Sarbanes-Oxley, Basel II, and data mining opportunities in compliance systems

G. Allen
Research and Development, TOWER Software

Abstract

New legislative and industry governance directives have sparked development of systems to tackle Information Lifecycle Management, and related compliance automation for regulations including Sarbanes-Oxley, Basel II and ISO 15489 records management. Many of these systems use a "store now, sort later" philosophy, whereas others capture related information in the context of business processes.

Unusual avenues emerge from these systems for data mining and knowledge discovery in the hybrid database of structured and unstructured information that form the supporting compliance systems. This paper will explore the data mining possibilities inherent in compliance systems using classification systems not designed with such mining opportunities in mind. Examples include identifying customers likely to be early adopters of new software based on product and customer support systems; and opportunities for knowledge discovery and mapping, and identifying subject matter experts within taxonomies designed to support financial business functions.
Keywords: Sarbanes-Oxley, Basel II, enterprise content management, records management, data mining.

1 Introduction

New legislative and industry governance directives have sparked development of systems to tackle Information Lifecycle Management, and related compliance automation for regulations including Sarbanes-Oxley, the new Basel Capital Accord (Basel II) and the ISO 15489 Records Management standard.

Two broad approaches to compliance systems are pursued in contemporary markets. The first attempts to retrofit or re-engineer bespoke compliance



features into existing systems. The second approach uses a common compliance infrastructure and repository as a foundation for all line of business systems [1].

This paper presents some atypical data mining opportunities possible when using a common compliance infrastructure for all line of business systems. A case study describes the outcomes from exploratory mining of a Sarbanes-Oxley compliance support system. Opportunities for data mining and knowledge mapping are then discussed within the framework of a Basel II compliance system.

2 Research methodology

A typical approach in such scenarios would be to mine using decision trees, clusters, neural networks and other traditional approaches, seeking to use the various aspects of financial and other information captured by the compliance system to ascertain high-value customers, repeat business opportunities, purchasing trends, patterns, and the like [2].

The research undertaken for this paper instead deviated from this safe trajectory, choosing to perform exploratory mining, using quite simple, well-proven data mining approaches and algorithms with unusual business requirements as driver, and outcomes measured in different ways. The primary mining algorithms used were a common Decision Trees algorithm, and a Clustering algorithm using expectation-maximisation [3]

3 Case study - data mining and Sarbanes-Oxley compliance systems

The Sarbanes-Oxley Act mandates a wealth of checks and balances to ensure integrity of financial information from source through to board level. Many aspects of a business' day-to-day operations are affected by these requirements. Section 404 of the Act describes in somewhat ambiguous terms the audit and data collection requirements of all supporting information technology systems [4]. This data set of configuration changes, patches, installation, and fixes to production systems is usually overlooked in the rush to mine the financial data within the systems themselves.

As part of its Sarbanes-Oxley commitment to clients, a multination software vendor utilised a common repository system for all line of business applications, including finance, sales and order processing, customer relationship management, and application development management. To satisfy the simultaneous needs of tracking support incidents, and the subset of these incidents covering requests for new software as part of ongoing maintenance contracts, the ECM product was configured such that an overly-generous set of information about the current IT operating environment, and the proposed environment targeted for new software releases, was recorded. It was this rich source of data that presented several eccentric data mining avenues.

3.1 Business goal for mining

The business goal for data mining was articulated as identifying what co-factors in an IT environment influenced upgrades to the software package, and by correlation, what was lacking in those clients' environments most tardy at upgrading. The cost imperative was still a motivating factor - those clients delaying upgrades induce a "late kick" in the cost of support, as systems age, expertise is lost, and risk escalates [5].

3.2 Model development

A data set was harvested from the compliance system, leading to over 150 distinct data elements being tracked for a given software client. With this scale, it was considered overly ambitious to attempt to build a single star or snowflake data model to represent all data in the first instance. Research instead focused on slices of client “metadata” with 10 to 15 dimensions at a time, exploring those collections that showed early promise.

