules. These include laws and regulations, industry standards, guidelines and best practices, and organizational bylaws. Laws and regulations are developed by governmental bodies and are the only authoritative rules that an organization must comply with. Industry standards represent agreed-upon practices by experts in a particular industry. There exist several standardization bodies such as ISO (*ISO, 2009*), OMG (*OMG, 2009*), etc. that manage the development and evolution of standards. Complying with an industry standard is not an obligation but it is highly recommended for the organization to be competitive. Best practices and guidelines represent reference work in a particular field. These practices are not standardized and as such they can be customized to fit the needs of a particular organization. Organizational

bylaws consist of internal rules that regulate the way an organization conducts business. In this thesis, the emphasis is put on regulations and laws, although we believe that the techniques presented here are equally applicable to other authoritative rules.

The scope and focus on laws and regulations vary depending on whether they are defined at the federal, state, or international level. Their impact on organizations, however, is very similar. They require from organizations to be able to provide evidence, through reporting and auditing measures, that they have implemented the required provisions and clauses. For example regulations such as SOX, HIPAA and GLBA in the U.S. require organizations to provide supporting documents that ensure that confidentiality, integrity, and availability of electronic data and customer information have been implemented.

2.2 Complying with Multiple Regulations

Organizations often have to comply with many regulations at the same time. A naïve approach to support these regulations would be to extract requirements for each regulations independently from the others and create policies and procedures that ensure that these re for each regulation.

This approach does not take into account the fact that the set of rules in multiple regulations are not mutually exclusive, there exist many overlapping content. In Figure 2-1, we show three regulations SOX, GLBA, and HIPAA, which although they focus on different areas, they contain many overlapping provisions.

Considering the overlap between multiple regulations can help, on one hand, in reducing regulatory compliance costs by reducing the number of policies and eliminating

implementing the same policy twice while increasing the return on investments (ROI) in compliance tools and technologies. On the other hand, it can help in reducing the non-compliance risk by detecting conflicts between regulations so that the companies can implement the right policies and procedures to deal with the conflicting situations.

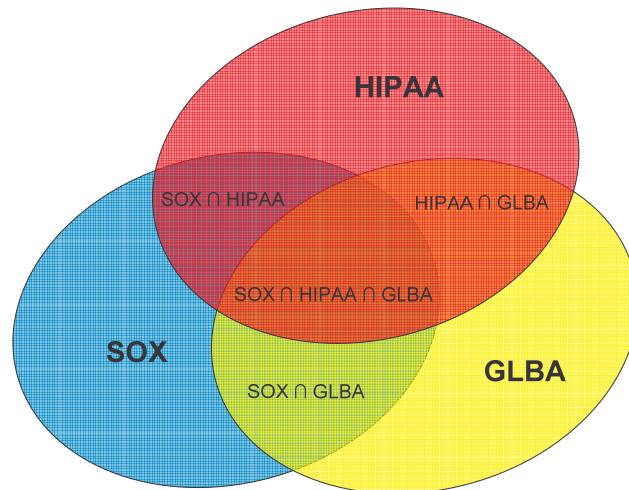


Figure 2-1 Overlap between regulations

2.3 Related Work

The high risk of non-compliance is a major driver for the growing popularity of the field of regulatory compliance. The multi-disciplinary nature of the research issues which covers broad aspects including law, computer science, business management and linguistics, led researchers, from various disciplines, to address the problem. In this section, we survey the studies that we found are closest to ours.

T. D. Breaux proposed what he called a Frame-Based Requirements Analysis Method (FBRAM) to obtain legal requirements from U.S. regulatory documents. Beaux's'

method intended to help software engineers to understand legal requirements in order to avoid non compliance (*Breaux & Anton, 2008*). Breaux used domain-independent upper ontology, natural language phrase heuristics, a regulatory document model and a frame-based mark-up language to represent legal requirements, while maintaining tractability between the regulations and the formal representation in the frame based model. In his methodology Breaux studied the sources of legal ambiguity, like the effect of facts, definitions, constraints, language ambiguity and cross-references. His study of the effect of cross-references was limited to defining a set of patterns to extract internal references, to support mappings between legal requirements (*Breaux, Vail, & Anton, 2006*) (*Breaux T. , 2009*).

Perhaps the most related work to our approach is the work done by REGNET project at Stanford University. The project objective is to build a formal information infrastructure for regulatory information management and compliance assistance to supports U.S. federal and state regulations. Their focus was on environmental regulations and related documents (*Lau, Kerrigan, Wang, Law, & Wiederhold, 2004*) (*Lau, Law, & Wiederhold, 2006*). 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. (*Lau, Wang, & Law, 2006*)(*Lau, Law, & Wiederhold, 2006*) (*Kerrigan & Law, 2005*). For example in (*Lau, Wang, & Law, 2006*) REGNET presented a comparative approach to compare different provisions to find the “degree of similarity” between two documents. REGNET “relatedness measure” was based on feature matching where they defined a feature as “The evidence of relatedness between two provisions, which could contain domain-specific information”, the domain specific information can

be exceptions, definitions and concepts. Although the work of the REGNET approach does not focus on citation analysis, we think that their techniques complement the ones we present in this paper.

Jacobson developed an Interactive Legal Citation Checker, a proprietary tool to extract citations from legal documents, the aim is to help authors of legal documents write citations that comply with a citation style (*Jacobson, 2006*). Seeing how this tool only extracts the citations and does not detect the various relations among citations (i.e., assertion and amendments) (*Jacobson, 2006*). We intend to investigate how we can extract a citation graph such as presented in this paper. Our work differs from Jacobson's in the focus on the understanding and analysis of legal document rather than checking citation formats.

The Semantics-Based Legal Citation Network is a prototype visualization tool that focuses on case law citations. The tool was proposed by Zhang et al. (*Zhang & Koppaka, 2007*). The goal is to help attorneys and other people who work on the legal field to study cases and legal issues without the need to go through the whole cases or the daunting task of manual citation search. The citation viewer focuses on one legal issue and allows an attorney to navigate (forwards and backwards) through the citation network to study how the issue was developed and handled in different cases over the years. While citation networks aim to simulate the way attorneys study law documents, and propose a "general attorney behaviour model" (*Sutton, 1994*). Our work uses citation analysis which is based on citation graphs; to improve the understanding and analysis of legal documents by users who are not necessarily lawyers. We focus on detecting overlaps

between regulations through citation relatedness analysis, as an attempt to mitigate the risk of non-compliance by detecting the sources of the overlaps in regulation acts.

Chapter 3: Citation Analysis Framework

“Policy-based controls are fantastic, but you now have a separate constraint coming in from the side which is compliance. More powerful management tools give you easier changes, but then you have to beef up the audit trail and show your managers and auditors the mapping between your stated policies and the measurements coming up from what's really happening.”

Dave Clarke

3.1 Citations in Legal Documents

According to Merriam Webster dictionary (*Merriam-Webster Online Dictionary, 2008*), a citation is “*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 of 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 (*AALL, 2004*). 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 law 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 and used to grant authority to administrative agencies to adopt and amend administrative (*OAL, 2008*). The focus of this research is on statutes and administrative laws (that we refer to interchangeably as laws and regulations).

3.2 Citation Styles

A citation style is a standardized format that uses abbreviations and special expressions to facilitate the writing and reading of citations. Figure 3-1 is a classification of most

commonly used citation styles in different areas of studies (Roberts, 2009). The figure shows that in-text citations can be divided into two main citation systems; The Note citation system (e.g. footnote styles) which uses numerical numbers in sequence through the text, and the parenthetical citation system that uses abbreviations to represent important information, which can provide information on the cited document. The Harvard citation style is an example of one of the most common styles that uses the parenthetical citation system (Roberts, 2009).

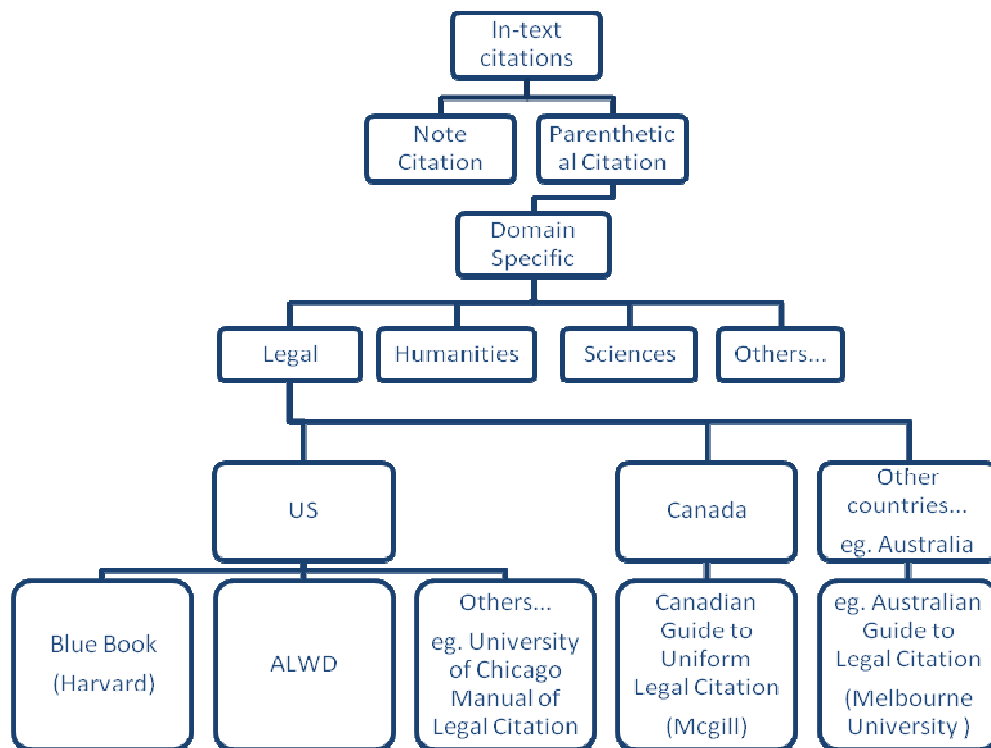


Figure 3-1 Classification of Citation Styles

Legal citation styles have evolved significantly since the “citation reform” that has been undertaken in the U.S. and Canada in 1990 (AALL, 2004). The aim was to create styles that can be readily processed using Information Technology tools so as to leverage electronic storage of laws and documents. The citation reform also tackled the problem of

standardizing citation styles, and making them a public domain, vendor neutral, and medium neutral.

The result of this effort has led to two main citation systems known as the Bluebook from the Harvard Law Review Association (*Harvard Law Review Association, 2000*) and the ALWP (Association of Legal Writing Directors) manuals (*Dickerson, Darby; Association of Legal Writing Directors, 2003*). The two manuals bare a slight difference, which reflect the differences in the environments represented by the law documents used to create the manuals. In Canada, the most common citation manual is the Canadian Guide to Uniform Legal Citation published by McGill Law Journal (*McGill Law Journal, 2006*), which is very similar to the Bluebook citation manual.

These citation manuals contain a comprehensive set of rules to represent most citation styles in the field of legal research. They have been used to cite all sorts of legal sources such as cases opinions, statues and administrative regulations.

Figure 3-2 shows an example of two citations used in U.S. federal regulations using the Bluebook manual. The first citation 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 (U.S.C), which is a repository for U.S. codes, where each law is codified and its code is added to this repository (*Law Reform Commission, 2008*).

1 C.F.R. § 20.

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

Figure 3-2 Example of two citations using the Bluebook standard

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 3-2 indicates that the citation refers 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). The third component of a citation consists of the section number, which is usually preceded by the section sign §. The first citation of Figure 3-2 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 so-called standard paragraphing hierarchy (*Department of Justice - Canada, 2009*) (*United States Congress Data Dictionary of Legislative Documents*). For example, a subsection is denoted by a lower case alphabet put between round brackets or parentheses. The second citation of Figure 3-2 refers to Section 78u of the United State Code, Subsection d, Paragraph 3, Subpargarph 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 automatically extract the citations and the relationships among the citing and cited provisions. 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 referenced. Another source of ambiguity is the one shown in Figure 3-3 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 3-3 Example of a provision taken from 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 3-4 shows a provision taken from SOX, the usage of the terms “this act” and “that act” make understand which document is being cited difficult. Also in this example, the “this act” is considered as an implicit internal reference, since it does not 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 3-4 Example illustrates citations ambiguity found in SOX

3.3 Citation Analysis

3.3.1 Definition

We define citation analysis as a technique that aims to facilitate the understanding and analysis of regulatory compliance documents through the exploration of the relationships among the provisions they contain. Citation analysis uses a citation graph extracted from multiple regulatory documents to analyze the content of these documents and the relationship among them. Dealing with multiple regulations is one of our approach's distinctive advantages. Most existing studies (*Breaux, Vail, & Anton, 2006*) (*Breaux T. , 2009*) in this area focus on only one regulation. We anticipate that citation analysis can help in the analysis and understanding of regulatory compliance documents by:

- Uncovering overlaps and possible conflicts that exist between multiple regulatory documents through analyzing the provisions (and the respective acts) they have in common.
- Detecting important provisions by ranking them based on their relevance. For example, knowing that a particular provision is cited by many other provisions within the same document is a good indicator of its importance. These provisions are most likely the ones that a company needs to focus on. This is particularly important for companies that do not have sufficient resources to comply with the entire set of acts and provisions that apply to them. From the risk analysis perspective, these companies must have a quick way to identify the provisions that are high risk.

