AN INTELLIGENT DOCUMENT ANALYSIS SYSTEM FOR EVALUATING CORPORATE GOVERNANCE PRACTICES BASED ON SEC REQUIRED FILING

By

Ying Zheng

A Dissertation

Presented to the Faculty of the Graduate School

of Towson University

in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF SCIENCE

Department of Computer and Information Sciences

TOWSON UNIVERSITY

Towson, Maryland 21252

July, 2016

DISSERTATION APPROVAL PAGE

This is to certify that the dissertation prepared by **Ying Zheng,** entitled **"an Intelligent Document Analysis System for Evaluating Corporate Governance Practices Based on SEC Required Filing",** has been approved by this committee as satisfactorily completing the dissertation requirements for the degree **Doctor of Science** in Information Technology.

7-20-2016

Dr. Harry Zhou
Chair, Dissertation Committee
Date

7/20/2016

Dr. Sungchul Hong
Member, Dissertation Committee
Date

7/20/2016

Dr. Chao Lu
Member, Dissertation Committee
Date

7/20/2016

Dr. Ziying Tang
Member, Dissertation Committee
Date

7/20/2016

Dr. Chao Lu
Chair, Department of COSC
Date

Ramesh K Karne 7-20-2016

Dr. Ramesh Karne
Doctoral Program Director
Date

Janet V Delany 8-8-16

Dr. Janet V. Delany
Dean of Graduate Studies
Date

Dedicated

To

My Grandparents

Table of Contents

Abstract

An IDA is an Intelligent Document Analysis System for Evaluating Corporate Governance Practices capable of retrieving required documentation of public companies from the Securities and Exchange Commission (SEC) and performing analysis and rating in terms of recommended corporate governance practices.

A desired IDA system must be loose coupling, cost-effective, efficient, accurate, as well as operate in real-time. Such a sophisticated system can help individual and institutional investors with evaluating individual companies' corporate governance practices. With the techniques of analogical learning, local knowledge bases, databases, and question-dependent semantic networks, the IDA system is able to automatically evaluate the strengths, deficiencies, and risks of a company's corporate governance practices based on the documents stored in the SEC EDGAR database. (U.S. Securities and Exchange Commission 2013) A produced score reduces a complex corporate governance process and related policies into a single number which enables concerned government agencies, investors and legislators to easily review the governance characteristics of individual companies.

The present manual process not only incurs huge overhead costs, it also has the risk associated with human error and bias in their ratings. To reduce the cost, some companies outsource the job to other countries with a low labor cost. But the problems of how to timely respond to a changing world and constantly updating information remains unanswered. To address those challenges and corresponding issues, first, we developed a system with a user friendly interface and a collection of

knowledge bases, databases and semantic networks, a rating engine and web portal. Then, we ran iterative tests and experiments to identify the drawbacks. Subsequently, we developed effective techniques to address the uncertainty of finding the correct answers, with the purpose of improving the system's performance and ensuring accuracy. Lastly, we investigated and implemented machine learning techniques to efficiently improve the process of the IDA System by using different methods of evaluating similarity.

Acknowledgement

I would like to express my sincere gratitude to my academic advisor, Dr. Harry Zhou, who is a renowned researcher in the field of data mining, machine learning, expert systems, decision-support systems, applications of artificial intelligence in finance and investment. He frequently shares with me potential research topics and encourages me to explore different perspectives of thinking. I would have never been able to earn my doctorate degree without his guidance. Due to his motivation and advice, I believe I will have great success in academic research.

I would also like to thank Dr. Chao Lu, who encourages me to continue my high level study in Towson University. He encourages me to practice by myself and helps me to build confidence in my research.

I would also like to acknowledge Dr. Sungchul Hong, Dr. Chao Lu, and Dr. Ziying Tang for providing guidance on my dissertation. In addition, I would like to thank my colleagues and our research group members who collaborates with me and provided help in my academic endeavors.

I would like to thank my husband, Mr. Zhijiang Chen, for standing by me and helping me during difficult times. Finally, I would like to thank my parents for continuous their support and their deep love.