Several decision tree models were attempted using data profiles drawn from the following data elements

- Licencing financial information, such as Licence Revenue, Licence Fee, Local Tax Amount, Local Tax Jurisdiction, Billing Region, Maintenance Date
- Licencing metrics, such as Licence Seat Type, Corporate Licence Model, Licence Geography
- Support information, such as Maintenance Date, First Line of Support, Second Line of Support, Number of Support incidents
- Demographic information, such as Institution Type, Number of Users
- Product version information, including Current Installed Product Version, Latest Version Issued, Version Upgrade History
- Server IT environment, including Server Operating System, Server Database System, Networking protocol, Server hardware type
- Client IT environment, including PC Operating System, virus software in use, desktop word processing, spreadsheet and mail software.

The resultant models showed no clear pattern or relationship that would indicate influence over the "Current Installed Product Version".

Nine decision tree models were explored, using a selection of different metadata slices, with different biases, from varying categories. In most cases, overfitting became a problem, with the sheer number of discrete software versions causing excessive forking at the first or second branch level. When corrected to apply better stunting of leaf cases, the resulting models still showed no well-defined or apparently useful patterns. This outcome was deemed largely to be a shortcoming of the method used to employ the mining algorithms, rather than a flaw in the algorithms themselves. The abundance of data coupled with time constraints meant that no systematic exploration of tree models could be attempted, nor better data enhancement and cleaning that may have helped deliver more useful decision tree models.



Clustering models were explored next, and early successes with these models influenced the decision to explore these to a state that could support a hypothesis, while avoiding data overfit and other problems that affect the decision tree models.

Thirty variations on clustering model were explored. The most promising model used a small data profile selection, as follows.

Table 1: Cluster Model, data profile, (model 3 of 30).

Organisation Identifier	Support Status
Server Operating System	First Line of Support
Network Operating System	Last Version Issued
Database System	Last Version Requested
Database Size	Current Version Installed

This collection focused on the technology at the server tier, support characteristics, and software version profile.

The cluster can be visualised in Figure 1.

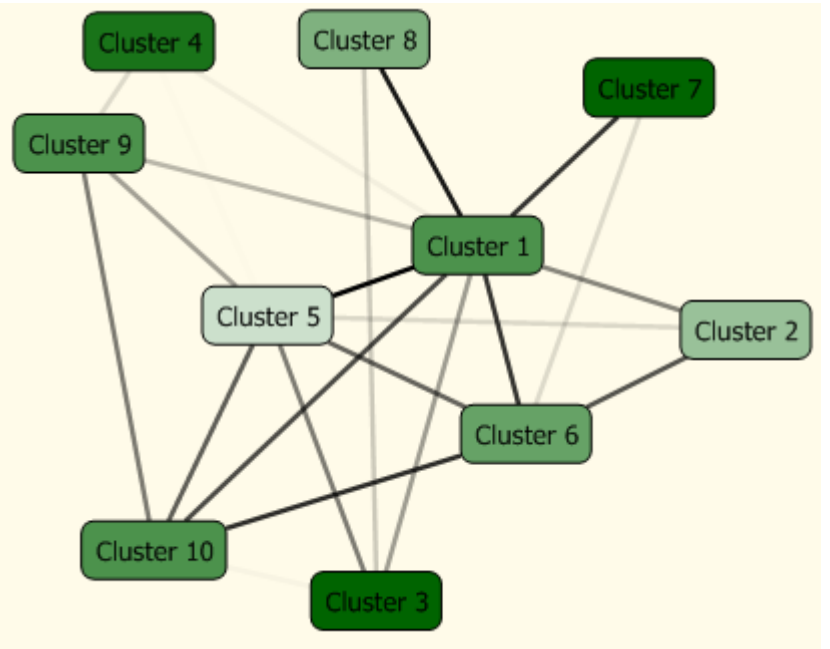


Figure 1: Clustering visualisation, model 3 of 30.

Two clusters, Cluster 5 and Cluster 1, showed very strong patterns of influence on the current version of software installed. Cluster 5's properties indicated a common use of a now-deprecated database technology, coupled with an old version of the software deployed, but a recent request for new software to

be issued as per maintenance contracts. Cluster 1's profile showed a grouping that covered a particular business partner offering First Line of Support, and Current Version Installed only slightly trailing the Latest Version Sent.

3.3 Discussion

From the perspective of the company, and its business goal, the clusters uncovered useful behavioural patterns amongst clients.

