Mining Executive Compensation Data from SEC Filings

Chengmin Ding
Institutional Shareholder Services
2099 Gaither Rd, Suite 500
Rockville, MD 20850
chengmin.ding@issproxy.com

Ping Chen
Dept of Computer and Mathematical Sciences
University of Houston-Downtown
One Main St.
Houston, TX 77002
chenp@uhd.edu

ABSTRACT

In recent years, corporate governance has become a more and more important concern in investment decision-making. As one of the most important factors in evaluating corporate governance, executive compensation study has drawn a lot of attention. Most companies with excessive executive pay are linked with scandals or corporate failures. At the same time, there are a lot of cases that executive compensation has no direct relationship with the company's stock performance, which harms the company's growth and sacrifices shareholders' interests. So collecting and evaluating executive compensation data becomes a pressing issue.

This paper presents a text mining system ECRS (Executive Compensation Retrieval System) to automatically extract executive compensation data from the SEC proxy filing. An analysis based on the extracted data is provided and some samples on using the raw data to derive useful information for the financial analysts are also presented

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data mining*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text Analysis*

General Terms

Text Mining, Corporate Governance

Keywords

Executive Compensation, Ontology, SenseNet, ECRS

1. INTRODUCTION

On December 2nd 2001, Enron, the U.S. seventh biggest company in revenue, filed bankruptcy, the biggest corporate bankruptcy in U.S. history at that time. The collapse of Enron soon started a wave of skepticism about U.S. top executives and the businesses under their management. A series of accounting scandals were exposed to the public, together with the stories of how self-interested senior executives enriched themselves by hiding and deceiving shareholders and then walked away with huge option profits while shareholders were left to suffer the consequences. With each revelation of corporate fraud, faith in Corporate America built up during the past decades has eroded a little more, giving way to widespread suspicion and mistrust.

To rebuild the confidence of shareholders, Sarbanes-Oxley Act was passed by the US congress on July 30, 2002. It demands the truthfulness and thoroughness of a company's financial reporting and it enforces a set of good corporate governance guidelines. So the challenge facing the investor world has gradually shifted from little disclosure, cooked books etc to truthful but overwhelmed amount of information. All the information are buried in tens of thousands of public text filings in the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system. Now how to sort out the useful information from all the public filings has become an ever-pressing issue.

In the commercial world, several EDGAR data vendors exist, e.g. EDGAR online, 10-K Wizard etc. They are used on portal sites for general financial information (e. g. at Yahoo, BigCharts). These EDGAR agents, however, usually can only extract well-structured information. They are either unable to provide the desired information or charge expensive fees for customized solutions which is not affordable for small firms and individual investors. In the academic field, some researches have been conducted and published in extracting company annual reports(10-K) [3] , but a google search has revealed the researches and applications in this field are still very limited.

The authors have been working on extracting the corporate governance related information from EDGAR over the past several years. We will present a new method based on SenseNet technology to automatically extract executive compensation data from the SEC proxy filing. In the following sections, we will first introduce Corporate Governance as a business context; then we will take a look at the structure of SEC proxy filing which is the text document we are going to mine on; next we will explain our ECRS system in technical details; finally we will present out experiment results and sample usages of the raw data. It worth pointing out that our ECRS system is an ongoing project and the results shown here are preliminary.

2. COPORATE GOVERNANCE

Only fifteen years ago, there was little academic interest let alone practical usage in corporate governance. The modern trend of developing corporate governance guidelines and codes of best practice began in the early 1990's in the United Kingdom, the United States and Canada in response to problems in the corporate performance of leading companies, the perceived lack of effective board oversight that contributed to those performance problems, and pressure for change from institutional investors. Now there is an almost-exponential increase in studies in a wide field of corporate governance influence. Scholars from the fields of business, accounting, economics, finance, law and political science are pouring out studies. Global and regional forums and seminars have been held to encourage dialogues between academic studies and industry implementations and to promote best practice all over the world.

Then what is corporate governance? It is defined as the relationship among various participants in determining the direction and performance of corporations. The primary participants are (1) the shareholders, (2) the management (led by the chief executive office), and (3) the board of directors. [1]

The fundamental concern of corporate governance is to ensure the means by which a firm's managers are held accountable to capital providers for the use of assets. When a firm's management is separate and distinct from the providers of the firm's capital, managers have a responsibility to use assets efficiently in pursuit of the firm's objective.

