

# Leveraging Structure to Query Text and Data in Regulatory Filings

Thomas Y. Lee  
University of Pennsylvania  
thomasyl@wharton.upenn.edu

## INTRODUCTION

The Government Paperwork Elimination Act, enacted in October 1998, has driven a steady migration from paper to electronic filings in response to regulatory requirements. By many measures, the effort has proven a success. For example, through August 2006, the U.S. Securities and Exchange Commission (SEC) has received more than 4.5 billion electronic filings and the U.S. Environmental Protection Agency (EPA) has received nearly 73 million filings.

Auditing these submissions involves integrating text and data from multiple filings. In addition to numerical figures, a comprehensive review of regulatory submissions would include the analysis of mandatory and optional prose that report information such as "forward-looking" statements (SEC 10-K Item 7.)[1] or "hazard evaluation prevention" (EPA ARIP 21b. 22b.)[2]; and because a single document may include references to other filings and because a single company may file multiple documents over a single fiscal year, the audit process would integrate both text and data from multiple documents by a single filer. Moreover, requirements for industry-specific comments such as "risk-factors" (SEC 10-K Item 1A.) or "mitigation response" (EPA ARIP 18.) require the comparison of filings between multiple companies within a single industry to assess sector-specific norms.

## MOTIVATION

Due in part to complexity and sheer volume, comprehensive reviews of electronic filings are not common. In the case of the SEC, even with Sarbanes-Oxley and electronic-filing, mandated three-year audits for every filer may be limited to narrow spot-checks [3]. Electronic filing is no panacea. Current SEC submission guidelines are limited to text or optional HTML formatting [4]. Although XML element definitions are being studied by all government agencies, such proposals do not address the selective, textual elements of regulatory filings (see, for example, XBRL and proposed data elements for Institutional Controls [5, 6].

The objective of this work is to automatically extract and integrate content from the text segments of regulatory filings for the purpose of compliance checking. We propose to leverage knowledge about document structure to learn schemas and constraints. Rather than relying upon an explicit schema or a training set of representative document instances, we begin with the actual legislation and corresponding regulatory text. From the regulations, we manually identify a set of mandatory and optional labels (headings) and preliminary hierarchical structure (sections, sub-sections, etc.). Extending the structured retrieval (information retrieval) and wrapper induction (information extraction) literature, we automatically learn regular expressions corresponding to mandatory and optional labels, the ordering of optional labels, and both firm and industry-specific inclusion constraints between elements.

## RELATED WORK

In structured text retrieval, document word models are extended with structural context [7]. Early work modeled structure as a single sequence of contiguous, non-overlapping fragments. Other work supports multiple structural hierarchies. In either case, query results are refined by structural context (e.g. where in the document a term appears). Unlike information retrieval, we wish to query specific document fragments within each filing rather than ranking all submissions. More significantly, rather than assuming knowledge of the document hierarchy and its associated text patterns, we are given structural labels from the regulations and learn the corresponding patterns.

Chang et al. recently summarized techniques to automatically learn extraction patterns from text [8]. Supervised information extraction approaches rely upon users to identify interesting elements within representative document instances where unsupervised techniques automatically identify elements based

upon repetition within and between instances. The goal is extraction patterns for one or more attribute-value records. Unfortunately, although regulators may specify a standard set of labels, different firms and industries deviate from the instructions in idiosyncratic ways (see Fig 1). Moreover, the labels themselves change over time as regulations evolve. We learn candidate patterns from the regulatory text and relax those patterns for individual document instances. Patterns for individual labels are combined to match the entire, hierarchical document model as in structured retrieval rather than the slots of record(s) embedded within the text.

Given the document model, there are techniques for learning structural relationships that approximate semistructured constraints [9, 10]. We seek to learn constraints that define relations between structured data and unstructured text within the same or linked filings. For example, only firms within specific SIC codes might report "store openings" as a "forward-looking" financial event. Likewise, certain industries may never require remediation measures for airborne pollutants from accidental chemical releases.

SEC General Instructions for filing Form 10-K, last updated 12/05	
<b>Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operation.</b>	
Furnish the information required by Item 303 of Regulation S-K (§ 229.303 of this chapter)	
<b>Item 7A. Quantitative and Qualitative Disclosures About Market Risk.</b>	
Furnish the information required by Item 305 of Regulation S-K (§ 229.305 of this chapter)	
<b>Item 8. Financial Statements and Supplementary Data.</b>	
Furnish financial statements meeting the requirements of Regulation S-X (§ 210 of this chapter) ...	
Kohl's 10-K, 1997	
Item 7.	MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS
...	
Item 8.	FINANCIAL STATEMENTS AND SUPPLEMENTARY DATA
...	
Federated 10-K, 1998	
ITEM 7. MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS	
...	
ITEM 7A. QUANTITATIVE AND QUALITATIVE DISCLOSURES ABOUT MARKET RISK	
...	
ITEM 8. CONSOLIDATED FINANCIAL STATEMENTS AND SUPPLEMENTARY DATA.	
...	

**Figure 1. Comparing labels in the regulatory instructions [1] to 10-K filings from different companies in different years.**

## REFERENCES

- [1] U.S. Securities and Exchange Commission, "Annual Report Pursuant to Section 13 or 15(d) (Form 10-K)," December 2005.
- [2] U.S. Environmental Protection Agency, "Accidental Release Information Program (ARIP)," 1999.
- [3] M. Leone, "RX for Fraud: More SEC Checkups," in *CFO.com*, 2003.
- [4] U.S. Securities and Exchange Commission, "EDGAR Filer Manual (Volume II)," 3 ed, 2006.
- [5] U.S. Environmental Protection Agency, "Draft IC Data Element Registry," Chicago IC Data Exchange Training Workshop, 2003.
- [6] U.S. Securities and Exchange Commission, "FAQ: XBRL Voluntary Filing Program," 2006.
- [7] R. Baeza-Yates and G. Navarro, "Integrating Contents and Structure in Text Retrieval," *SIGMOD Record*, vol. 25, pp. 67-79, 1996.
- [8] C.-H. Chang, M. Kaye, M. R. Girgis, and K. F. Shalhan, "A Survey of Web Information Extraction Systems," *IEEE TKDE*, pp. 1411-28, 2006.
- [9] P. Buneman, S. Davidson, W. Fan, C. Hara, and W. C. Tan, "Keys for XML," WWW, 2001.
- [10] K. Wang and H. Liu, "Discovering Structural Association of Semistructured Data," *IEEE TKDE*, pp. 353-71, 2000.