Cluster 5 uncovered a suspected, but previously only anecdotal, concern amongst a client subset seeking guidance over the deprecation of support for one particular database type, and consequently delays over upgrading their very old version of the software

Cluster 1 revealed the influence of a service provider within the Asia Pacific region providing First Line of Support, coupled with a relatively recent software version, and issue of the most recent software version. The model highlighted the efficacy of this service provider's policy of keeping their clients up to date.

The use of data mining techniques on the compliance system was seen as a very useful exercise. The mining techniques used were very simple, and while not providing instant monetary return, new sales, or startling insight, it has allowed the company to embark on a new support model that promotes services and support to keep clients current with their software installation. A set of new exploratory goals for mining is being developed for a future project.

In relation to the Sarbanes-Oxley Act, organisations and their compliance systems will track a similar, if smaller, set of audit information covering maintenance practices for software and supporting information technology systems, including upgrade process and outcomes. This case study illustrates data mining possibilities in such a system.

4 Basel II and ISO 15489 records management: opportunities for innovative data mining

To most observers, Basel II is fundamentally about capital adequacy rules, risk identification and management, and transparent processes in these areas as a mechanism for mitigating risk: e.g. fraud. Buried in its implementation details are data capture requirements regarding operational risk. In the USA, these requirements started in 2004, and are to provide a contiguous body of data by 2007 [6]. Other jurisdictions will stager the data collection requirements to synchronise with adoption timeframes. This will provide a rich vein of information for data mining, and this study illustrates data mining and knowledge discovery advantages inherent in using a sound records management system as the basis for the compliance application.

Internationally recognised electronic document and records management systems (EDRMS), such as those compliant with regimes like DoD 5015.2 and UK PRO standards (based on ISO 15489), are by design centred on long-term data capture of every business transaction performed by an organisation [7, 8]. While this can include financial and monetary details down to the line-item,



supporting structures will typically allow for classification of such business tasks within a pre-defined taxonomy, with concomitant registration of participants, workflow and activities, as well as unstructured documentation and communication accompanying the transaction.

4.1 Classification taxonomies as groundwork for data mining

For instance, a prototype classification system designed for financial transactions can be illustrated as per figure 2.