Nowadays, more and more institutional investors are using Corporate Governance as one of the most important screening criteria. A lot of analyses have been conducted on the corporate governance related filings by the investors. Most of the relevant information for corporate governance is in the proxy filing and we will discuss it in the next section.

3. OVERVIEW OF PROXY DOCUMENT

Every year, stock shareholders will receive a proxy statement in the mail. It can also be downloaded from SEC's website. The official name of the document is DEF-14A.

The purpose of the proxy statement is to provide investors with enough disclosure to make important shareholder votes. At least once a year, at the annual meeting, shareholders will be able to vote for important agenda such as electing directors and issuing new stocks.

The structure of a proxy statement includes [2]:

- Short letter from the CEO, explaining when the annual meeting will be held and the issues to be voted on
- Description of voting rights and the process of voting
- Detialed description of the matters to be voted on
- List of the major shareholders
- Executive compensation
- Audit committee report
- Compensation committee report
- Shareholder proposals

Among all this information, we believe the executive compensation information is one of the most important piece to examine.

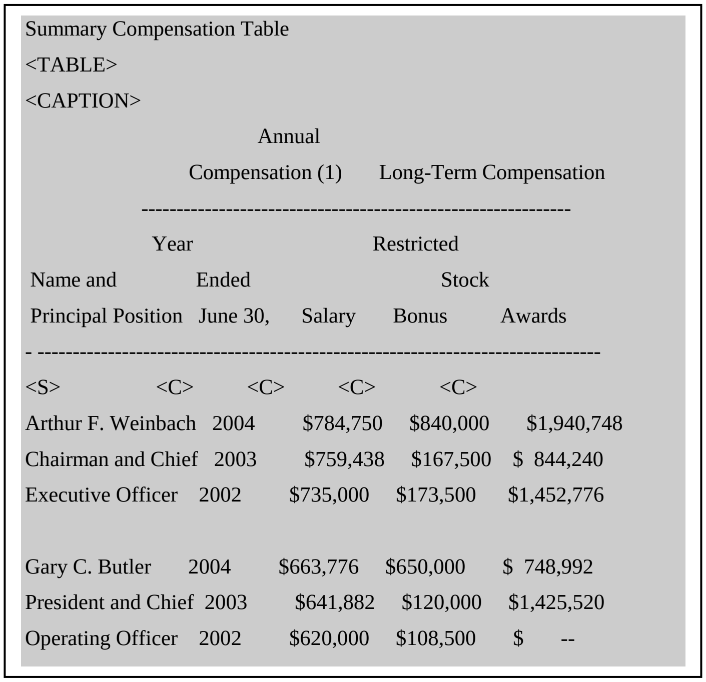
First of all, executive compensation study is a core part of corporate governance [1][2]. An excessive pay package usually signals a potential corporate governance problem. For example, if the grant size of a stock option is 10 million shares, it is probably verging on the obscene. After all, if the stock goes up a mere $1, the executive gets a nice $10 million boost to net worth. This might be a potential cause of the executives to cook the book and raise the stock price temporarily – at the cost of thousands of shareholders.

Secondly, Pay for performance is now a major topic in the corporate world. Only if an executive's compensation is linked to the company's stock performance, the executive's interest is aligned with that of the shareholders. For example, the former CEO of WorldCom, Bernie Ebbers got a $65,000 hike in his salary and a $10 million bonus even though the stock fell 70 percent [2]. He wouldn't think about the shareholder loss since he still got good money. The result, WorldCom filed the largest bankruptcy in the US's history.

Furthermore, compensation issues present shareholders with some of their most cost-effective (highly leveraged) opportunities for "investing" in shareholder initiatives. A shareholder can submit a shareholder proposal about executive compensation for little more than the cost of a stamp. [1]

The role of the shareholders with regard to compensation starts with one simple point: compensation presents an investment opportunity. The executive compensation plan is a clear indicator of the company's value as an investment. Both corporate governance practitioners and financial analysts are very interested to find out such information.