- Assessing the impact of change in a particular act. This applies in situations where an act on which many other acts depend on changes. A citation graph can be readily used to understand the impact of these changes by showing in a visual way (using a citation analysis tool) the acts that are potentially impacted by the changes.
- Checking the consistency of multiple acts that make different modifications to a common act on which they depend. For example, if two acts A and B depend on a third act C and that A modifies C's provisions then B is affected by the changes of A. From our experience, most conflicts that exist in regulations are due to lack of tools that assist users in doing consistency checks. We believe that citation analysis, if supported in a tool, can be a powerful solution to this problem.
- Understanding regulations by allowing easy navigation through its provisions and the ones from other acts on which it depends on. To achieve this, a tool that supports citation analysis should allow for features such as searching and slicing based on external or internal citations, limiting the depth of a citation (e.g., section, subsection, paragraphs, etc.), and so on.

3.3.2 Building Citation Graphs

A citation graph is a directed non-ordered graph $G = (V, E, R)$ 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.
- R = Represents the relation between two vertices, more explanation about the types of relations existing in citation graphs is presented in Section 3.3.2.3.

3.3.2.1 Extracting the provisions

Our definition of a provision is similar to the definition provided by Eugen Ehrlich in his classic paper “The Sociology of Law” (*Ehrlich, 1922*) where he defines a provision as “an instruction framed in words addressed to courts as to decide legal cases or a similar instruction addressed to administrative officials as how to deal with particular cases” (*Ehrlich, 1922*). A provision can be in the form of a clause, sentence, or paragraph of a legal document, which provides information for a particular matter. Provisions are considered the core of regulations since they contain the rights and obligations, assets, and liabilities stipulated in the corresponding act. Citations are embedded in the provisions as one of their main components, and play a critical role in understanding the provisions (*Martin, 2007*).

Our process of extracting provisions from regulatory documents is based on the fact that most North-American legal documents use the paragraph hierarchy style, which consists of an indentation system that indicates the beginning of a section, its subsections, paragraphs, etc. (*Department of Justice - Canada, 2009*). However, since most documents are usually saved in an unstructured manner (mostly in PDF and HTML) and come without a table of content, we needed to create a set of regular expressions that can

identify the various components of a provision during the parsing of the law document under study. These regular expressions are shown in Figure 3-5. Each provision is uniquely indexed using a combination of the document title, section, subsection, and other subcomponents.

```
TITLE I: TITLE\s?[IVX]{1, 3}
SEC.: SEC.\s?\d{1,5}
(a): \([a-hj_z]\)
(1): \(\d{1,2}\)
(A): \([A-HJ-Z]\)
(iii): \([ivx]{1,3}\)
(II): \([IVX]{1,3}\)
(aa): \([a-h][a-h]\)
```

Figure 3-5 The regular expressions used to extract provisions according to the paragraph hierarchy structure

Once the document is parsed and its various components are identified, we save it in XML using tags that distinguish between the components of the provisions for further processing.

3.3.2.2 Extracting the Citations

The next step is to extract the citations. We limit ourselves to the citations that follow standardized citation styles, such as the Bluebook and ALWD. For this purpose, we have developed a set of regular expressions that capture the structure of citations expressed in these standards. The resulting regular expressions are represented in the automata of Figure 3-6. We experimented with these regular expressions on SOX, GLBA, and

HIPAA (the regulations used in our case study) and obtained a very high success rate in terms of identifying the citations they contain.

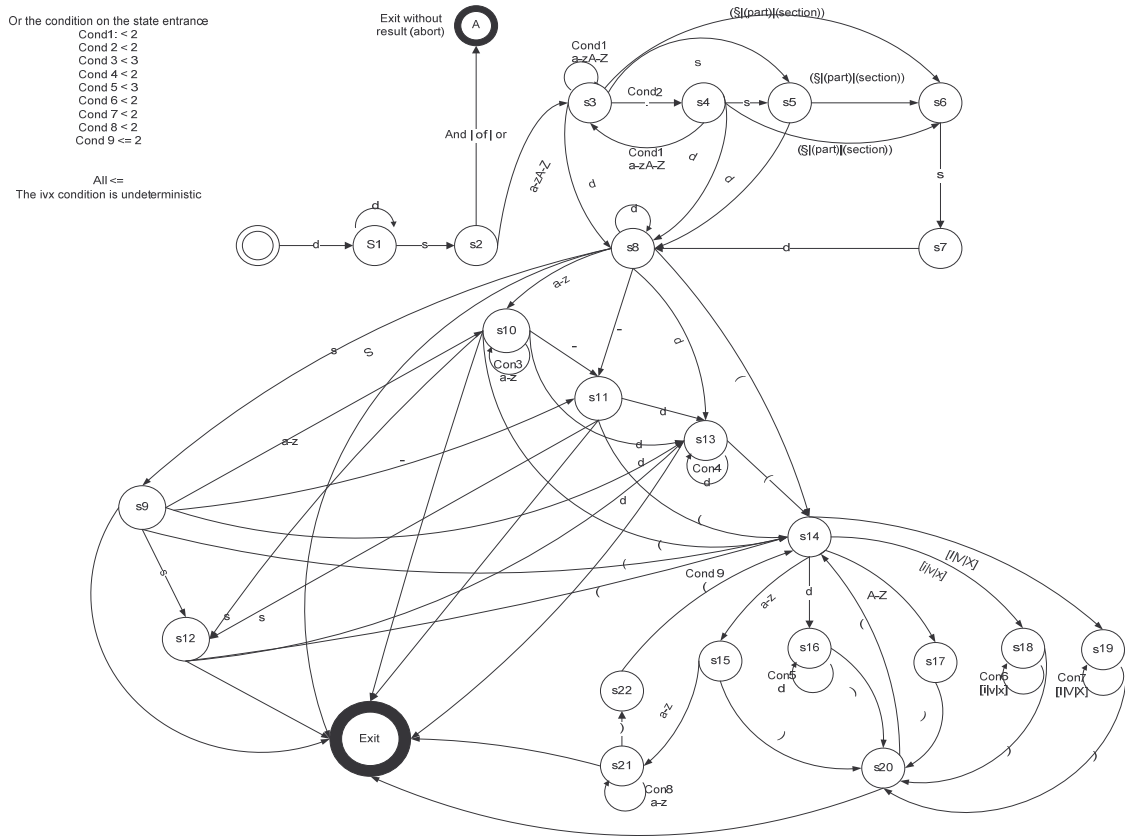


Figure 3-6 Automata for the regular expression that captures the standardized citation styles

3.3.2.3 Extracting the Relations

The next step is to determine the relations between the citing and the cited provisions. There exist many types of relations that link a citing provision to a cited one. After manually inspecting many regulations, we found that these relations can be grouped in two categories: Assertions and Amendments. A relation is considered as an assertion if it refers to a provision that is used to support the author’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: 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 further divided into subtypes including:

- Amendment by insertion: The citing provision adds more detail or completes 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 reader's attention by crossing the information about a cited provision that is no longer valid and inserting new parts or details.
- Amendment by redesignation: This occurs when the citing provision changes the name of the cited provision. The new name is then reflected in the citing provision.

Through the exploration of many regulatory documents, we have noticed that amendments appear more frequently than assertions. This is due to the fact that they are used as a legal instrument for changing the content of a particular act to adapt it to

changing circumstances. Suber et al. note that “Amendments aspire to capture the inconsistent virtues of stability and flexibility, protecting what the enacting generation thinks wise, but permitting future generations to think otherwise” (*Suber, 1999*). An act can amend itself or other acts. In this study, we pay careful attention to amendments since they are the ones that can potentially generate overlaps and conflicts (*Antoniou, Billington, & Maher, 1999*).

We can identify the relation that a citation represents by exploring the text surrounding it, searching for specific verbs such as “amends”, “defines”, etc. To extract these verbs, we use a text tagging technique based on Brill’s Transformation-based learning tagger to perform the Part of Speech (POS) tagging process (*Brill, 1995*). The objective is to divide the text into words, and assign each word its corresponding part of speech (i.e., noun, verbs, preposition, etc.), based on both its definition, as well as its context. These words can later be explored individually for the purpose of extracting the relations among provisions. An example of tagging the provision of Figure 3-7a is shown in Figure 3-7b.

<p>TITLE XI SEC.1103(b)</p> <p>TECHNICAL AMENDMENT.—Section 21C(c)(2) of the Securities Exchange Act of 1934 (15 U.S.C. 78u–3(c)(2)) is amended by striking “This” and inserting “paragraph (1)”</p>	<p>TITLE XI SEC.1103(b)</p> <p>TECHNICAL/NNP AMENDMENT.—Section/NNP 21C(c)(2)/CD of/IN the/DT Securities/NNP Exchange/NNP Act/NNP of/IN 1934/CD (*) is/VBZ amended/VBN by/IN striking/JJ “This” and/CC inserting/VBG “paragraph (1)”</p>
---	---

Figure 3-7 a) Example of a provision taken from SOX before tagging b) Example a tagged provision

We collect the verbs with direct relations to the citation of interest. Using the WordNet (*Miller, 1995*) lexical database cognitive synonyms, which is a set of one or more

synonyms that can be used interchangeably without changing the meaning of the context in which it is embedded, we are able to convert these verbs to one of the relations we discussed earlier (i.e., assertion or amendment). The morphological processor in WordNet finds the base form automatically and then returns all synonyms of the word. The last step is to match the meaning of the verb with one of the relations we already defined. An example of using WorldNet synonyms is shown in Figure 3-8., in which the synonyms of the word “Revise” (base form of “Revised”) are listed, among which the word “amend” appears, which classify the corresponding citation as an amendment

```

Word    ➔ Base  ---- (POS) synonym1 | synonym2
Revised ➔ Revise ---- (n) revise | rescript | revisal | revision
        |
        ---- (v) rewrite | retool | amend