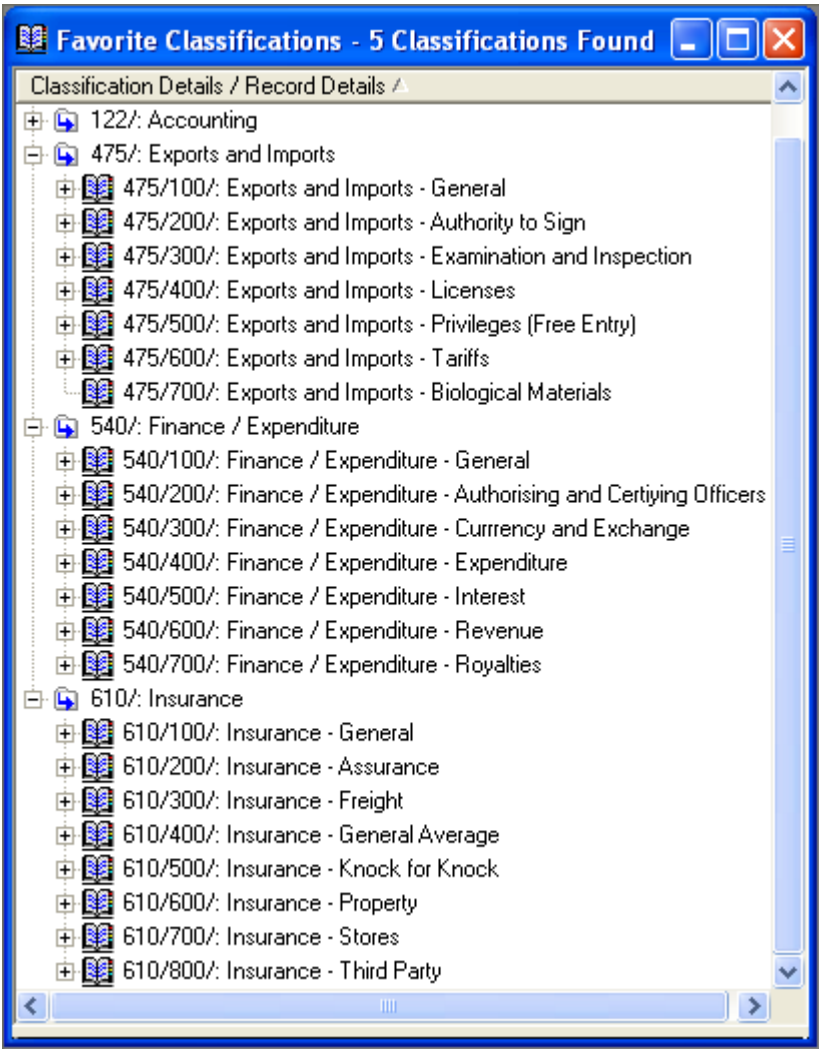


Figure 2: Prototype classification structure for financial transactions.

To the business user, this represents a functional description of task types (even in this overly simplified example), and a convenient long-term classification approach to data capture and retention. From a data mining perspective, a data repository is already formed, with the advantage that data has already undergone point-of-capture processing to facilitate data identification, cleansing, and some aspects of enrichment – such as calculation of additional data items including future destruction or archiving dates that would otherwise be performed as a discrete enrichment step in some data mining approaches. Clearly this does not obviate the need for a strong methodology to ensure cleanliness and appropriateness of data, but it can accelerate the process

Compliance-focused EDRMS also implement comprehensive data profiles of internal and external participants in business activities, together with complete audit trails of all transactions by these participants, as mandated by contemporary standards [7, 8].

Knowledge mapping opportunities are immediately apparent in this scenario. One key driver for contemporary knowledge mapping is to alleviate the problems caused to organisations by the transient nature of knowledge particular to combinations of business participants and business functions [9].

By combining the thorough data trail of all business participants, in conjunction with the strong taxonomies supported by the EDRMS, traditional knowledge mapping outcomes are readily supported. For example, islands of skill and expertise can be identified based on specific functional areas of the classification system, and previously unrecognised tacit and informal knowledge can make itself known

5 Conclusion

Compliance systems are evolving to meet the needs dictated by contemporary business environments, under the influence of legislation and regulations such as the Sarbanes-Oxley Act, and Basel II. Those compliance systems with strong enterprise content and records management features offer a wide range of innovative data mining opportunities often overlooked in the rush to glean immediate financial insight and advantage.

The examination of a compliance system used to support Sarbanes-Oxley outcomes by an international software firm illustrates some of the atypical possibilities in a business environment. With the aid of data mining and knowledge mapping, these compliance systems can be treated as valuable data assets, rather than regulatory cost centres.

References

[1] Managing corporate records for Sarbanes-Oxley, TechTarget.com, Storage Management Web Site. http://searchstorage.techtarget.com/tip/0,289483,sid5_gci932864,00.html

[2] Hirji, K. K. *Mining for Data*. Today's Engineer, 2(2), pp. 32-35



[3] Effective Strategies for Data Mining. Microsoft Web Site, Redmond, Washington http://www.microsoft.com/resources/documentation/sql/2000/all/reskit/en-us/part6/c2461.mspx
[4] Internal Computer Investigations as a Critical Control Activity Under Sarbanes-Oxley. Web Site, Thomson-Findlaw, http://library.lp.findlaw.com/articles/file/01015/009276/title/Subject/topic/Employment%20Law_Whistleblowing/filename/employmentlaw_1_489
[5] Central Sponsor for Information Assurance, Cabinet Office Web Site, London. http://www.knowledgenetwork.gov.uk/co/kimscsia.nsf/0/E78166CE814F839480256EA100386554?OpenDocument
[6] Preparing for the Pain of Basel II, Developer.com Web Site, http://www.developer.com/java/ent/article.php/3403901
[7] DoD 5015.2 Standard, Joint Interoperability Test Command, Web Site, Arizona. http://jitc.fhu.disa.mil/recmgt/index.html
[8] Electronic Records Management Functional Requirements for ERMS, Web Site, United Kingdom National Archive, London http://www.nationalarchives.gov.uk/electronicrecords/reqs2002/
[9] Knowledge Mapping: A Practical Overview. Smith-Weaver-Smith Web Site, Indianapolis. http://www.smithweaversmith.com/knowledg2.htm