Based on the above discussion, we can see the disclosure of executive compensation has recently drawn more and more attentions. To have a way to quickly get this information for the target company and its peers (in the same index or industry) are vital.



```
Summary Compensation Table
<TABLE>
<CAPTION>
                              Annual
                        Compensation (1)      Long-Term Compensation
                   ------------------------------------------------------------
                  Year                         Restricted
 Name and             Ended                              Stock
 Principal Position   June 30,     Salary      Bonus         Awards
- -------------------------------------------------------------------------------
<S>              <C>         <C>         <C>           <C>
Arthur F. Weinbach   2004        $784,750     $840,000      $1,940,748
Chairman and Chief   2003        $759,438     $167,500    $  844,240
Executive Officer    2002       $735,000     $173,500     $1,452,776

Gary C. Butler       2004       $663,776     $650,000      $  748,992
President and Chief  2003        $641,882     $120,000     $1,425,520
Operating Officer    2002       $620,000     $108,500      $      --
```

Figure 1. Summary Compensation Table from DEF 14A.

4. MINING THE PROXY DOCUMENT

We built an Executive Compensation Retrieval System (ECRS) to extract the executive compensation data. ECRS includes a web crawler scheduled to go to www.sec.gov to download the latest DEF 14A filings in text format. Then the ECRS engine will loop through the local repository and parse the documents. In DEF 14A, the Summary Compensation Table contains all the executive compensation information. ECRS will look for this table (c.f. Figure 1) and extract the executive compensation information from it.

As you can see in Figure. 1, there exist some variances within this simple table, and more variances are found with tables from our collection of DEF 14A forms submitted by public companies to SEC. Although these forms are usually carefully prepared, typos, errors and other noise still exist and usually will hurt quality of information extraction considerably. There exist a lot of work in table mining. Most of these techniques are developed for table analysis only, and our technique utilizes tables as a case study and can be generalized for regular text analysis. We use a technology called SenseNet which is a variation of hidden markov model. It includes the following techniques.

- Pattern matching with regular expressions
- An ontology knowledge base for word classification
- Data range checking
- Shallow word context analysis
- Predictive parsing with weights and probabilities

The next few sections discuss each of these subjects in more detail.

4.1 Matching with Regular Expressions

Regular expressions are used throughout the application for both finding text within the document and for performing document preprocessing. Preprocessing provides an effective way to condition the document prior to processing including the ability to strip unwanted text or formatting out of the document. For example, columns will sometimes use -- to represent a 0. With preprocessing, we can convert -- to 0. Preprocessing can also be useful for striping HTML tags from the document.

4.2 Ontology for Word Classification

As was shown in the summary compensation table excerpt (c.f. Figure 1), the first column of data contains both personal names and position titles combined together. It is necessary to split this information apart, and word ontology provides an effective way to do just that.

To create the personal name ontology, a list of personal names from the U.S. Census Bureau was used. According the Census Bureau, this list contains approximately 90% of all of the first and last names in use in the U.S. The list was partitioned by first and last name and the total number of entrees is 91,933. For the position and title ontology, the title words were manually extracted from about 25 financial randomly picked documents. Example title words include CEO, CFO, Chairman, Chief, and CIO etc. Unlike the personal names, this list is very small with only 40 members. The resulting ontology looks similar to the following:

```
                 ...
Gary is_a personal name
                 ...
President is_a title
                 ...
Butler is_a personal me
```

Figure 2. Ontology Example

4.3 Data Range Checking

Numerical range checking provides a simple means for data classification. For example it can be safe to assume that the year column for this project will contain numbers within the range of 2000 +/- 10 years. We can also check salary column in a similar way. For example, we expect that the salary will be a number greater than 0. This range checking will alleviate problem caused by misaligned table cells and ensure data quality.

4.4 Shallow Word Context Analysis

Checking word context provides another way to perform word classification [5]. Performing shallow context analysis means that only the adjacent previous and adjacent next words are considered in the analysis. For example, if given the word sequence:

`President and` **`Chief`** `Executive`

The application will classify the word Chief as a position title word because the word ontology indicates that the previous word `“and”` and the following word `“Executive”` are both title words.

4.5 Predictive Parsing with Weights and Probabilities

All of the previously mentioned techniques of pattern matching, ontology, range, and context need to be glued together in a way that allows the application to predict the categorization for each word in the table. This is achieved through the use of probabilistic weights and confidence levels. For example, the Title column has the following attributes and weights:

Attribute:	Pattern	Ontology	Context	Range	Confidence

Assigned Weight:	0.25	0.5	0.25	0	0.5
Cumulative Test Results	0.25 ✓	0.75 ✓	1.0 ✓	1.0	1.0

Table 1. Attribute and Weight Example

These weights represent probabilities. For example, if given the following word sequence:

```
President and **Chief** Executive
```

If we are testing to see if the current word Chief belongs in the Title column, then the following analysis will be performed:

The weights for all of the passing attributes will be added together. The cumulative test results are compared to the required confidence level. In this case the Title column has a required confidence of .5 and the word Chief passes all of the Title attribute tests. The word has a cumulative weight of 1.0. Since the cumulative weight is greater than the required confidence level, the word will be categorized as a Title word.