```

Figure 3-8 Example of the matching process used in extracting the relations

The final step in building the citation graph is to generate the graph itself. Figure 3-10 shows an example of a citation graph constructed from the SOX provisions shown in Figure 3-9. In this graph, we can see three SOX provisions that amend or redefine three other provisions from different law documents, namely, 15 U.S.C 17a(s), 15 U.S.C. 78a(m), and the SA of 1933 Section 19(b).

```

SEC. 2. DEFINITIONS
(a) IN GENERAL ...
    (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));

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.—

Figure 3-9 Example of two provisions taken from SOX

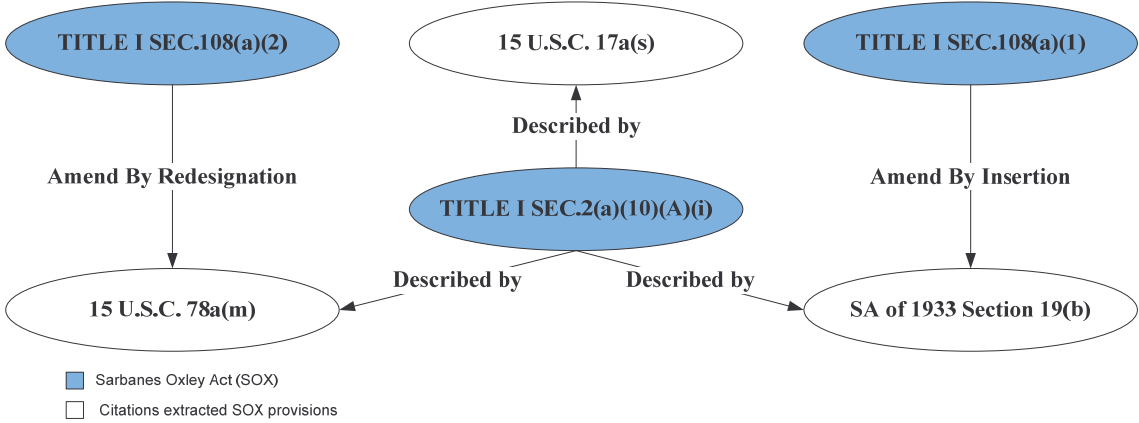


Figure 3-10 A citation graph extracted from the provisions described in Figure 4-9

3.4 Analysis of Citation Graphs

The information contained in citation graphs can be analyzed in several ways depending on the objective and the scope of the analysis. In this thesis, our objective is to facilitate the understanding and the analysis of multiple regulations. For this purpose, we have developed three analysis techniques:

- Citation Depth Limiting
- Fan-in and Fan-out analysis
- Relatedness Analysis

3.4.1 Citation Depth Limiting

As explained previously, legal citations follow a hierarchical structure called the paragraph hierarchy structure. Each citation represents the exact place of the cited information in the document. The title and the section number of a United States Code citation is enough to distinguish between different acts. Other information in the citation is needed to determine the exact provision within the act. If the objective of the analysis is to show the degree of coupling between two regulatory acts then we do not need the full citation representation, the title number combined with the section number are sufficient to draw the relation between two specific acts.

Citation Depth Limiting is one of the filtering techniques used in our approach to allow users to study the relations among regulations by specifying the citation level. In other words, by varying the amount of information that is contained in a citation; we also vary the level of granularity of the graph. For example, we can trim all citations to only the title number to focus on understanding the relationship between different acts without having to look at their provisions. Similarly, the more information used in the citation, the deeper granularity we achieve, and hence the more detailed citation graph we obtain.

3.4.2 Fan-In and Fan-Out Analysis

In our analysis, we used the concepts of fan-in and fan-out as explained in (*Yourdon & Constantine, 1979*) to measure a coupling relationship among the components of a software system. Similarly, in our analysis, we measure coupling between provisions of multiple acts. We define fan-in and fan-out as follows:

- Fan-in (c): Measures the number of provisions citing the provision c.
- Fan-out (c): Measures the number of citations c depends on (i.e., the provisions c cites).

In a citation graph, fan-in is represented by the number of incoming edges to a citation node, whereas fan-out represents the number of outgoing edges from a provision node. Our study shows that the coupling relation between different acts can be determined by analyzing the fan-in relations of the external citations, the higher the fan-in, the stronger the relation. This can help in ranking regulations that have a coupling relationship with the act under study. These supplementary documents should be read in conjunction with the act in question itself in order to extract the rules and requirements from that act (*Breaux , 2009*).

An example of measuring the importance of a provision by examining its fan-in is depicted in Figure 3-11, in which many provisions of the SOX act cite specific provisions of Securities Exchange Act of 1934, which are represented by the citation 15 U.S.C 78 followed with specific sections. This is an indication that one should study and understand the Securities Exchange Act provisions in order to understand SOX. By referring to the Securities Exchange Commission (SEC) website, we have concluded that the Securities Exchange Act is considered the parent law for SOX, as a result the two cannot be analyzed independently.

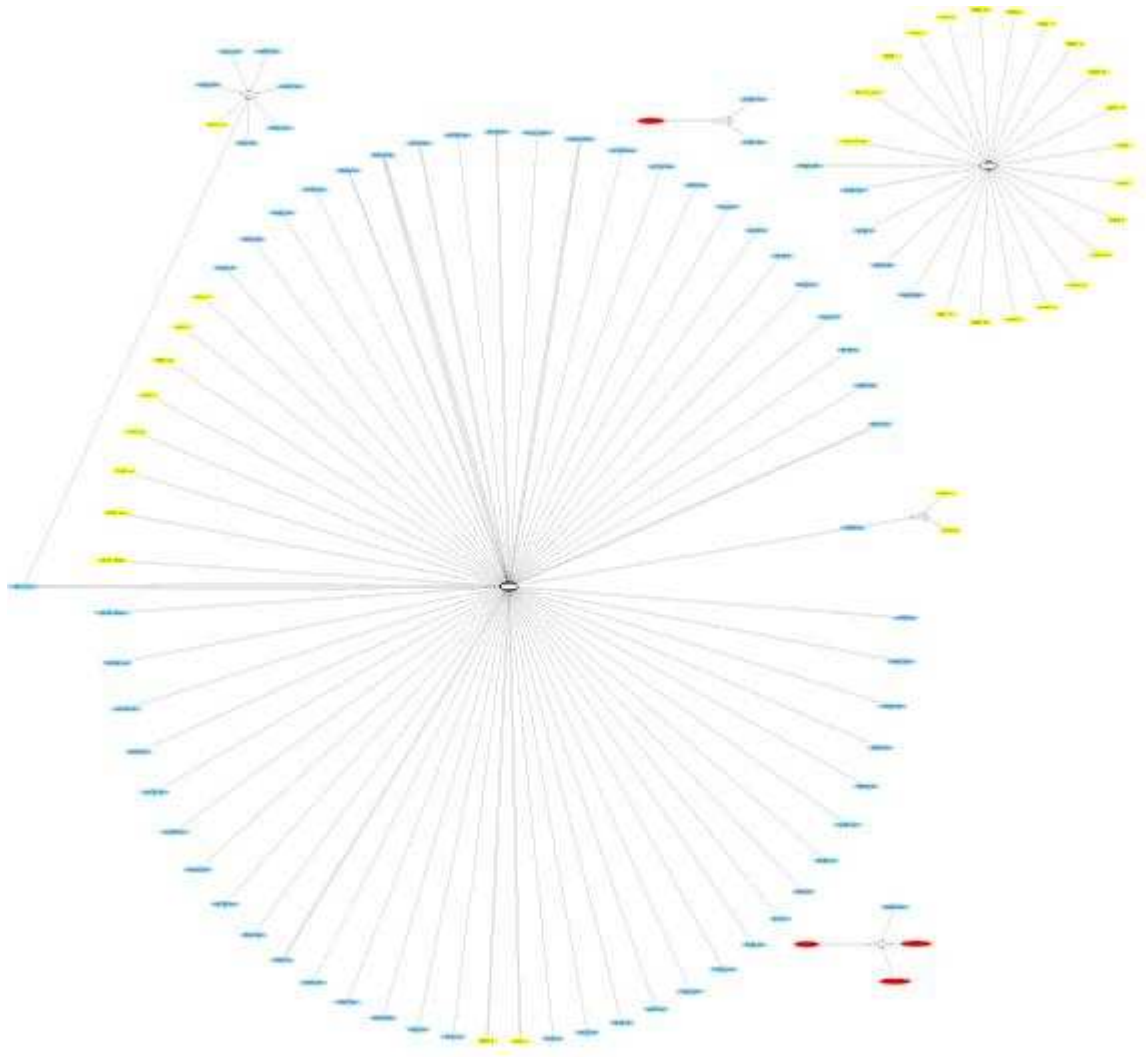


Figure 3-11 A citation graph extracted from SOX, shows (15 U.S.C 78) in the center, and a large number of SOX provisions that refer to this act

On the other hand, a high fan-out of a particular provision indicates that the provision is citing many other provisions. The more dependencies a provision has with other provisions (from the same or different acts), the higher the risk of inconsistencies. This is because the dependent provision is vulnerable to changes of any of the acts on which it depends. In our case study, we noticed that provisions that cause conflicts and overlaps are usually the same as the ones that have high fan-out.

3.4.3 Relatedness Analysis

The objective of the relatedness analysis is to detect any possible overlap or conflict between regulations. Relatedness analysis uses a top down approach to explore citation graphs, starting from studying if two or more regulations are sharing some common citations, or referring to common regulations, down to more detailed graphs that show the exact section or subsection that refer to these common regulations, and the type of relations that connect the citing and the cited regulations.

Citation relatedness analysis focuses on how different acts relate to each other. The process of applying citation analysis to detect how different acts relate to each other can be summarized in the following steps:

1. We analyze the citation graph to identify regulations that are related to each other through the citation relationship.
2. For each of the intersecting regulations, we dig one level deeper to study the provisions that relate to each other. This is because two citing provisions may refer to the same act, but to different provisions in the cited act.
3. For each related provisions, we analyze whether it is an amendment or an assertion. As previously mentioned, amendments, in particular amendments by deletion and striking, represent high risks of conflict since an act that deletes parts of another act will affect all other acts that depend on the modified act.

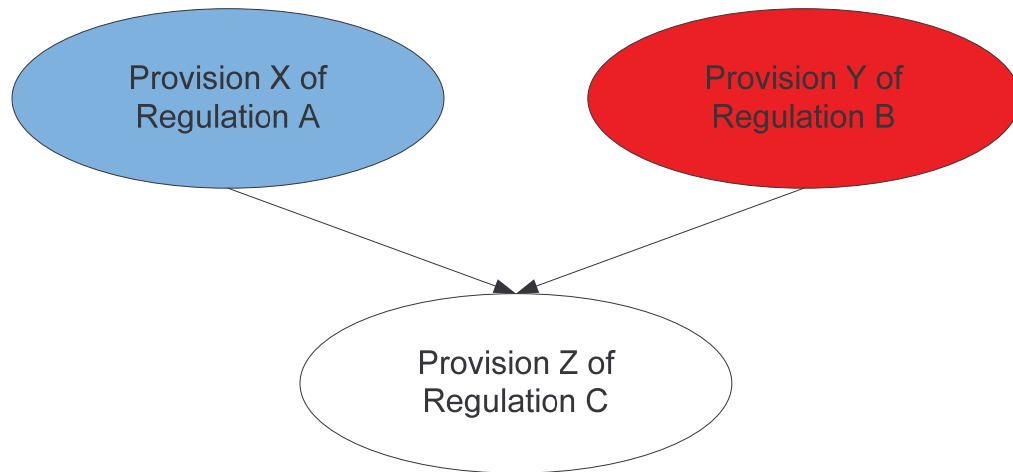


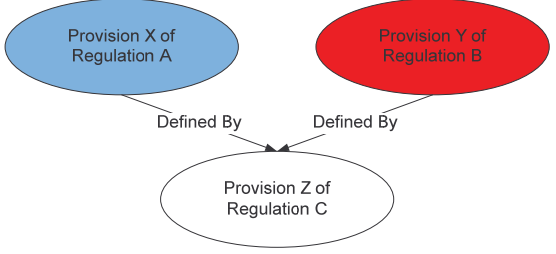
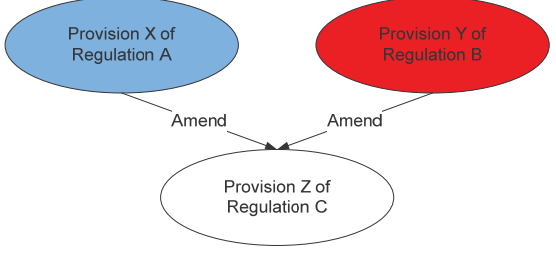
Figure 3-12 Explanation of overlaps at the provision level

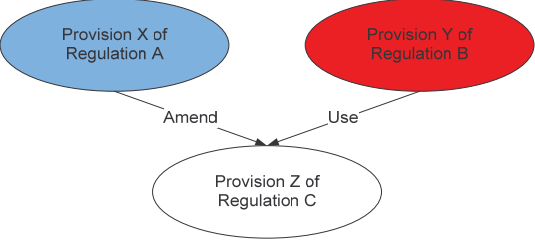
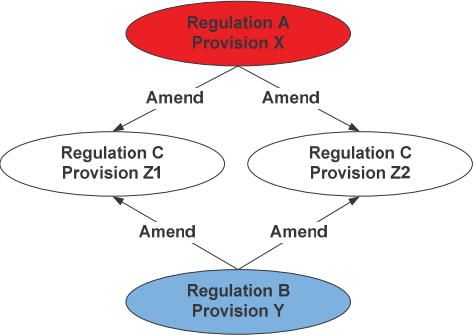
Cases of overlapping and conflicting provision happen often when two provisions of different acts (e.g. A and B) refer to the same content of another provision from a third act (e.g. C), as illustrated in Figure 3-12. An organization that is required to comply with A and B needs to pay careful attention to such cases in order to uncover the overlap and possible conflict in order to avoid recreating policies and procedures that deal with repetitive information that might be either identical (cases of overlap) or conflicting. In addition, this dependency among provisions usually leads to inconsistencies unless there is a mechanism that checks that changes made to the common provisions (Provision Z in our example – see Figure 3-12) are reflected in both provisions that depend on it. Citation graphs can be used to help build such mechanisms.

After studying several regulatory documents and the dependencies among them, we have designed a set of patterns that represent situations where possible overlaps and conflicts may occur. These patterns take into account the related provisions and the type of relations that connect them (i.e. assertion and amendment).

The following table 3.1 shows the list of identified patterns:

Table 3.1 Analysis Patterns

 <p>The diagram shows two ovals at the top: a blue oval on the left labeled 'Provision X of Regulation A' and a red oval on the right labeled 'Provision Y of Regulation B'. Arrows from both ovals point downwards to a central white oval labeled 'Provision Z of Regulation C'. Each arrow is labeled 'Defined By'. Below the diagram is the text 'Define By - Defined By Pattern'.</p>	<p>When two provisions X and Y from two different regulations A and B are respectively using the definition or the concept explained in another regulation C, this indicates a similarity between provision X and provision Y. As a result, there is a possibility to find common rules between the two provisions.</p>
 <p>The diagram shows two ovals at the top: a blue oval on the left labeled 'Provision X of Regulation A' and a red oval on the right labeled 'Provision Y of Regulation B'. Arrows from both ovals point downwards to a central white oval labeled 'Provision Z of Regulation C'. Each arrow is labeled 'Amend'. Below the diagram is the text 'Amend-Amend Pattern'.</p>	<p>When two provisions X and Y from two different regulations A and B respectively amend a provision from another regulation C, this may result in a conflict, if they modify it in a contradictory way. Amendment by redesignation does not affect the content so it not a source of conflict. However; amendment by striking, which is usually followed by an amendment by insertion or deletion, can lead to a conflict.</p>

	<p>Amendment by insertion and deletion affect the content and as a result may be a source of conflicts.</p>