Towson, Maryland Ying Zheng

July, 2016

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In recent years, corporate governance has become an important concern in investment decision-making. The spate of corporate scandals has sent investor confidence plummeting to an all-time low, and governance of public corporations has moved in a more shareholder-centric direction. Boards need to adjust and prioritize the focus of their corporate agenda, given the above concerns as well as their unique corporate circumstances. Although the details will vary across corporations, the main focus should be on:

- Corporate performance
- CEO selection, shareholders' compensation
- Internal controls, risk oversight and compliance
- Shareholder activism and engagement

Corporations create wealth for shareholders, but their contributions to the economy extend well beyond the return of profit. They can provide employment, support innovation, purchase goods and services, pay taxes, and support various social and charitable programs. Given the important role that corporations play in our society, concerns about the use of corporate power and expectations for the board continue to expand, especially those related to the oversight of risk management, compliance and social responsibility.

In response, boards need to approach these issues with an objective stance and a fiduciary mindset. Boards should work with their management team to ensure that the corporate culture is one that, among other things, encourages employees to come forward with concerns. Boards should assess the quality of the corporation's messaging and communicate with every opportunity that internal reporting is expected, valued and critical to the corporation's success.

Management integrity is also key to building trust with customers, suppliers, employees, regulators and investors. Integrity and trust can be difficult to assess, but should be of particular concern if corporations aim to achieve long-term interests of its shareholders, balance a host of competing special interests and pressures, and address the expectations of the broader society.

Since 1934, the Securities and Exchange Commission (SEC) has required public companies to disclose meaningful financial and other information to the public. In 1984, the office of Investor Education started to implement the Electronic Data Gathering Analysis and Retrieval (EDGAR) system [1], which electronically receives, processes, and disseminates more than 500,000 financial statements every year. The office also maintains a very active website that contains a wealth of information about the Commission and the securities industry, and also hosts the EDGAR database for free public access [2]. The problem is that nobody can go through thousands of documents every quarter to evaluate companies' governance policies. The manual process not only incurs huge overhead costs, it also has the risk associated with human error and bias in their ratings. To reduce the cost, some companies have outsourced this process to other countries with a low labor cost. But the problem of how to timely respond to a changing world and constantly updating information remains unanswered.

1.2 Challenges and Issues

Using natural language to communicate software requirements can be very problematic. While natural language can be easily understood, and has been used widely to write user requirements, it has limitations when being used to communicate system requirements, which are often much more detailed and complex. Therefore, if we were to achieve clarity, it is nearly impossible to explain and convey the complicated details without making the document lengthy and reiterative; on the contrary, if we were to ensure a concise and efficient reading experience for system users, natural language specifications can be confusing and hard to understand:

- Natural language understanding relies on the readers and writers using the same words for the same concept. The ambiguity of natural language will lead to misunderstandings between individuals.

- Natural language requirements specification is overly flexible. The same concept can be related in variety of different methods. It is up to the reader to find out when requirements are the same and when they are distinct.

- There is no easy way to modularize natural language requirements. It may be difficult to find all related requirements. To discover the consequences of a change, you may have to look at every requirement rather than at just a group of related requirements.

Because of these problems, requirements specifications written in natural language are prone to misunderstandings.

1.3 Existing Solution and Its Limitations

The existing solution is a manual process. The problem is that nobody can go through thousands of documents every quarter to evaluate companies' governance policies. The current manual process not only incurs huge overhead costs, but also has the risk associated with human error and bias in their ratings. To reduce the cost, some companies have outsourced this process to other countries with a low labor cost. But the problems of how to timely respond to a changing world and constantly updating information remain unanswered.

1.4 Significance of Proposed Research

In this dissertation, we propose to design and implement an intelligent documents analysis system(IDA) to enable efficiency and accuracy. To do so, we make several contributions:

- First, we extract paragraphs from free text file to answer pre-defined questions. For each question, the IDA system extracts paragraphs from the SEC document, analyzes the paragraph, retrieves the answer, highlights the key words in the synonym list of the semantic net, and displays the score result in the web portal.

- Secondly, we propose to automatically analyze the SEC documents, which can be achieved by implementing the following techniques:

    - We design a question-dependent knowledge base in which rules are used to analyze the possible answer phrases that may exist in a company's proxy filing. A question-dependent database provides the question-related data and information that are needed when retrieving relevant passages. A question-dependent semantic net is manually developed for each individual question in which question-related key words and their synonyms are used to retrieve answers from the company document database.