5. EXPERIMENTAL RESULTS

5.1 Experiment Results

The experiments were conducted using randomly picked S&P 500 companies. The companies are grouped into the following industries based on Global Industry Classification Standards (GICS) developed by Standard & Poor's and Morgan Stanley Capital International.

- Banks
- Capital Goods
- Energy
- Food Beverage & Tobacco
- Health Care Equipment & Services
- Household & Personal Products
- Insurance
- Materials
- Pharmaceuticals & Biotechnology
- Retailing
- Software & Services
- Technology Hardware & Equipment
- Telecommunication Services
- Utilities

The test sample documents were retrieved from www.sec.gov in its raw text format and are the companies' 2004 Proxy filings.

At least 2 companies of each industry were selected and the total number of companies tested is 36. The total number of compensation records retrieved by ECRS is 323. We conducted a manual retrieval of each of the record for the target companies and used this result as the basis for comparison. It was found out the successful retrieval rate of ECRS is about 81%. This is a very encouraging result considering the DEF-14A document contains a lot of noises and variations across different companies. We have also repeated the tests against 2003 and 2002 Proxy Filings and noticed the successful extraction rates are a bit higher than the 2004 extraction rate. The reason for this slight decreasing rate of successful extraction is due to the fact that more and more companies start to include various HTML tag into their raw text filings, which cannot be recognized by ECRS at this time. Since ECRS is designed to be an open system, studying and enhancing the regular expression pattern base for preprocessing should solve this problem.

5.2 Usage of the executive compensation data

In corporate governance analysis, extremely high or low executive compensation usually implies some kind of hidden problem. An exaggerate amount of stock options grant may cause the executive to cook the financial book and exaggerate the profit. After the stock rise high, he or she can then cash in the stock options and make a good fortune at the sacrifice of shareholders' value. On the other hand, a much lower total cash compensation (salary + bonus) compared to its peer companies may make the executive unwilling to perform. So learning the industry trend of the executive compensation can provide guidance to evaluate compensation issues for specific target company.

Based on the data between 2002 through 2004 retrieved by ECRS, the average salary (c.f. Figure 3), average stock options (c.f. Figure 4) and average total cash compensation (c.f. Figure 5) per executive officer across the 14 industries are charted below.

So we can see that during the past 3 years, Capital Goods and Household & Personal Products industries enjoy the highest salary for its executives; retailing industry tends to give large stock options to its executives while Insurance Companies' executives get the highest total cash compensation. With these analyses, it makes analyst job easier to identify anomalies among individual companies and identify the ones that needs further scrutiny.

There are also many other use cases of executive compensation data. For example if we includes the company's market capitalization and fiscal year end stock price, we can examine the Total Shareholder Return (TSR) and its correlation with the executive compensation. This kind of pay-for-performance analysis right now is a very hot topic in the corporate world.