 <p>The diagram shows three ovals. A blue oval on the left contains 'Provision X of Regulation A'. A red oval on the right contains 'Provision Y of Regulation B'. A white oval at the bottom contains 'Provision Z of Regulation C'. An arrow labeled 'Amend' points from the blue oval to the white oval. An arrow labeled 'Use' points from the red oval to the white oval.</p> <p style="text-align: center;">Amend-Use Pattern</p>	<p>If Provision X from an act A amends a provision Z from C and that a provision Y of act B uses the provision C then changes made by A may affect B. This may result in errors or conflicts.</p>
 <p>The diagram shows four ovals. At the top is a red oval labeled 'Regulation A Provision X'. Below it are two white ovals: 'Regulation C Provision Z1' on the left and 'Regulation C Provision Z2' on the right. At the bottom is a blue oval labeled 'Regulation B Provision Y'. Arrows labeled 'Amend' point from the red oval to both the Z1 and Z2 ovals. Arrows labeled 'Amend' point from the blue oval to both the Z1 and Z2 ovals.</p> <p style="text-align: center;">Generalized Amend-Amend Pattern</p>	<p>We have noticed through the study of many regulations that related provisions from different acts usually refer to the same set of provisions. In other words, if a provision Y from B refers to provision Z1 of regulation C, whenever the provision X of regulation A refers to Z1, then it is most likely that provision X will also refer to provision Z2 of regulation A when provision Y refers to Z2. This is a more generalized case of the Amend-Amend pattern.</p>

An example of applying these patterns (and hence relatedness analysis) to detecting conflicts among regulations is to use them to detect a well documented conflict in the area of law, in particular, between two copyright laws, The Digital Millennium Copyright Act (DMCA) of 1998, and the Technology Education and Copyright Harmonization Act (TEACHA) of 2002. The problem was first reported in 2003 by a group representing the college media centers, which sent a warning message to the U.S. Copyright Office about a possible conflict between these two federal laws.

The conflict consists of the fact that while DMCA restricts access to the electronic copyrighted material, TEACHA allows access to the same material under the “fair use” cover for pedagogical purposes, more specifically for online education and distance learning. The “fair use” umbrella aimed to solve the problem if TEACHA did not restrict the conversion of materials from analog to digital format, which will be considered a violation of the anti-circumvention provision which was added to the copyright act by the DMCA.

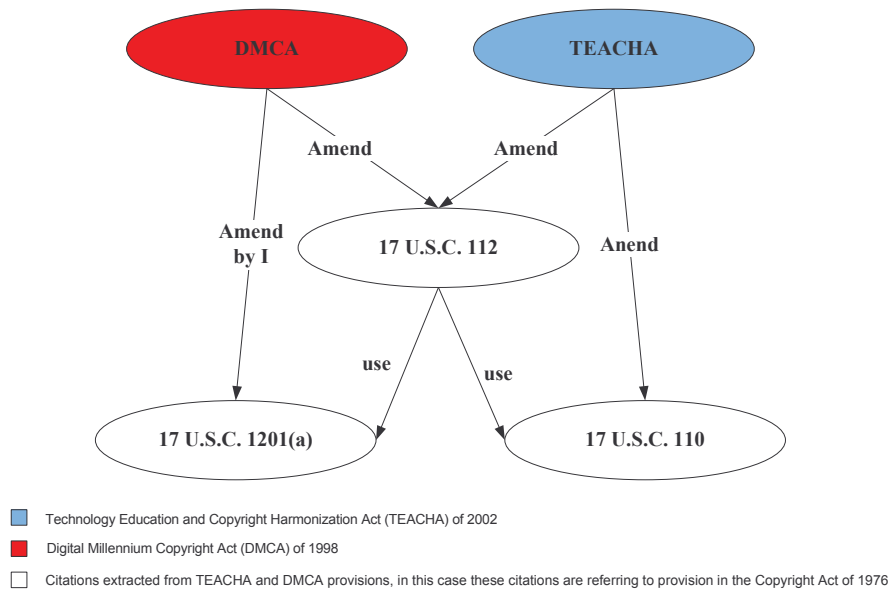


Figure 3-13 Citation graph that shows a situation where a possible conflict between DMCA and TEACHA may exist

We applied citations relatedness analysis on DMCA and TEACHA to uncover this conflict. Figure 3-13 shows a partial graph representing the provisions related to access to the electronic copyrighted material in both laws. The graph combines the first three patterns explained before. Using this graph, we can see that TEACHA amended Section 110 as well as Section 112 of the Copyright Act Title 17 of the U.S. code, at the same time the DMCA amended Section 112 and added a new section, Section 1201 to the Copyright Act. Based on citation relatedness analysis, this situation represents a high risk of non-compliance due to the dependencies between Sections 112, 110 and 1201. Further analysis is needed to see if this dependency yields a conflicting situation.

According to the graph the Copyright Act of 1976 Title 17 of the U.S. code was amended by both DMCA and TEACHA. The Copyright Act is an old act; hence it does not cover the issues brought up by the new technologies such as protection of digital content. The objective for TEACHA amendment of Section 110 (Limitations on exclusive rights:

Exemption of certain performances and displays) is to allow instructors who work in non-profit educational institutions to use portions of movies, or audiovisual work in online courses without looking for permission or paying fees to copyright holders. The specified amendment clarifies the conditions under which such usage will not be considered as infringement of the copyright.

DMCA added Section 1201(a) which is referred to as the anti-circumvention provision. The provision makes it illegal to circumvent technologies that prevent access to copyrighted material such as ripping Macrovision or Content Scramble System DVDs. On the other hand the DMCA released some of the liabilities of transmitting organizations and added a statutory license subsection to Section 112.

TEACHA allows using the copyrighted materials in distance learning, as a result there is a need to have the copyrighted materials in digital format. DMCA restricts the circumvention of technologies and converting materials from analog to digital format. There is a need for some exceptions for distance learning so that TEACHA can achieve its goals. The amendment by TEACHA to Section 112 - Limitations on exclusive rights: Ephemeral recordings - restricted the conversion of materials from analog to digital format unless the following conditions hold.

*“(A) no digital version of the work is available to the institution; or
(B) the digital version of the work that is available to the institution is subject to technological protection measures that prevent its use for section 110(2).”¹*

¹ Subparagraph (A) and (B) of Subsection (f) of Section 112 of title 17, United States Code as amended by TEACHA

These conditions complicate the problem especially that not all digital material that is protected has a non-digital version that educational institutions can convert to make use of TEACHA.

The contradictions between the two copyright laws did not provide the instructor's with a safe path while practicing distance learning, and kept the liability of their actions up to the judges' decision based on the "fair use" act.

3.5 Compliance Decision Support System (CompDSS)

The Compliance Decision Support System (CompDSS) is a prototype tool that supports the framework presented in this chapter. Using CompDSS, an analyst can load multiple regulations for exploration, display the list of provisions they contain, generate composite citation graphs to investigate the relationships among regulations, and apply the analyses techniques discussed earlier.

3.5.1 CompDSS Architecture

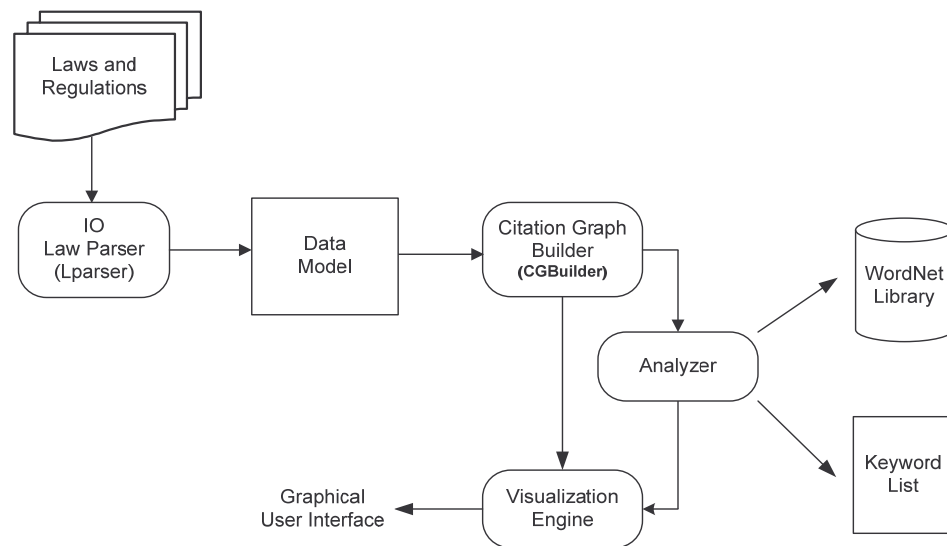


Figure 3-14 Overall Architecture of CompDSS components that support citation analysis

3.5.2 Law Parser

The role of the parser is to extract the provisions and their constituents from an input law document and build a data model that can later be processed by the other components.

This requires some pre-processing steps to clean up the input documents, which are usually saved in unstructured formats such as PDF and HTML, by removing headers, footers and other unnecessary data. The parser relies on the set of regular expressions discussed in Section 3.3.2.1 and which are designed specifically to recognize the various elements of a North-American law. The user has, however, the flexibility to modify these regular expressions to adapt them to other laws.

The steps bellow illustrates how the Law Parser in CompDSS works:

Step1: Read the file that contains the regulation Line by Line and store the result into a list

Step2: Build the paragraph hierarchy by identifying indentations. In this step, the law parser creates a list of regular expression to capture the paragraph hierarchy patterns which represents the set of all indents used in the document. For example, the expression $\backslash(\backslashd{1,2})\backslash$ is equal to one or two digits between two brackets which represents a subsection according to the paragraph hierarchy structure. The tool provides the user with the default regular expressions list according to United States Congress Data Dictionary of Legislative Documents and allows the user to add his or her own regular expressions.

The snapshot in Figure 3-15 shows the paragraph hierarchy builder screen. The tree on the right hand side of the snapshot represents the default set of regular expressions that works with all North American laws that follow the standard paragraph hierarchy. However, as the left hand side shows, the user has the capability to modify this list, or create his own list to capture any other document structure.

Step 3: Clean the document from table of contents, footers, headers, empty lines and other unnecessary information. The program also allows the user to add any pattern they want to delete from the whole document.

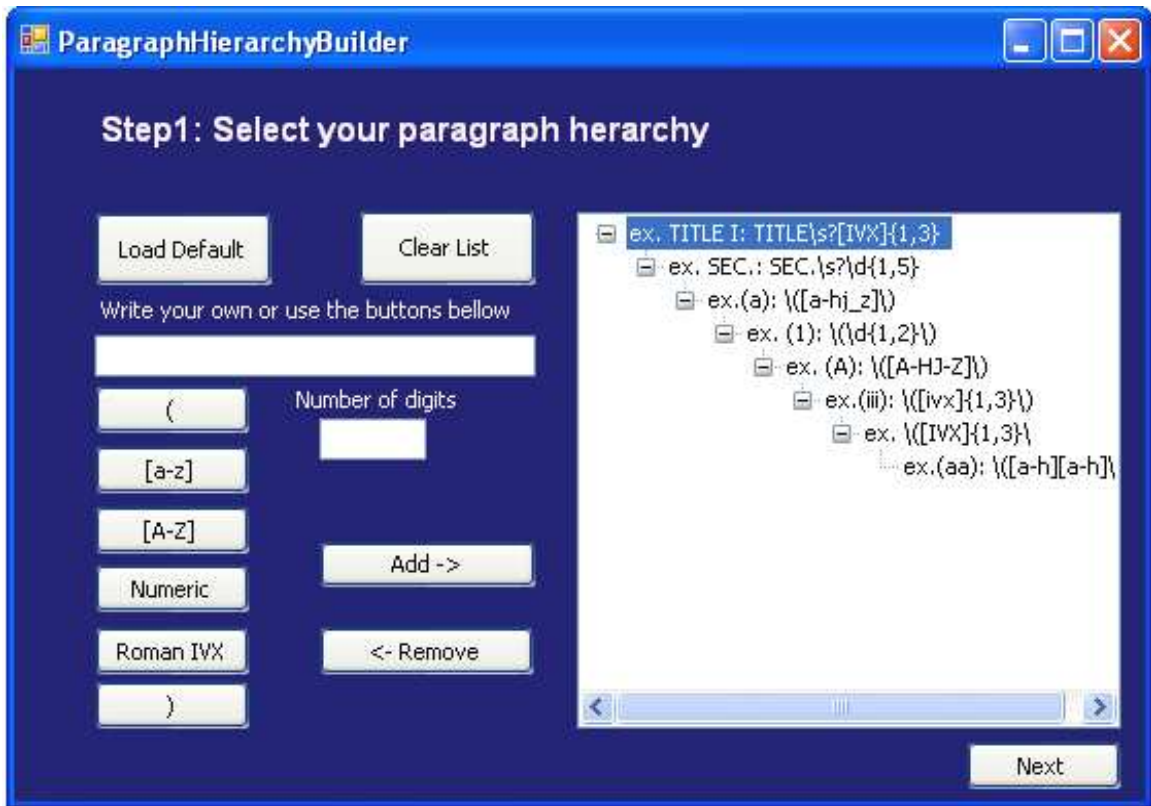


Figure 3-15 A snapshot of one screen from the CompDSS tool, representing the paragraph hierarchy builder (one phase in the LParser)

Step 4: Make each provision as one Line

The content of a provision is converted into one line to facilitate its processing. For example, this would help extract citations in the correct format since a citation may be distributed into more than one line.

To convert the whole provision paragraph into one line, we check if the line is a ProvisionStartingLine or not. Where a ProvisionStartingLine is a line that starts with a

paragraph hierarchy indent. If the line is not a ProvisionStartingLine we concatenate it with the previous line and so on.