    - We test and compare several different possible ranking models such as: Vector Space Model (TF-IDF), BM25F Model, Factored Language Model (FLM), Divergence from Randomness (DFR) Models, and Information-Based Models. With experiments, Information-based Models were found to outperform average language models. Based on the result, the Implement Information-Based Model was chosen to be the ranking model for the IDA System.

    - We utilize Fuzzy Logic that represents knowledge and reasons in an imprecise or fuzzy manner for reasoning under uncertainty. Unlike classical logic, fuzzy logic incorporates an alternative way of thinking, which allows the modeling of complex systems using a higher level of abstraction originating from our knowledge and experience.

- Implement an analogical learning and retrieval process. Analogical learning is concerned with the improved performance of a system over time without much human intervention. It is a heuristic approach in which reasoning can be guided, solutions to unfamiliar problems can be constructed, and abstract knowledge can be generalized. We suggest that people use old data in their memory when faced with a problem in which the person must recognize, extract, and compile, from the data, a limited set of features and use it to recall relevant experience.

1.5 Organization of Dissertation Research

This dissertation is structured as follows. In Chapter 2, we discuss the problem domain and possible approaches. In Chapter 3, we conduct a literature review of the Local Knowledge Base, Semantic Net, Fuzzy Logic, Machine Learning, and Ranking Models. In Chapter 4, the system architecture is introduced and the workflow of the system is explained in detail. In Chapter 5, we introduce the local knowledge base in detail, and present examples, which demonstrate the use of local knowledge base. In Chapter 6, semantic net is introduced in details, and we represent question examples to demonstrate the use of semantic net. In Chapter 7, we present hypothetical examples that indicate the necessity and feasibility of fuzzy logic. In Chapter 8, we also introduce machine learning in detail, discuss the necessity and feasibility of machine learning, and represent question examples. In Chapter 9, we represent the implementation of the IDA System, and the final evaluation results based on experiments. Finally, we conclude the dissertation in Chapter 10.

Chapter 2

Problem Domain

2.1 Background

Corporate governance essentially involves balancing the interests of the many stakeholders in a company, including its shareholders, management, customers, suppliers, financiers, government and the community. Since corporate governance also provides the framework for attaining a company's objectives, it encompasses practically every sphere of management, from action plans and internal controls to performance measurement and corporate disclosure.

2.1.1 Importance of Corporate Governance

Following a string of corporate scandals in the United States, legislators and regulators rushed to enact corporate governance reforms, which resulted in the passage of the Sarbanes-Oxley Act of 2002.

Corporate governance is the way a corporation polices itself, and it is important for a variety of reasons:

- Increase the accountability of a company.
- Help the decision making process of a company.
- Avoid massive disasters before they occur.
- Help outside investors to protect themselves against expropriation by the insiders.

- Lower the risk of scandals, fraud, and criminal liability of the company.
- Maintain public image.

2.1.2 Purpose of Corporate Governance Evaluation

The purpose of corporate governance is to persuade, induce, compel, and motivate corporate managers to keep promises they make to investors. Corporate governance is about reducing any actions by management or directors that are at odds with the legitimate, investment-backed expectation of investors. Good corporate governance is simply about keeping promises. Bad governance, then, is simply a promise-breaking behavior. The following is a list of corporate governance principles:

- **Treat all shareholders equally**. Good corporate governance seeks to make sure that all shareholders get a voice at general meetings and are allowed to participate.
- **All stakeholder interests should be recognized** by corporate governance. The stakeholders are important members including people as investors, creditors, customers, suppliers, and employees that don't hold any shares. Taking the time to address stakeholders' interests can help the company establish a good and positive relationship with the community and the press.
- **Board of directors' responsibilities must be clearly identified.** All board members must be on the same page and share a similar vision for the future of the company.

- **Ethical behavior** violations in favor of higher profits can cause massive civil and legal problems down the road. A code of conduct regarding ethical decisions should be established for all members of the board.
- **Business transparency** is the key to promoting shareholder trust. Financial records, earning reports and forward guidance should all be clearly stated without exaggeration or “creative” accounting.