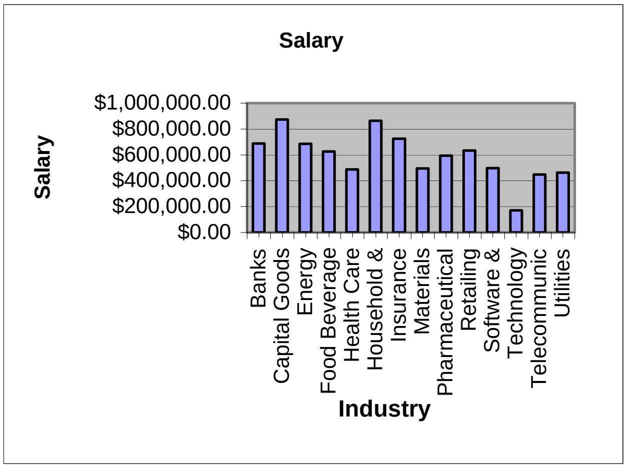


Figure 4. Average Executive Salary By Industry

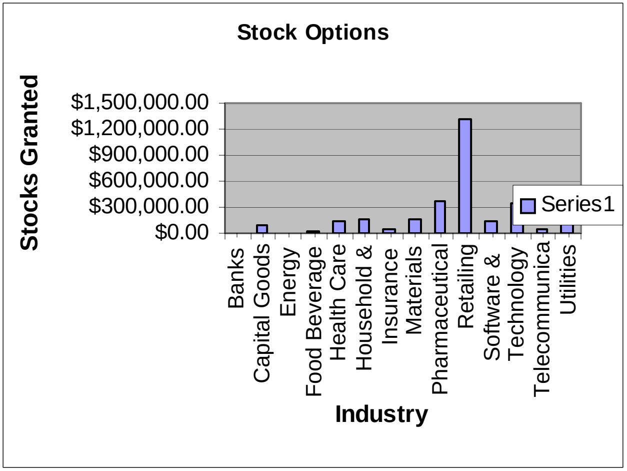


Figure 5. Average Stock Options By Industry

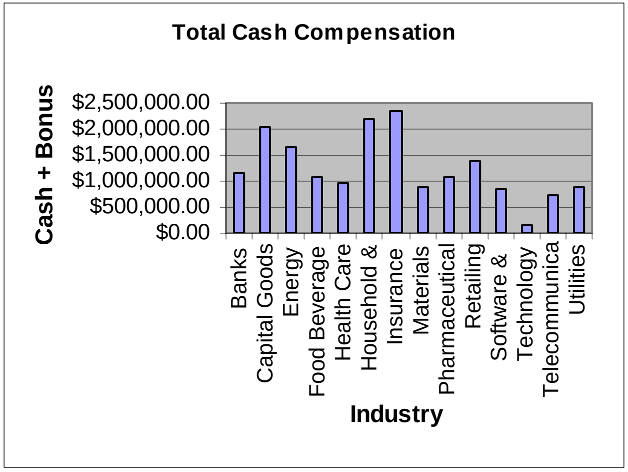


Figure 6. Average Total Cash Compensation By Industry

6. CONCLUSION AND FUTURE WORK

Our experiments have shown SenseNet technology is a promising and practical method to extract executive compensation information from SEC proxy filings. It currently can achieve a 80-90% successful rate in extracting executive compensation records from the summary compensation table in proxy document. The same technique can also be used in extracting regular text such as company related information (e.g. address, contact information) and stock price related information instead of tables. An initial testing result shows the successful extraction rate for such information is well over 90% since such information has less variation and noises.

ECRS is an open system and its pattern-matching library used in preprocessing can be expanded to cover different variations of the documents. Our future enhancements in it will focus on the following directions:

- Expand the ontology for name and positions. Possibly interface ECRS with online providers such as WorldNet[4]. XML Topic Map[6]] is a new data representation and linking technology that may help to keep our ontology comprehensive and update to date.
- Like an Artificial Intelligence (AI) system, ECRS use SenseNet technology to simulate human being to extract information. But as any other AI systems, it can never reach the accuracy level of human being. So in order to make it practical for commercial use, a cooperative system rather than a fully automated system needs to be built. Critical data point checks and alert mechanisms need to be built into the system so that any data anomalies can be captured and corrected by human intervention.

7. ACKNOWLEDGMENTS

Thanks Aaron Stone for the implementation of Java code used in the table analysis case study.

8. REFERENCES

[1] Monks R.A.G., Minow, N., *Corporate Governance 2nd Editon*, Blackwell Publishing, Malden, MA, 2001

[2] Taulli, T., *The EDGAROnLine Guide to Decoding Financial Statements*, J. Ross Publishing, Boca Ration, Florida, 2004

[3] Leinnemann, C. and Schlottmann, F., *Automatic Extraction and Analysis of Financial Data from the EDGAR database*, http://generalupdate.rau.ac.za/infosci/conf/thursday/Leinnemann.htm

[4] Fellbaum, S. (editor), WordNet: An Electronic Lexical Database, published by Bradfords Books, ISBN 0-262-06197-X, 1998.

[5] Reinberger, M.-L. and Daelemans, W., Is shallow parsing useful for the unsupervised learning
of semantic clusters? In *Proceedings CICLing03*. Springer-Verlag, 2003

[6] Park, J. (editor), *XML Topic Maps*, Addison-Wesley, Boston, MA, 2003