```
FOR EACH RegularExpression X IN paragraphList
  IF X has a Match in (line) AND IF Match is the first word in the
  (line)
    Then The line is ProvisionStartingLine
  Else
    The line is NOT PovisionStartingLine
```

Figure 3-16 A procedure to detect the start and the end of a provision then converting the whole provision to one line

The output of this step is a document, where each line in this document representing a provision.

Step 5: Create provision identifiers

This is the last step in the LParser, which converts the document into a structured format, where each provision will be assigned a provision identifier. The provision identifier is the composite of all the indents that form a full and unique path of that provision in the document (its exact location). For example the provision identifier “Title I. Sec.102(a)” of SOX determines how the reader should navigate through the document to reach that specified provision.

We create provision identifiers and assign them to the corresponding provisions for easier navigability by applying the algorithm in Figure 3-17.

Definitions:

- `regexListIndex`: In step 2 we created a paragraph hierarchy List the `regexListIndex` is the index of that list.
- `provosionCarrierList`: A list that stores the indents and the level of the indent starting from level zero. This list is used to build the provision later using the `createProvision` function. The `provisionCarrierList` is a temporary list. The value of count depends on the depth of the paragraph hierarchy in a specific line.

```
For each provision
  Set regexListIndex to 0
  For each RegularExpression X IN paragraphList
    Find a matche IN the line and store it IN the MatchString
    regexListIndex = the index of the matched pattern
    IF MatchString is the first word IN the line and not empty
      IF regexListIndex == The number of items IN the provosionCarrierList
        Insert the MatchingString in provosionCarrierList
      END IF
      IF regexListIndex < number of elements IN the provosionCarrierList
        WHILE
          regexListIndex < number of elements IN the provosionCarrierList
            provosionCarrierList.RemoveAt (provosionCarrierList.Count - 1)
          END WHILE
        Replace the last element IN provosionCarrierList with
        MatchString
      ELSE
        Error in the document hierarchy
      END IF
      Call createProvision(line, MatchString, regexListIndex)
    END IF
```

Figure 3-17 An algorithm to create provision identifier

The `createProvision` function assigns values to each member of the provision object. It saves the depth level of that provision, separates the indent from the provision text, saves the text of the provision, and calls the `getFullProvisionName` function which concatenates all the indents in the `provisionCarrierList` to form unique provision identifiers.

Step 6: save the result in XML format, and present it in the program in tree format

The output of the parser is an object model that instantiates a simple and yet expressive data model that we built to characterize the content of law documents. The data model is shown in Figure 3-18 in the form of a UML class diagram. In this model, an act is composed of several provisions, which contain various elements organized in a hierarchical way including sections, subsections, clauses, sub-clauses, and so on. This hierarchical structure is captured using the composite design pattern (*Gamma, Helm, Johnson, & Vlissides, 1994*) and is modeled using three classes, namely, `ProvisionElement` (abstract class), `LeafElement`, and `CompositeElement`.

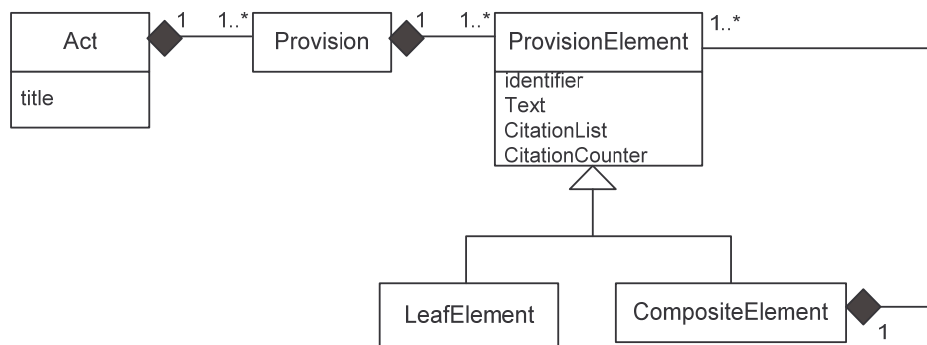


Figure 3-18 A UML class diagram of the CompDSS data model.

3.5.3 Graph Builder (CGBuilder)

The citation graph builder component (CGBuilder) operates on the object model to construct the citation graph. For this purpose, we implemented the regular expression automata in Section 3.3.2.2 which allows extracting citations that follow the Bluebook and ALWD standards. The user has the flexibility to modify this regular expression to adapt it to other citation style standards. The object model in Figure 3-18 shows that each provision has a CitationList, which is a list of all the citations embedded in the provision text. The CGBuilder checks each provision text for matches, citations that match the regular expression are then added to the CitationList of that provision.

As explained in Section 3.3.2, the next step in building the citation graph after extracting the citations is extracting the relations between the provisions and the citations embedded between their lines. In Section 3.3.2, we proposed using a Part of Speech (POS) tagging technique in order to find the verbs that have direct relations with the intended citation. The approach for finding the relations can be summarized in the steps illustrated in Figure 3-19.

In Steps 1 and 2, before the tagging process starts, we replace all citations by a (*) symbol so they can be treated as one word. We push the replaced citations in a stack to maintain the order so as to retrieve them later after the tagging process is complete.

Each provision is partitioned into a list of words. After that we apply a text tagging technique based on Brill's Transformation-based learning tagger to perform the Part of Speech (POS) tagging process (*Brill, 1995*).

1. For each provision in the act, replace citations in the provision text with stars (*).
 2. Push the replaced citations in a stack so that it can be retrieved later in the right order.
 3. Divide the provision text into a list of tokens
 4. Apply POS tagging on tokens
 5. Find the base form of each word by applying word stemming
 6. Apply word sense disambiguation to find the most appropriate sense for each word in the sentence
 7. Extract the verbs surrounding the star symbols, and reinsert the corresponding citation to the right place.
 8. For each verb find all its synonyms using WorldNet
- Based on the synonyms and the predefined relations list, match each verb to one of the relations.

Figure 3-19 The steps to determine the relationships between the Act provisions and the citations embedded in these provisions

The natural language processing (NLP) literature provides different approaches for POS Tagging. Van Guilder in her handout for LING 361 (*Van Guilder, 1995*) classified the different POS Tagging models. Figure 3-20 shows a slightly modified version of that classification. POS taggers can be either supervised, which needs a pre-annotated corpus for training to learn how to tag similar situations in the future, or unsupervised which

does not need corpus for training, and use advanced computational techniques instead (i.e. Baum-Welch algorithm) (Huang, Acero, & Hon, 2001) (Baum, Petrie, Soules, & Weiss, 1970). Both supervised and unsupervised techniques can be further classified into rule based, stochastic, or neural techniques.

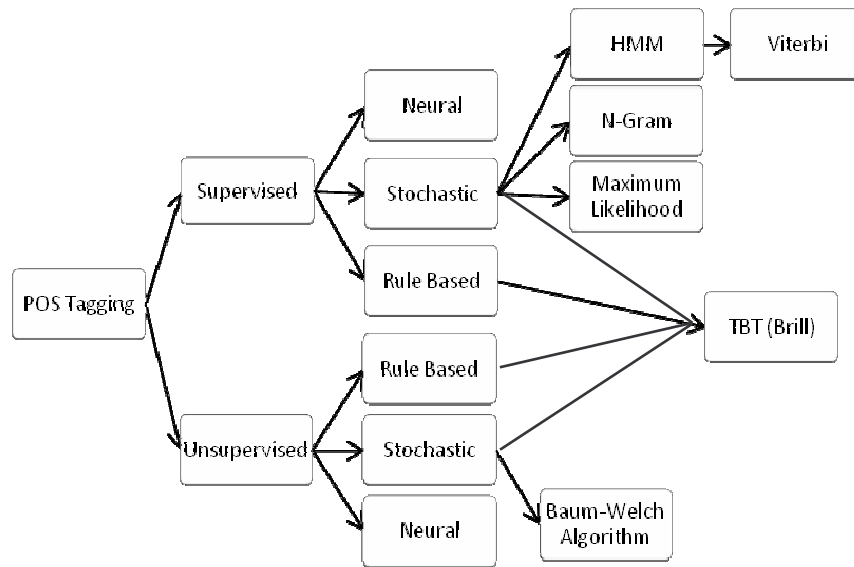


Figure 3-20 Classification of POS taggers based on (Van Gulder 1995) classification

Unsupervised rule based taggers use hand written rules (context frame rules), while the supervised ones use automatically generated rules based on the corpus during training, to assign each word a single grammatical interpretation (tag) (Pradeep Varma, Rakesh, & Sanyal, 2007) (Brill E. , 1992).

Stochastic taggers use statistics, probability, frequency and mathematical formulas and equations. Many stochastic methods were developed to disambiguate tags like Hidden

Markova Model (HMM), Tag Context Model (N-Grams) and Maximum Likelihood Estimation (MLE) (*Pradeep Varma, Rakesh, & Sanyal, 2007*).

The third and last tagging approach is the use of artificial intelligence neural network for tagging (*Schmid, 1994*).

We implemented the Brill's transformation-based learning tagger algorithm described in (*Brill E. , 1995*) to be able to perform the part of speech (POS) process and extract the verbs that surround the citations. Brill's transformation tagger is a supervised hybrid technique that combines both the rule-based and stochastic taggers. It uses predefined rules as rule based, but also stochastic-like methods to induce new rules from the data. Brill tagger requires tagged training corpus to achieve this purpose. There are many training corpus available, the most famous ones are the Brown Corpus, the British National Corpus, the Wall Street Journal corpus, the New York Times corpus which is part of the Gigaword corpus, and the Reuters News corpus. In our case we used the Brown and Wall Street Journal corpus.

Another important aspect of the CGBuilder component is that it accesses the WorldNet library to retrieve synonymous words. WordNet is a freely and publicly lexical database that provides a large repository of English lexical items. It establishes connection between four types of (POS), noun, verb, adjective and adverb (*WordNet 3.0 Reference Manual 2009*). WordNet groups all words that correspond to a specific meaning together in Synonym-Sets (synset). The concept that a synset represents is called the Gloss, and the meaning of a word under particular (POS) is called the sense. Synsets are connected to each other through semantic relations such as those explained in Table 3.2.

Table 3.2 Some of the relations between synsets

Relation	Description	Example
Hyponym - Hypernym	is kind of relation	Salmon and Cod are Hyponym of Fish
Troponym- Hypernym	is way to do relation	To trim and to slice are Troponyms of to cut
Meronymy-Holonymy	is part of relation	Trunk is part of tree
Antonymy - Synonymy	Has opposite meaning	up is Antonym of down

Knowing the POS of the words surrounding the citation can help retrieve the corresponding synset (sense) from WordNet, however, to do so we first need to retrieve the base form of the verb. We applied a porter stemming algorithm as described in (*Porter, 1980*) to obtain the base form of the verb.

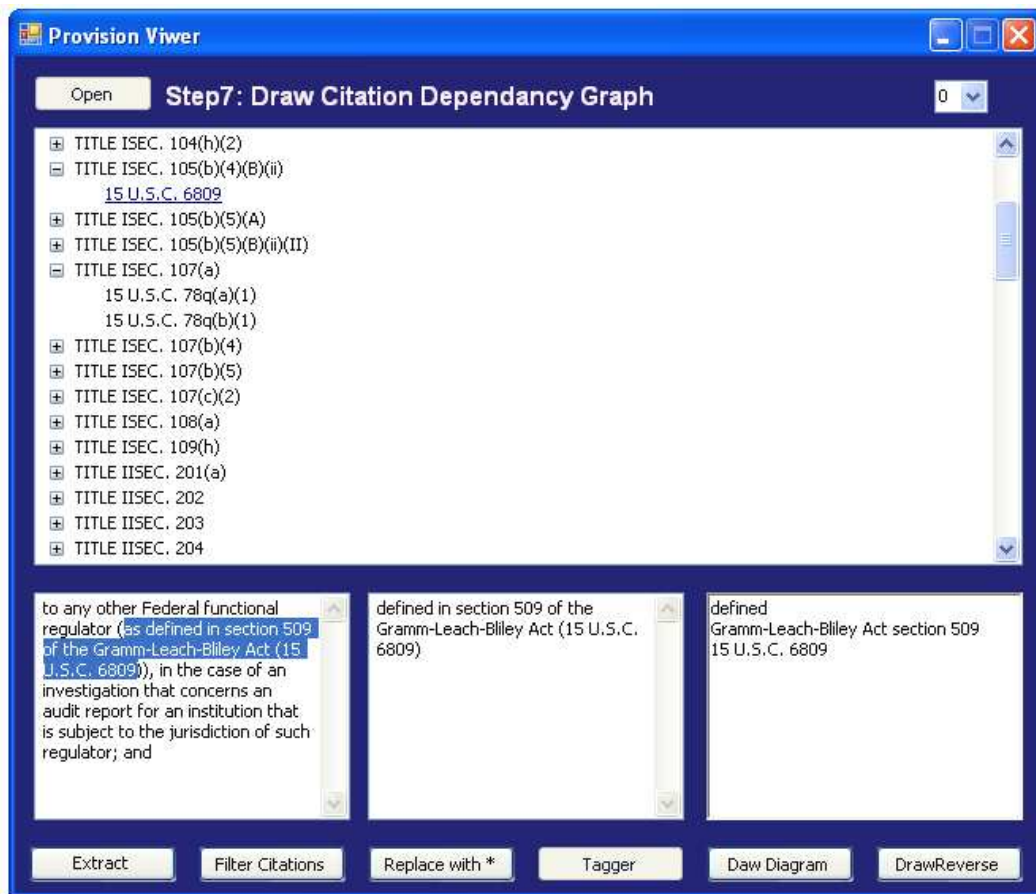


Figure 3-21 A screen snapshot of the CompDSS tool that shows the extracted provision of the tagging process

The verbs and their synonyms are then matched to a list of keywords that we created to be able to classify the citation types into either assertions or amendments. At any time, the user can update the keyword list by adding, modifying, or deleting words.

In order to fully automate the verb matching step, we applied an algorithm that is similar to Michael Lesk's algorithm (*Lesk, 1986*) to give a sense score in cases where the synset of the verb does not directly match one of the predefined relations. In such cases, we count the number of words that are shared between the gloss of the synset and the definition of each of the predefined relations. The relation is then determined based on the higher sense score.

Figure 3-21 shows a screen snapshot where the provisions extracted with the parser are shown on the main screen. The lower view panels are used for tagging the document to prepare it for the graph building phase.