The SEC’s mission is to protect investors; maintain fair, and efficient markets; and facilitate capital formation [3]. One of the major tasks that the SEC conducts is to review corporate financial securities filings at the end of fiscal year ending September 30 each year. The required documents by the SEC including the 8-k, 10-K and DEF 14A (proxy statement) contain passages that are of interest in terms of corporate governance practice. Among the key elements included in the corporate governance section are: bylaws, code of conduct and ethics, board of directors’ information, board committee mandates, basis for executive compensation, and information on insider purchases and sales. Relevant corporate governance and disclosure data should be woven into the company’s 10-K and proxy statement. These filings are an important source of documents for the rating system.

Since the publication of “The 2003 report on Corporate Governance Best Practices: A Blueprint for the post-Enron Era was intended as a sourcebook of leading governance practices for board members and management”, there has been considerable development in the field of corporate governance. In addition to bringing the 2003 report up-to-date, the “Corporate Governance Handbook 2005: Developments in Best

Practices, Compliance, and Legal Standards" aims to provide boards and management with a compendium of the leading corporate governance practices.

Major topics covered in the 2005 book [4] include the following:

- Setting a New Standard for Corporate Governance
- Corporate Governance Practices
- Nominating Corporate Governance Committee Practices
- Compensation Committee Practices
- Audit Committee Practices
- Disclosure, Compliance and Ethics
- Strategy and Enterprise Risk Oversight

From these topics, we have compiled a question list with a selection of 200 plus questions, and manually stored each question in our Question List database. The questions are in the area of Governance, Compensations, and Auditing and Accounting. Some examples include "The Company has poison pills such as preferred stock", "Compensation committee use industry standards to decide on type of compensation and level of compensation", and "Audit committee has a member or an expert on financial statements or GAAP Accounting".

These questions are used to retrieve valuable information about the companies to help investors understand corporate governance of the companies they're willing to invest in. Some of these 200 questions are listed below:

- Does the company provide employment agreements that protect executive officers?
- Is management required to hold accumulated stock for the long-term?
- Does the compensation committee have policies and programs to recapture incentives from management in case of malfeasance?
- Does the company use industry standards to inform their decisions about type of compensation and level of compensation for executives?
- Does the company measure and track its adherence to ethical conduct?

The questions listed above have been addressed through the process of analyzing and retrieving answers from companies' 8-k, 10-k and DEF 14a.

2.2 Approaches

2.2.1 Manual Approach

The manual process can put significant pressure on people, who may be expected to be correct in all details of their work at all times; however, it is only human to make mistakes and overlook things at times. With a manual process the level of service is dependent on individuals and this puts a requirement on management to run training continuously for staff to keep them motivated and to ensure they are following the correct procedures. Regardless of such trainings and policy reinforcement, it is unavoidable that some staff can accidentally mix up details and end up with inconsistency in data entry. This has the effect of making information unusable for reporting or finding answers with data discovery. Nevertheless, reporting and checking data accuracy can be extremely time-consuming and expensive. This is often an area where a significant amount of money can be wasted. We discover the following major disadvantages when using a manual process:

1. Inconsistency in data entry, room for errors, keying mistakes
2. Large ongoing staff training cost
3. Process is dependent on good individuals
4. Reduction in sharing information
5. Time consuming and costly to produce reports
6. Lack of security

To reduce the cost, some companies have outsourced the process or task to other countries with a low labor cost. But the following problems still persist:

1. The inability to keep up with the changing world
2. Delayed information updating
3. Language barriers

2.2.2 Natural Language Software

A system with natural language processing capabilities must address the problems associated with processing very large unstructured text documents. Natural language processing programs struggle with terminology because of term variation, when a concept is expressed in several different ways; and term ambiguity, when the same term is used to refer to multiple concepts. Term variation and ambiguity may cause irrelevant information to be retrieved and relevant information to be overlooked. The most serious problem faced by a natural language processing system is the time complexity problem. That is, the computational time grows exponentially when the number of words increases. It is a common belief that such a system is very difficult, even impossible, to be developed. In response to these research challenges, we developed the IDA System that employs machine learning, question-dependent knowledge bases, databases and question-specific semantic networks in order to keep the problem at a manageable size.

2.2.2.1 Common Difficulties and Problems

Natural language processing (NLP) has come very far in the last thirty years; however, the technology has not yet achieved a revolutionary impact on society. Several critical issues have never been adequately addressed in either theoretical or applied work.

