

Paper summary: 2007 Winter Conference on Business Intelligence

Title: Learning from Legislation to Query Regulatory Filings

Author: Thomas Y. Lee

Institution: University of Pennsylvania, The Wharton School

Email: thomasyl@wharton.upenn.edu

Word count: (921 excluding references; 1052 total)

Learning from Legislation to Query Regulatory Filings

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 been a success. For example, through August 2006, the U.S. Securities and Exchange Commission (SEC) has received more than 4.5 billion electronic filings. These filings are read by institutional and individual investors for risk management, by industry partners and competitors for strategic planning, and by Federal auditors for enforcing compliance with public policy.

In addition to numerical figures, a review of regulatory submissions would include the analysis of mandatory and optional prose that report information such as "forward-looking" statements and factors affecting market risk (SEC 10-K Item 7. and 7A.)[1]; moreover, requirements for industry-specific comments such as "risk-factors" (SEC 10-K Item 1A.) require the comparison of filings between multiple companies within a single industry to assess sector-specific norms.

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 [2]. Electronic filing is no panacea. Current SEC submission guidelines are limited to text or optional HTML formatting [3]. 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 [4]).

RESEARCH QUESTION AND APPROACH

In this paper, we automatically learn to extract and integrate content from the text segments of regulatory filings for the purpose of competitive analysis and regulatory audit. We leverage knowledge about document structure derived from the regulatory policy documents to learn extraction patterns that are robust to the idiosyncrasies of individual filings.

Unlike the prior literature that relies upon an explicit schema or a training set of representative document instances, we begin with the actual legislation and corresponding regulations. From the policy documents, we identify, by hand, the legislatively mandated reporting structure and corresponding elements (sections, sub-sections, etc.). Extending the structured retrieval (information retrieval) and wrapper induction (information extraction) literature, we construct a two-level finite-state transducer that first generates regular expressions corresponding to mandatory and optional elements and second defines integer constraints to order those elements.

RELATED WORK/EXPECTED CONTRIBUTIONS

In structured text retrieval, document word models are extended with structural context [5]. Structure is modeled as a single sequence of contiguous, non-overlapping fragments or by multiple such hierarchies. In either case, the structure is assumed and query results are refined relative to that context (e.g. where in the document a term appears). Structured retrieval techniques assume that documents are already fragmented or that fragment separators are known. We define document fragments as semistructured elements; the element regular expression patterns are learned from the regulatory instructions. In addition, we query elements within a filing rather than rank the relevancy of that filing.

Chang et al. recently summarized techniques to automatically learn extraction patterns from text [6]. 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 elements, different firms and industries deviate from the instructions in idiosyncratic ways (see Fig 1). Moreover, the elements 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. Finally, because a particular pattern may match multiple times within a document (e.g. internal cross-referencing), structural context in the form of *where* a pattern matches in a document relative to other elements is critical.

Given the document model, there are techniques for learning structural relationships that approximate semistructured constraints [7, 8]. 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.

CURRENT STATUS OF MANUSCRIPT

A preliminary technical implementation is complete as is the collection of an initial data set of all 10-K filings from 1995-2005. Preliminary evaluation applied to three unrelated SIC codes is on-going as is an extension to combine Part-of-Speech tagging and Noun-Phrase Verb-Phrase text chunking to discover additional, optional elements.

SEC General Instructions for filing Form 10-K, last updated 12/05	
Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operation.	
Furnish the information required by Item 303 of Regulation S-K (§ 229.303 of this chapter)	
Item 7A. Quantitative and Qualitative Disclosures About Market Risk.	
Furnish the information required by Item 305 of Regulation S-K (§ 229.305 of this chapter)	
Item 8. Financial Statements and Supplementary Data.	
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 elements in the regulatory instructions [1] to 10-K filings from different companies in different years.

REFERENCES

- [1] U.S. SEC, "Annual Report Pursuant to Section 13 or 15(d) (Form 10-K)," December 2005.
- [2] M. Leone, "RX for Fraud: More SEC Checkups," in *CFO.com*, 2003.
- [3] U.S. SEC, "EDGAR Filer Manual (Volume II)," 3 ed, 2006.
- [4] U.S. SEC, "FAQ: XBRL Voluntary Filing Program," 2006.
- [5] R. Baeza-Yates and G. Navarro, "Integrating Contents and Structure in Text Retrieval," *SIGMOD Record*, vol. 25, pp. 67-79, 1996.
- [6] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shalan, "A Survey of Web Information Extraction Systems," *IEEE TKDE*, vol. 18, pp. 1411-28, 2006.
- [7] P. Buneman, S. Davidson, W. Fan, C. Hara, and W. C. Tan, "Keys for XML," WWW, 2001.
- [8] K. Wang and H. Liu, "Discovering Structural Association of Semistructured Data," *IEEE TKDE*, vol. 12, pp. 353-71, 2000.