3.5.4 Analyzer

The role of the Analyzer component is to compute statistical information about the regulations as well as the citation graphs such as the number of provisions in a regulatory document, the number of provisions that are common between specific laws, the number of provisions on which a particular provision depends on (its outgoing edges), the number of provisions that depend on a particular provision (its incoming edges), etc. The objective is to guide the analysts when browsing large citation graphs by indicating places of interest based on statistical data.

3.5.5 Visualization Engine

The Visualization Engine relies on the Graphviz library (*Graphviz - Graph Visualization Software*) to draw the citation graphs. We used various visualization techniques to present the graph in a usable way such as color-coding techniques to distinguish among various regulations, the ability to vary the level of the details displayed in the graph by, for example, changing the tree depth of the provision hierarchical structure, and so on. At a very high-level, the tool can show a graph that exhibits only the name of the studied regulations and the relationship among them. The other extreme will be to show a detailed graph where all the provisions and the relationship among them are displayed. We expect the analysts to vary this level of detail based on the objective of the analysis and their knowledge of the regulations.

Figure 3-22 shows the citation graph generated from processing SOX, HIPAA, and GLBA. The graph is filtered to show only the intersections among these regulations. The blue color represents SOX provisions, the yellow color represents GLBA provisions while HIPAA is represented using the red color. The white ovals are those citations embedded in the provisions.

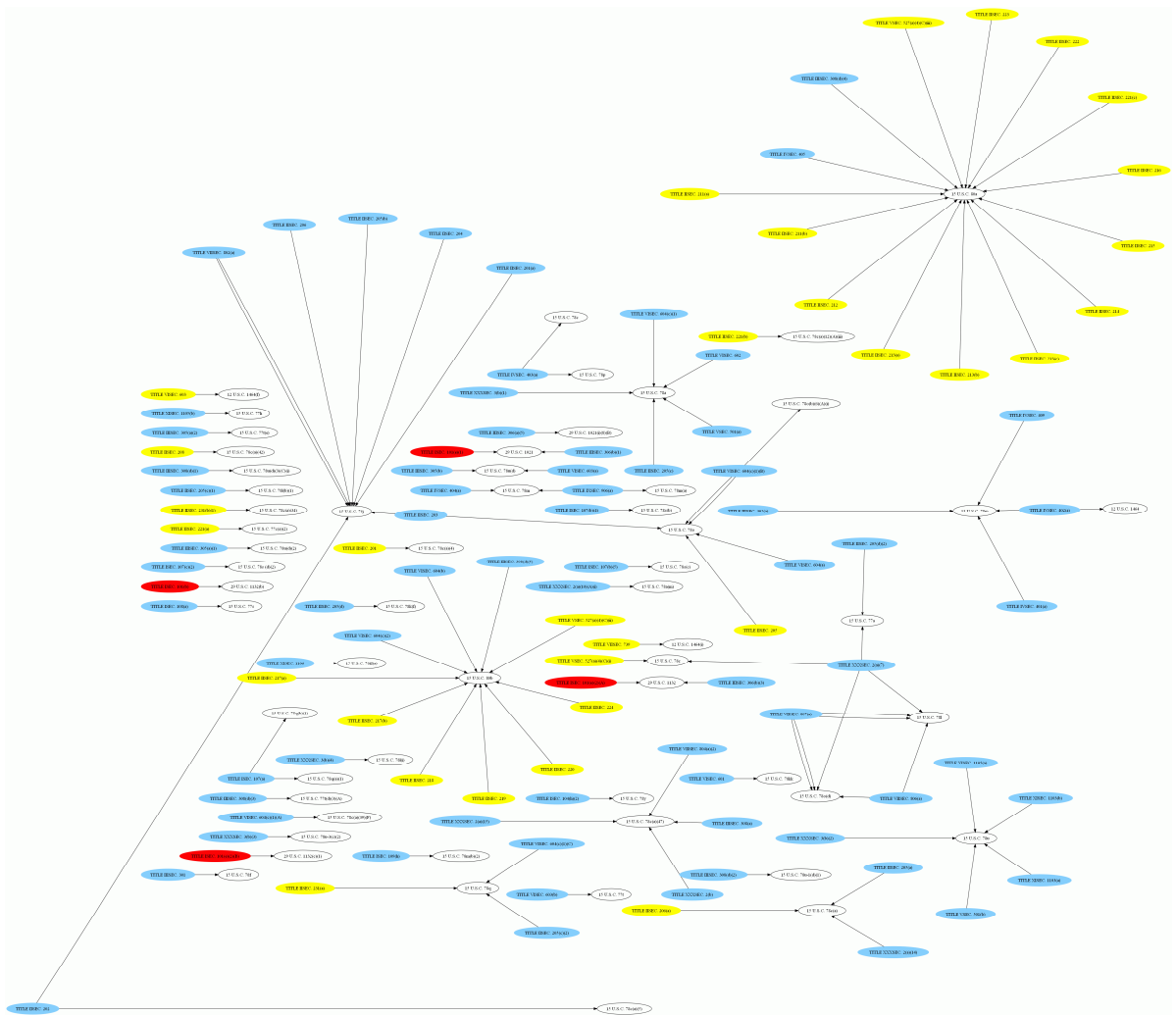


Figure 3-22 A citation graph that shows the relationship between SOX, GLBA and HIPAA

Chapter 4: Application

"Key to simplifying compliance management and reducing costs is leveraging the appropriate tools and technologies,"

Brandon Dunlap

4.1 Target Regulations

In this section, we discuss the target regulations which consist of three regulations namely, the Gramm-Leach-Bliley Act (GLBA), The Health Insurance Portability and Accountability Act (HIPAA), and The Sarbanes-Oxley Act (SOX). GLBA and SOX are two acts used in the financial sector. HIPAA is used to regulate how information such as patient records should be handled by health providers. However, all of these acts contain provisions that protect peoples' private information whether the information is used by health or financial institutions (*Kairab, 2004*). When regulators created HIPAA they put into consideration that HIPAA should not interfere with any privacy or security rules from GLBA. Each of these acts is discussed further in what follows:

GLBA:

The Financial Services Modernization Act, which is also known as the Gramm-Leach-Bliley Act (GLBA) was enacted in November 1999 to allow better synergy among

financial institutions including commercial banks, investment companies, brokerage firms, and insurance companies (*Kairab, 2004*). As a financial act, most of GLBA's seven titles are concerned with financial issues. However, GLBA has another key element which is concerned with privacy and security requirements for personal financial information. Title V of the act focuses on the fact that companies must have a complete security program to protect their customers' financial information, as noted in Section 501 of the act "...each financial institution has a continuing obligation to respect the privacy of its customers and protect the security and confidentiality of those customers' non-public personal information" (*Herrmann, 2007*).

GLBA contains several provisions that govern the way organizations should handle the security of their customers' financial records. For example, GLBA mandates financial institutions to have a comprehensive information security program based on standards to enforce technical and physical safeguards for protecting the information of customers, to ensure confidentiality of customer records, and protect the information from loss, damage, or fraudulent accessibility from none authorized people or parties.

HIPAA:

The Health Insurance Portability and Accountability Act (HIPAA) was enacted in August 1996 as part of the healthcare reform during President Clinton's second term to ensure security, privacy and portability of medical insurance and health information (*Bhandari & Computing, 2006*). HIPAA consists of five titles that cover the standardization of healthcare delivery process in an attempt to make it more efficient, setting rules to ensure the privacy of patient information and setting standards for securing this information against abuse or other electronic risks. HIPAA Title II, in particular, refers to the

“Administrative Simplification” provisions that address five security and privacy rules that must be financed by healthcare institutions. These rules ensure proper protection of patient electronic records against attacks and fraudulent use. HIPAA security requirements are mapped to the ISO 17799 standard (Herrmann, 2007).

SOX:

The Sarbanes-Oxley Act (SOX), which is also known as the Public Company Accounting Reform and Investor Protection Act was enacted in July 2002. This act was a reaction to a series of corporate scandals due to false reporting of financial and accounting statements (*Bhandari & Computing, 2006*). SOX requires from organizations to have an internal control framework to monitor the processes that have direct impact on financial reporting. Many aspects of this internal control framework impact the way financial and accounting records are handled from the security perspective. In order for auditors, for example, to be able to assess the adherence to SOX requirements, they must first understand the internal control framework of the organization in question. This requires studying the access control of the applications used to ensure authenticity, security and integrity of the information flow between systems, the confidentiality of electronic records, non-repudiation of electronic signatures, etc. The information collected in an internal control report can be seen as requirements for information security programs and can help in security assessment.

4.2 Applying Citation Analysis

4.2.1 Building the Citation Graph

Figure 3-22 shows the citation graph extracted from SOX, GLBA, and HIPAA. The graph also shows the relationships among these regulations through the provisions from other acts they refer to.

4.2.2 Result of Applying Citation Analysis to GLBA, HIPAA and SOX

In this sub-section, we present the quantitative and qualitative results of applying citation analysis to SOX, GLBA, and HIPAA.

4.2.2.1 Quantitative Analysis

As shown in Table 4.1, SOX and GLBA have four regulations in common. The regulations are named using their corresponding U.S. code (U.S.C), which is a U.S. repository where every legal document is codified. For example, “15 U.S.C. 78” refers to the Title 15 of the Commerce and Trade Act, which is Securities Investor Protection Act Section 78. SOX also intersect HIPAA in two regulations.

Table 4.1 Common regulations between SOX, GLBA, and HIPAA

Citation	GLBA	SOX	HIPAA
12 U.S.C.1464	2	1	0
15 U.S.C.77	1	7	0
15 U.S.C.78	10	69	0
15 U.S.C.80	20	5	0
29 U.S.C.1021	0	2	1
29 U.S.C.1132	0	1	3

A closer look at the provisions that are commonly cited by these regulations Table 4.2 shows that despite the fact that most provisions of SOX and a large number of GLBA provisions refer to “15 U.S.C. 78”, there are only four provisions of this act that they share in common, which are “15 U.S.C. 78q”, “15 U.S.C. 78o”, “15 U.S.C. 78c”, and “15 U.S.C. 78c(a)”.

Table 4.2 Common regulation sections between SOX, GLBA, and HIPAA

Citation	GLBA	SOX	HIPAA
15 U.S.C. 78q	1	2	0
15 U.S.C. 78o	2	3	0
15 U.S.C. 78c	1	1	0
15 U.S.C. 78c(a)	1	2	0
15 U.S.C. 80a	13	2	0
15 U.S.C. 80b	7	3	0
29 U.S.C. 1021	0	1	1
29 U.S.C. 1132	0	1	1

Table 4.3 shows statistical information extracted from the three acts. We can see that amendments represent an average of 75% of the total number of relations with external provisions, which supports the fact that any analysis techniques (such as the ones we propose in this thesis) need to focus on studying amendments.

Table 4.3 Statistical information about the three regulations

	SOX	HIPAA	GLBA
Number of External Citations	110	86	128
Number of Pages	66	169	144
Number of Provision Contain citations (full provision)	80	71	124
Number of Provision Contain citations (Section level)	43	29	74
Number of Amendments	63	66	118

Table 4.4 shows the result of applying fan-out analysis to the three regulations. The table is sorted in a descending order based on the fan-out.

Table 4.4 Fan-out analysis

SOX	fan-out	HIPAA	fan-out	GLBA	fan-out
TITLE VI SEC. 604	11	TITLE I SEC. 101	13	TITLE VI SEC. 606	13
TITLE 0 SEC. 2	9	TITLE II SEC. 231	11	TITLE I SEC. 121	8
TITLE II SEC. 205	7	TITLE IV SEC. 421	6	TITLE I SEC. 107	6
TITLE IV SEC. 402	7	TITLE II SEC. 215	5	TITLE II SEC. 231	4
TITLE III SEC. 308	6	TITLE II SEC. 262	5	TITLE I SEC. 103	4
TITLE I SEC. 107	5	TITLE II SEC. 211	4	TITLE I SEC. 104	4
TITLE VII SEC. 807	5	TITLE I SEC. 102	4	TITLE VI SEC. 604	4
TITL 0 SEC. 3	4	TITLE II SEC. 214	3	TITLE II SEC. 213	3
TITLE III SEC. 306	4	TITLE II SEC. 221	3	TITLE II SEC. 221	3
TITLE VII SEC. 807	4	TITLE II SEC. 205	3	TITLE I SEC. 102	3
TITLE I SEC. 105	3	TITLE II SEC. 211	3	TITLE V SEC. 527	3
...		

Table 4.5 shows fan-in analysis for the three regulations. The table is sorted in descending order based on the fan-in.

Table 4.5 Fan-in analysis

SOX	fan-in	HIPAA	fan-in	GLBA	fan-in
15 U.S.C. 78	70	42 U.S.C. 1320	23	15 U.S.C. 80	20
15 U.S.C. 77	7	42 U.S.C. 1395	15	15 U.S.C. 78	10
15 U.S.C. 80	5	42 U.S.C. 300	8	12 U.S.C. 1843	8
12 U.S.C. 1813	3	42 U.S.C. 1301	5	12 U.S.C. 1841	5
15 U.S.C. 6809	2	8 U.S.C. 1481	3	12 U.S.C. 1844	4
29 U.S.C. 1021	2	22 U.S.C. 2504	3	12 U.S.C. 1811	4
15 U.S.C. 1602	2	29 U.S.C. 1132	3	12 U.S.C. 371	4
5 U.S.C. 552	1	42 U.S.C. 242	3	12 U.S.C. 1430	4
15 U.S.C. 17	1	29 U.S.C. 1144	2	12 U.S.C. 1467	3
15 U.S.C. 1637	1	29 U.S.C. 1161	2	15 U.S.C. 1693	3
29 U.S.C. 1002	1	29 U.S.C. 1021	1	15 U.S.C. 1011	2
29 U.S.C. 1131	1	29 U.S.C. 1022	1	517 U.S. 25	2
29 U.S.C. 1132	1	29 U.S.C. 1024	1	12 U.S.C. 1842	2
12 U.S.C. 1464	1	29 U.S.C. 1003	1	12 U.S.C. 1821	2
12 U.S.C. 375	1	29 U.S.C. 1136	1	12 U.S.C.1464	2
...		

4.2.2.2 Qualitative Analysis

We analyzed the provisions that are commonly referred to by three regulations. We found many cases of overlaps and potential risks of consistency issues that can lead to conflicts.

We chose to report on two cases that are most representative to our findings.