We focus on the following areas, which will have maximum impact when combined with our proposed IDA system:

1. Knowledge acquisition from natural language texts of various kinds, from interaction with a human being, and from other sources. Language processing requires lexical, grammatical, semantic, and pragmatic knowledge. Current knowledge acquisition techniques are too slow and too difficult to use on a wide scale or on large problem. Knowledge bases should be many times the size of current ones.

2. Partial understanding gleaned from multi-sentence language, or from fragment of language. Approaches to language understanding that require perfect

> input or that try to produce perfect output seem doomed to fail because novel language, incomplete language, and erroneous language are the norm, not the exception.

In recent years, from several commercially available systems for database question-answering, it is visible that the result in NLP of transferring technology developed in the 1970s and early 1980s, have been successfully used to improve productivity. The success has depended on the fact that sufficient coverage of the language is possible with relatively simple semantic and discourse models. The semantics are bounded by the semantics of the relations used in the databases and by the fact that words have a restricted number of meanings in one domain.

Natural language processing is successful in meeting the challenges as far as syntax is concerned. However, it still has a long way to go in the areas of semantics and pragmatics.

Since syntax is without doubt the most mature field in both computational linguistics and the closely related field of linguistics. We will only take a closer look in the semantics and pragmatics phenomena of NLP in the following two sections.

2.2.2.1.1 Semantics

For database access, the semantics of the system can be confined to that of the individual entities, classes of entities, relationships among the entities, attributes of the entities, and the typical operations that are performed in database retrieval [5]. This simplifies the problem of semantics in at least the following ways: first, the meaning of individual words and of the phrases they compose can be restricted to the domain-specific meanings actually modeled in the database. Instead of needing to come up with very general semantics for each word, a very literal semantics providing the mapping from the words to the entities modeled in the database is all that is required, as the database could not provide additional information even if more general semantics were available. Second, problems of semantic ambiguity regarding alternative senses for a word are reduced, for only those word senses corresponding to entities in the database will contribute to the search space of possible alternatives. It is hard to deal with exceptions. Representing and inferring world knowledge, commonsense knowledge in particular, is difficult. Semantics of discourse segments is a difficult problem.

For the task of database updating from messages, a key simplifying condition is that the information sought can be characterized ahead of time. For example, we can support that the goal is to update automatically a database regarding takeover bids, or suppose further that the information desired is the date of the bid, the bidder, the target, the percentage of stock sought, the value of the offer, and whether it is a friendly or hostile bit.

2.2.2.1.2 Pragmatics

There are challenges in pragmatics as well. The modeling of context and using context in understanding language are the most difficult, and therefore the least well-understood, areas of natural language processing. Context is all-pervasive and very powerful in natural language communication.

Context is fundamental to communicating substantial information with a few words. For example, a simple declarative sentence stating a fact, *there are compensation arrangements*, is not only a statement of fact but also serves some communication function. The function may be to inform, to mislead about a fact or the readers' belief about a fact, to draw attention, to remind the previously mentioned even or object related to fact, etc. So, the pragmatic interpretation seems to be open ended.

2.2.2.2 IDA Approach

In line of these difficulties, we propose to implement the IDA System with the following features utilizing techniques such as local knowledge base, semantic net, fuzzy logic, and machine learning, also shown in Figure 1:

1. Automatically retrieving related financial document from the SEC website.
2. Analyzing the document based on the question selected form our pre-defined question list using:
    a. A rule-based local knowledge base implemented with simple crisp logic, and fuzzy logic.
    b. A local semantic net stores a synonym list associated to each key word of a question from the question list.
    c. Answer Extraction algorithms to improve the accuracy of the results.
    d. Machining learning techniques to automatically increase the synonyms of the local semantic net.
3. Scoring the retrieved answers, and displaying the most relevant ones.

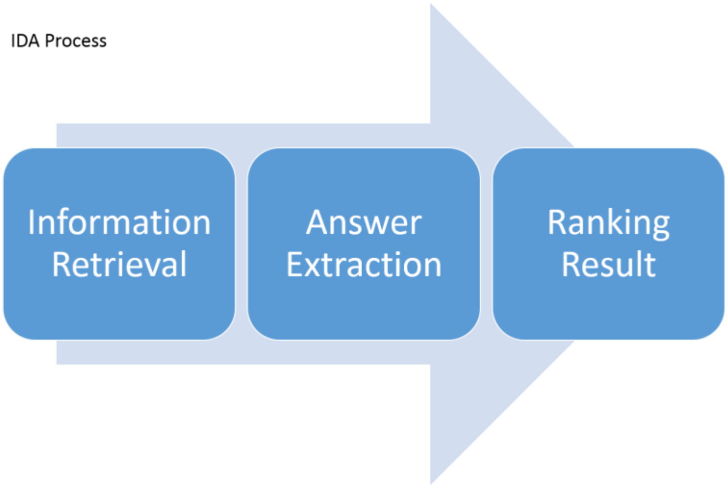


Figure 1: IDA Process

2.3 Summary

In this chapter, first we identify the purpose and importance of corporate governance, and list some required SEC reports containing passages that are of interest in terms of corporate governance practice. Then we present and compare the current approaches with our IDA approach, we illustrate the features in the IDA system, and show the IDA process.

Chapter 3

Literature Review

In this section, we briefly review the related work. Given the multi-disciplinary agenda of our proposed research, we only cover the most related work in this section.

Many research projects on question answering have been carried out. Artificial Intelligence Laboratory of University of Chicago developed FAQFinder, and they have done a lot of research work for English complex questions processing [6]. City University of Hong Kong developed a practical CQA in 2007[7]. Research Institute of Microsoft in Asia retrieved question by identifying the theme and focus of question in CQA [8]. Emory University studied how to judge whether an answer can meet the needs of the questioner [9]. Meanwhile, some scholars predict user satisfaction for question answers in CQA [10]. After submitting a question, they predict user satisfaction based on the answer time, the answer quality and other factors [11].

3.1 Local Knowledge Base

Several areas have focused on expanding or updating a knowledge base given large collections of documents. In the context of question answering, Schlaefer et al. use web retrieval to identify Wikipedia documents with content with extracted 'text nuggets' [12]. They achieved significant improvements in recall using external sources. The TREC knowledge Base Acceleration track [13] performs filtering on a stream of news documents to identify new citation worthy documents from known entities and detects changes in slot values over time. In both of these scenarios the focus is on a single entity and does not include a query topic. In contrast, we focus on identifying documents and entities for a specific user's information needs.

Recent research shows that text query expansion using data extracted from Wikipedia can significantly improve retrieval effectiveness for a variety of retrieval tasks [14] [15]. It is used to enrich keyword representations with explicit semantics (ESA) from Wikipedia [16] to improve clustering and classification tasks. Egozi et al. [17] used pseudo-relevance feedback from ESA annotated text documents to identify concepts and experiment with fusing text and concept-based scores. Instead of mapping all words to concepts, we link entity mentions explicitly.

Wick and McCallum [18] propose query-aware MCMC which focuses inference on a subset of variables in a graphical model. We similarly use the user information needed to focus inference on relevant portions of the document and entity distributions. We use retrieval as a mechanism to measure dependence upon the query.

Dalton and Dietz [19] propose sketch models which are applicable to a board range of corpora and knowledge bases. Jointly modeling a query-specific knowledge sketch from a given general-purpose knowledge base and collection of documents.

Traditionally, textual databases have been allowed to search their contents (words, phrases, etc.) or their structure (e.g. by navigating through a table of contents), but not both at the same time. Recent years, many models have appeared that allow mixing both type of queries.

Mixing contents and structure in queries allows to pose very powerful queries, being much more expressive than each mechanism by itself. By using a query language that integrates both types of queries, the retrieval quality can be potentiated.

Because of this, we see these models as an evolution from the classical ones. Suppose, for example, a typical situation of "visual memory": a user remembers that what he/she wants was typed in *italics*, shortly before a figure that said something about "earth". Searching for the word "earth" may not be a good idea, as well as searching all figures or all the text in italics. What would really help is a language in which we can say "I want a text on italics, near a figure containing the word 'earth'". This query mixes content and structure of the database, and only new models can handle it. [20]