The first situation is with respect to the relationship between provision 306(b)(3) of SOX and provision 101(e)(2)(A) of HIPAA that is depicted in the citation graph shown in Figure 4-1 and Figure 4-2, and which consists of the fact both these laws amend Section 502 of The Employee Retirement Income Security Act of 1974 (code 29 U.S.C. 1132). The amendment consists in both cases of striking and inserting paragraphs, in particular to Subsections (c) and (a)(6) of Section 502.

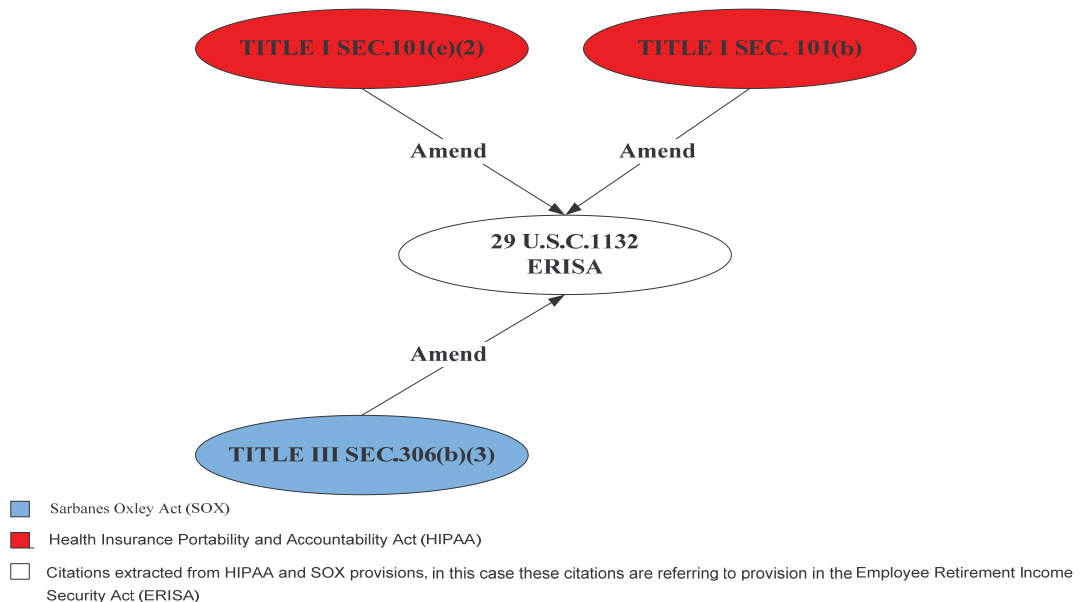


Figure 4-1 A graph showing SOX and HIPAA provisions amending the same provision of the 29 U.S.C 1132 law

Case 1: HIPAA and SOX

SOX - SEC. 306. INSIDER TRADES DURING PENSION FUND BLACKOUT PERIODS.

(1) IN GENERAL.—Section 101 of the Employee Retirement Income Security Act of 1974 (**29 U.S.C. 1021**) is amended by *redesignating* the second subsection (h) as subsection (j), and by *inserting* after the first subsection (h) the following new subsection:

“(i) NOTICE OF BLACKOUT PERIODS TO PARTICIPANT OR BENEFICIARY UNDER INDIVIDUAL ACCOUNT PLAN.—

...

(3) CIVIL PENALTIES FOR FAILURE TO PROVIDE NOTICE.— Section 502 of such Act (**29 U.S.C. 1132**) is amended—

<p>(A) in subsection (a)(6), by <i>striking</i> “(5), or (6)” and <i>inserting</i> “(5), (6), or (7)”;</p> <p>(B) by redesignating paragraph (7) of subsection (c) as paragraph (8); and</p> <p>(C) by inserting after paragraph (6) of subsection (c) the following new paragraph:</p> <p>“(7) The Secretary may assess a civil penalty against a plan administrator of up to \$100 a day from the date of the plan administrator’s failure or refusal to provide notice to participants and beneficiaries in accordance with section 101(i). For purposes of this paragraph, each violation with respect to any single participant or beneficiary shall be treated as a separate violation.”.</p>
<p>HIPAA- SEC.101. REPORTING AND ENFORCEMENT WITH RESPECT TO CERTAIN ARRANGEMENTS.—</p> <p>(b) ENFORCEMENT WITH RESPECT TO HEALTH INSURANCE ISSUERS.—</p> <p>Section 502(b) of such Act (29 U.S.C. 1132(b)) is amended by <i>adding</i> at the end the following new paragraph:</p> <p>“(3) The Secretary is not authorized to enforce under this part any requirement of part 7 against a health insurance issuer offering health insurance coverage in connection with a group health plan (as defined in section 706(a)(1)). Nothing in this paragraph shall affect the authority of the Secretary to issue regulations to carry out such part.”.</p> <p>...</p> <p>(e)</p> <p>(1) IN GENERAL.—Section 101 of such Act (29 U.S.C. 1021) is amended—</p> <p>(A) by <i>redesignating</i> subsection (g) as subsection (h), and</p> <p>(B) by <i>inserting</i> after subsection (f) the following new subsection:</p> <p>“(g) REPORTING BY CERTAIN ARRANGEMENTS.—The Secretary may, by regulation, require multiple employer welfare arrangements providing benefits consisting of medical care (within the meaning of section 706(a)(2)) which are not group health plans to report, not more frequently than annually, in such form and such manner as the Secretary may require for the purpose of determining the extent to which the requirements of part 7 are being carried out in connection with such benefits.”.</p> <p>(2) ENFORCEMENT.—</p> <p>(A) IN GENERAL.—Section 502 of such Act (29 U.S.C.1132) is amended—</p> <p>(i) in subsection (a)(6), by <i>striking</i> “under subsection (c)(2) or (i) or (l)” and <i>inserting</i> “under paragraph (2), (4), or (5) of subsection (c) or under subsection (i) or (l)”;</p> <p>(ii) in the last 2 sentences of subsection (c), by striking “For purposes of this paragraph” and all that follows through “The Secretary and” and inserting the following:</p> <p>“(5) The Secretary may assess a civil penalty against any person of up to \$1,000 a day from the date of the person’s failure or refusal to file the information required to be filed by such person with the Secretary under regulations prescribed pursuant to section 101(g).</p> <p>“(6) The Secretary and”.</p>

Figure 4-2 Provisions taken from SOX & HIPAA to illustrate the first case of overlap and possible conflicts

Although SOX and HIPAA agree on the rights and obligations of the Secretary, who is defined as an officer of the health and human services, in SOX, the secretary has a wider authority than in HIPAA. This is indicated by the fact that SOX added a new paragraph (7) to the Employee Retirement Income Security Act of 1974 by allowing the Secretary

to assess a civil penalty against a plan administrator who refuses or fails to provide notice to participants and beneficiaries, which is not the case in HIPAA.

The second situation, which is illustrated in the citation graph of Figure 4-3, shows an example of an overlap between SOX and HIPAA with respect to Section 101 of ERISA (29 U.S.C. 1021). Both HIPAA and SOX amend the same section but two different subsections. Although this decreases the possibility of a conflict, it enforces the fact that there is a strong relation between the policies that HIPAA and SOX mandate in these provisions. In this case, both regulations are adding new obligations on the plan administrator regarding the disclosure and reporting of information. SOX Section 306(b) amended Section 101(i) of ERISA (29 U.S.C. 1021) to ensure that plan administrators notify participants at least 30 days in advance of “blackout periods” (a temporary period of at least 3 days during which contracts or policies are suspended). On the other hand HIPAA Section 101(e)(1) amended Section 101 of ERISA (29 U.S.C. 1021) by adding a new paragraph to establish a filing requirement that was required by the Multiple Employer Welfare Arrangements (MEWA). This requirement consists of a new obligation on plan administrators that mandates reporting and filing annually to The Employee Benefits Security Administration (EBSA) information related to the requirements in Part 7 of ERISA which is added by HIPAA and other similar regulations such as Mental Health Parity Act to ensure compliance by MEWA. These requirements are not needed for SOX compliance.

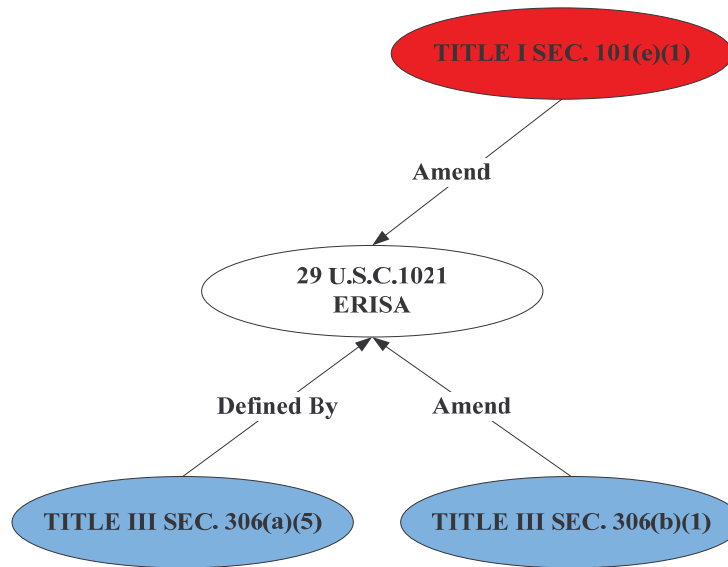


Figure 4-3 An example of SOX and HIPAA provisions amending the same provision of the 29 U.S.C 1132 law but different sections

Case 2: SOX and GLBA

SOX and GLBA overlap by addressing record keeping and the protection of investors' private information. The relationship between SOX provision 205(c)(2) and GLBA 231(a) (shown in Figure 4-4 and the corresponding graph in Figure 4-5) consists of the fact that they both amend the same provision of Section 17 of the Securities Exchange Act of 1934 (code 15 U.S.C 78q). Before the modification made by SOX to this section, it was sufficient from GLBA's perspective for an accountant that certifies the financial documents and reports of an organization to be an independent public account. SOX, on the other hand, amended Section 17 of the Securities Exchange Act of 1934 in such a way that the accountant must be working in a firm which must be registered by the Public Company Accounting Oversight Board (PCAOB) (PCAOB, 2009). PCAOB is a private non-profitable corporation created by SOX to oversee the auditors of public companies to protect the interest of investors (PCAOB, 2009). This amendment made by SOX and

which affected GLBA requires that all tool features, audit processes, and organizational policies that were compliant with the previous version of GLBA be updated.

<p>SOX - TITLE II—AUDITOR INDEPENDENCE</p> <p>SEC. 205. CONFORMING AMENDMENTS</p> <p>(c) OTHER REFERENCES.—The Securities Exchange Act of 1934 (15 U.S.C. 78a et seq.) is amended—</p> <p>(1) in section 12(b)(1) (15 U.S.C. 78l(b)(1)), by striking “independent public accountants” each place that term appears and inserting “a registered public accounting firm”; and</p> <p>(2) in subsections (e) and (i) of section 17 (15 U.S.C. 78q), by striking “an independent public accountant” each place that term appears and inserting “a registered public accounting firm”.</p>
<p>GLBA -Subtitle C—Securities and Exchange Commission Supervision of Investment Bank Holding Companies</p> <p>SEC. 231. SUPERVISION OF INVESTMENT BANK HOLDING COMPANIES</p> <p>BY THE SECURITIES AND EXCHANGE COMMISSION.</p> <p>(a) AMENDMENT.—Section 17 of the Securities Exchange Act of 1934 (15 U.S.C. 78q) is amended—</p> <p>(1) by redesignating subsection (i) as subsection (k); and</p> <p>(2) by inserting after subsection (h) the following new subsections:</p> <p>“(i) INVESTMENT BANK HOLDING COMPANIES.—</p> <p>“(1) Elective supervision of an investment bank holding company not having a bank or savings association affiliate</p> <p>...</p> <p>“(3) Supervision of investment bank holding companies</p> <p>“(A) Record keeping and reporting</p> <p>...</p> <p>“(ii) Form and contents Such records and reports shall be prepared in such form and according to such specifications (including certification by a registered public accounting firm), as the Commission may require and shall be provided promptly at any time upon request by the Commission. Such records and reports may include—</p>

Figure 4-4 Provisions taken from SOX and GLBA to illustrate a case of overlap that can lead to a conflict unless careful monitoring of the changes is performed

The relationship between SOX provision 205(c)2 and GLBA 231(a) shown in Figure 4-5, consists of the fact that they both amend the same provision of Section 17 of the Securities Exchange Act of 1934 (code 15 U.S.C 78q). Before the modification made by SOX to this section, It is sufficient from GLBA’s perspective for an accountant that certifies the financial documents and reports of an organization to be an independent public account. SOX, on the other hand, amended Section 17 of the Securities Exchange Act of 1934 in such a way that the accountant must be working in a firm which must be

registered by the Public Company Accounting Oversight Board (PCAOB). PCAOB is a private non-profitable corporation created by SOX to oversee the auditors of public companies to protect the interest of investors (*PCAOB, 2009*).

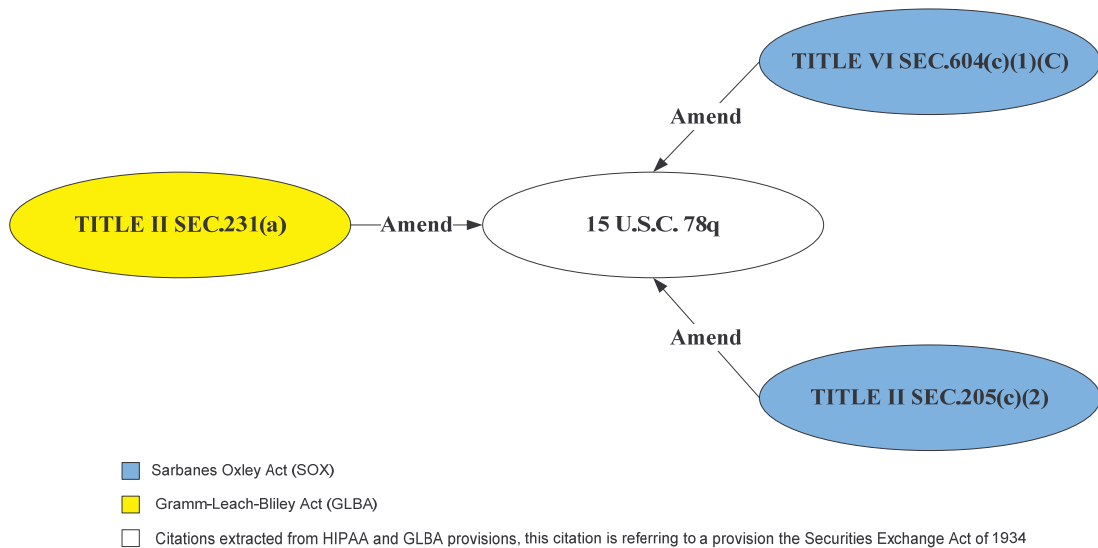


Figure 4-5 Example of SOX and GLBA provisions amending the same provision “15 U.S.C 78q”

Chapter 5: Conclusion and Future Work

“If I had eight hours to chop down a tree, I’d spend six sharpening my axe.”

Abe Lincoln

In this thesis, we argued that there is a need to investigate techniques and tools to help software companies deal with a large number of regulatory compliance requirements. This is because many software systems they develop are for organizations that are heavily regulated and must follow various local and international laws, which dedicate key features that these software systems must support or even the process by which they are to be built.

We discussed how citation analysis can be used to understand and analyze multiple regulations as well as detect overlaps and possible conflicts among regulations. We presented a detailed approach on how to build citation graphs from multiple documents, which involves parsing the content of multiple regulations to identify the provisions, citations, and the types of relations among citations. We discussed the fact that amendments (one of the relations among the citing and the cited provisions) represent risks of conflict. We suggested that this type of relation should be carefully studied

We also introduced a tool called CompDSS that aims to support citation analysis. We used the tool to generate a citation graph from two different sets of regulations. The first

set consists of three regulations that concerned with information privacy, namely, SOX, HIPAA, and GLBA. While the second set represents two copyright acts, namely, DMCA and TEACHA. We were able to detect overlaps and risks of conflicts between these regulations in both cases and as a result verify our approach.

Although the approach presented in this paper and the results of the case study are promising, there are many issues that need to be addressed:

- Citation analysis cannot be applied to analyze the dependencies among laws from different countries since it is unlikely that these laws refer to the same common laws. This limits our approach from being used by corporations that wish to apply it to detect overlaps and conflicts at the international level. We do not have the solution to this problem for the time being.
- Our process for extracting citations relies on citation style standards. However, there might be many citations that do not rely on the citation style standards. One possible solution to this issue is to use text mining techniques to extract such citations. Future research should focus on investigating techniques such as the ones found in text processing so as to detect citations that do not follow standards.
- The size of a citation graph that involves many laws can be relatively large. This can hinder the effective analysis of these laws and the relationships among them. What we need is to investigate ways to improve the usability of these graphs by adding support of various visualization techniques that have been shown to be effective in other areas of software engineering.

- In this thesis, the detection of possible overlaps and conflicts is a user-intensive task. For example, it is up to the analyst to find places where various amendments of the same provisions occur. To alleviate the user from this task, it is possible to investigate patterns that can be categorized as high risk of generating conflicts or overlaps. This will require careful analysis of law documents.
- Finally, there is a need to conduct further experiments involving analysts working with multiple regulations to be able to improve our approach as well as the tool.

Bibliography

- AALL. (2004). *AALL Universal Citation Guide Version 2.1*. U.S.A.: Citation Formats Committee of the American Association of Law Libraries.
- Antoniou, G., Billington, D., & Maher, M. (1999). On the Analysis of Regulations using Defeasible Rules. *in Proc. 32nd Hawaii International Conference on Systems Science* (pp. 225-225). IEEE Press.
- Band, J. (1998). The digital millennium copyright act. *Morrison & Foerster LLP, Washington, DC* <http://www.dfc.org/html/jb-index.html>.
- Barnett, H. (2004). *Constitutional & Administrative Law*. Portland, USA: Routledge Cavendish.
- Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 164-171.
- Bhandari, S., & Computing, T. (2006). Information Trust and Compliance issues. *MS Technology Management - Information Security and Management*, 1-15.
- Breaux, T. D., & Anton, A. I. (2008). Analyzing Regulatory Rules for Privacy and Security Requirements. *IEEE Transactions on Software Engineering*, 34 (1), 5-20.
- Breaux, T. D., Vail, M. W., & Anton, A. I. (2006). Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with

Regulations. *RE'06: Proceedings of the 14th IEEE International Requirements Engineering Conference (RE'06)* (pp. 49-58). Washington, DC, USA: IEEE Society Press.

- Breaux, T. (2009). PhD Theses. *Legal Requirements Acquisition for the Specification of Legally Compliant Information Systems*. North Carolina, USA: North Carolina state university.
- Breaux, T., & Antón, A. I. (2005). Mining rule semantics to understand legislative compliance. *the 2005 ACM workshop on Privacy in the electronic society* (pp. 5-54). New York, NY, USA: ACM.
- Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the workshop on Speech and Natural Language* (pp. 112-116). NJ, USA: Association for Computational Linguistics Morristown.
- Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Comput. Linguist.*, 21 (4), 543-565.
- Chang, H., & Kim, K. (2005). Design of Inside Information Leakage Prevention System in Ubiquitous Computing Environment. *Lecture notes in computer science*, 3483, 128.
- Cougias, D., Halpern, M., & Herold, R. (2007). Say What You Do: Building a Framework of It Controls, Policies, Standards, and Procedures. Schaser-Vartan Books.
- Crews, K. (2004). New copyright law for distance education: The meaning and importance of the TEACH Act. *Copyright Management Center at Indiana*

University—Purdue University-Indianapolis, < [http://www. copyright. iupui.edu/teach_summary. htm](http://www.iupui.edu/teach/_summary.htm)>(retrieved May 28, 2004) .

- Department of Justice - Canada. (2009). *Paragraphing*. Retrieved March 25, 2009, from Department of Justice - Canada official website: <http://www.justice.gc.ca/eng/dept-min/pub/legis/n26.html>
- Dickerson, Darby; Association of Legal Writing Directors. (2003). *ALWD citation manual : a professional system of citation*. New York: Aspen.
- Ehrlich, E. (1922). *The Sociology of Law*. 36 (2), 130-145.
- Ernst & Young. (2006). *Achieving Success in a Globalized World Is Your Way Secure? - Global Information Security Survey*. Ernst & Young.
- Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design Patterns Elements of Reusable Object-Oriented Software*. Boston, MA, USA: Addison-Wesley Longman Publishing Co.
- *Graphviz - Graph Visualization Software*. (n.d.). Retrieved August 11, 2009, from <http://www.graphviz.org/>
- Hamdaqa, M., & Hamou-Lhadj, A. (2009). Citation Analysis: An Approach for Facilitating the Understanding and the Analysis of Regulatory Compliance Documents. *Sixth International Conference on Information Technology: New Generations* (pp. 278-283). Las Vegas, Nevada: IEEE Computer Society.
- Hamou-Lhadj, A. (2009). *Software Compliance Research Group*. Retrieved November 18, 2009, from <http://users.encs.concordia.ca/~abdelw/softwarecompliance.html>

- Hamou-Lhadj, A., & Hamou-Lhadj, A. (2009). A Governance Framework for Building Secure IT Systems. *International Journal of Security and Its Applications (IJSIA)* , 3 (2).
- Harvard Law Review Association. (2000). *The Bluebook: A uniform system of citation*. Cambridge, Mass: Published and distributed by the Harvard Law Review Association.
- Herrmann, D. S. (2007). *Complete Guide to Security and Privacy Metrics: Measuring Regulatory Compliance, Operational Resilience, and ROI*. FL: Auerbach Publications.
- Huang, E. (2007). A DVD Dilemma: Ripping for Teaching. *Convergence* , 13 (2), 129.
- Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall.
- Hutchinson, K. (2003). The teach act: Copyright law and online education. *NYUL Rev.* , 78, 2204-2305.
- ISO. (2009). Retrieved from <http://www.iso.org/iso/home.htm>
- Jacobson, R. L. (2006). *Patent No. 7028259*. U.S.
- Kairab, S. (2004). *A practical guide to security assessments*. FL: Auerbach Publications.
- Kerrigan, S., & Law, K. (2005). A regulation-centric, logic-based compliance assistance framework. *International Journal of Computing in Civil Engineering* , 19 (1), 1-15.

- Kerrigan, S., Lau, G., Zhou, L., Wiederhold, G., & Law, K. (2001). Information infrastructure for regulation management and compliance checking. *The First National Conference on Digital Government* (pp. 167-170). Citeseer.
- Lau, G., Kerrigan, S., Wang, H., Law, K., & Wiederhold, G. (2004). An information infrastructure for government regulation analysis and compliance assistance. *Proc. 5th Conf. on Digital Government Research* (pp. 1-2). Seattle: ACM.
- Lau, G., Law, K., & Wiederhold, G. (2006). A relatedness analysis of government regulations using domain knowledge and structural organization. *International Journal of Information Retrieval* , 9 (6), 657-680.
- Lau, G., Wang, H., & Law, K. (2006). Locating related regulations using a comparative analysis approach. *Proceedings of the 2006 international conference on Digital government research* (pp. 229 - 238). ACM.
- Law Reform Commission. (2008). Statute Law Restatement. Law Reform Commission.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation* (pp. 24-26). New York, NY, USA: ACM.
- Martin, P. W. (2007). *LII / Legal Information Institute*. Retrieved November 18, 2008, from Introduction to Basic Legal Citation: <http://www.law.cornell.edu/citation/>

- McGill Law Journal. (2006). *Canadian guide to uniform legal citation*. Toronto: Thomson Carswell.
- Merriam-Webster Online Dictionary. (2008). *citation*. Retrieved November 25, 2008, from <http://www.merriam-webster.com>
- Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM* , 38 (11), 39-41.
- OAL. (2008). *FAQs: What is the difference between a regulation and a statute?* Retrieved November 25, 2008, from Office of Administrative Law official website: <http://www.oal.ca.gov/faqs.htm#7>
- *OMG*. (2009). Retrieved from <http://www.omg.org/>
- PCAOB. (2009). Retrieved May 20, 2009, from The Public Company Accounting Oversight Board Mission: www.pcaobus.org
- Porter, M. (1980). An algorithm for suffix stripping. *Program* , 30 (14), 130-137.
- Pradeep Varma, D., Rakesh, M., & Sanyal, R. (2007). HMM-based Language-independent POS Tagger. *IICAI*, (pp. 1924-1935).
- Roberts, A. (2009). *ABC of Referencing - ABC of Citation*. Retrieved November 18, 2008, from ABC Study Guide Index: <http://www.studymore.org.uk/refer.htm>
- Schmid, H. (1994). Part-of-speech tagging with neural networks. *Proceedings of the 15th conference on Computational linguistics* (pp. 172-176). Kyoto, Japan: Association for Computational Linguistics.
- Silverman, M. G. (2008). *Compliance Management for Public, Private, Or Nonprofit Organizations, McGraw-Hill Professional*. McGraw-Hill Professional.

- Suber, P. (1999). Amendment. In C. B. Gray, *Philosophy of Law: An Encyclopedia* (pp. I.31-32). Garland Pub. Co.
- Sutton, S. A. (1994). The role of attorney mental models of law in case relevance determinations: an exploratory analysis. *45* (3), 186 - 200.
- *United States Congress Data Dictionary of Legislative Documents*. (n.d.). Retrieved October 1, 2009, from Legislative Documents in XML at the United States House of Representatives: <http://xml.house.gov/>
- Van Guilder, L. (1995). *Automated Part of Speech Tagging: A Brief Overview*. Georgetown University.
- Yourdon, E., & Constantine, L. L. (1979). *Structured design: fundamentals of a discipline of computer program and systems design*. NJ, USA: Prentice-Hall, Inc.
- Zhang, P., & Koppaka, L. (2007). Semantics-based legal citation network. *Proceedings of the 11th international conference on Artificial intelligence and law* (pp. 123-130). New York, NY, USA: ACM.