3.2 Semantic Net

Prior work in the field pursued three main directions: comparing text fragments as bags of words in vector space [21], using lexical resources, and using latent Semantic Analysis (LSA) [22]. The former technique is the simplest, but performs sub-optimally when the texts to be compared share few words, for instance, when the texts uses synonyms to convey similar messages. This technique is also trivially inappropriate for comparing individual words. The latter two techniques attempt to circumvent this limitation [16]. Saravanan [23] proposed sentence clustering which aims at grouping sentences with similar meanings into clusters; commonly, vector similarity measures, then a standard clustering algorithm can be applied to group sentences into clusters. One feature of text clustering we will discuss is the generation of relevant term clusters based on lexical semantic relatedness. Mannan, et al [24] proposed a novel integrated ontology approach using query term along with semantic terms for information retrieval. In this approach, classes in the ontology are derived as semantically related keywords for ranking the sections in the document. Searches based on the keywords provides limited capabilities to capture the concept of the user requirement. To solve the limitations of the keyword based search, Thangaraj, et al [25] introduced the idea of semantic search in the field of information retrieval (IR). Text mining is about discovering unknown facts and hidden truth that may exist in the lexical, semantic or even statistical relations of text collections [26]. Liu, et al [27] presents an approach to use context at the lower layer to select the exact meaning of key words, and employs a combination of context, co-occurrence statistics and a thesaurus to group the distributed but semantically related words within a topic to form basic semantic nodes. Most semantic-net-based hypertext systems leave the linking consistency of the net to individual users. Users without guidance may accidentally introduce structural and relational inconsistencies in the semantic nets. Wang, et al [28] tackles the above problems by integrating logical structure and domain semantics into a semantic net. Noah, et al [29] proposed an approach meant to assist in extracting and modeling the semantic information content of web documents using natural language analysis technique and a domain specific ontology. Nesic, et al [30] presents an ontology-driven approach to semantic annotation, indexing and retrieval of fine-grained units of document data. The document units and the user query are both represented by weighted vectors of ontological concepts. Tsatsaronis, et al [31] proposed SemaFor, an indexing algorithm for text documents, which uses semantic spanning forests constructed from lexical resources, and spectral graph theory in order to represent documents for further processing. Becks, et al [32] presented a visualization technique based on a modular approach that allows a variety of techniques from semantic document analysis to be used in the visualization of the structure of technical document collections.

3.3 Fuzzy Logic

Portmann, et al [33] proposed to apply semiotic (i.e. sub-syntactical) and inductive (i.e. probabilistic) methods for inferring concept associations in human knowledge. These associations can be combined to form a fuzzy semantic net representing a map of the knowledge in the web, and provide interactive visualization of these cognitive concept maps to end users, who can browse and search the Web in a human-oriented, visual, and associative interface. Khanum, et al [34] proposed a hybrid system composed of a blend of Fuzzy Logic and Case-Based Reasoning which can lead to a solution where the two approaches cover each other's weaknesses and benefit from each other's strengths. Florez, et al [35] made a variety of modifications to the fuzzy data mining algorithms in order to improve accuracy and efficiency. The authors described an algorithm for computing fuzzy association rules based on Borgelt's prefix trees, modifications to the computation of support and confidence of fuzzy rules, a new method for computing the similarity of two fuzzy rule sets, and feature selection and optimization with a genetic algorithm. The experimental results achieved better running time and accuracy.

3.4 Machine Learning

Wu, et al [36] represented the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in 2006: C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, and CART. The authors provided a description of the algorithm, discussed the impact of the algorithm, and reviewed current and further research on the algorithm. Ashtawy, et al [37] used the following machine learning techniques: linear regression (MLR), multivariate adaptive regression splines (MARS), k-nearest neighbors (kNN), support vector machines (SVM), random forests, and boosted regression trees (BRT)". Depending on the type of ranking/scoring function (SF), and algorithm used, ranking accuracy depends on the size and homogeneity of the training data, and the number of features. Most scoring functions such as force-field-based, knowledge-based, and empirical are not based on machine learning (ML) algorithms. The authors evaluated the performance of several machine learning algorithms with respect to increases/decreases in training sets, and number of features. Li, et al [38] used two machine learning approaches: K-means and Support Vector Machine (SVM) for developing an approach. Modeling steps: 1. Data collection and cleansing, 2. Text sentiment calculation and marking, 3. Hotspot detection based on K-means clustering, and 4. Hotspot forecast based on SVM classification. Joachims [39] explored and identified the benefits of Support Vector Machine (SVM) for text categorization. SVM is well-founded in terms of computational learning theory and very open to theoretical understanding and analysis. Mansuy et al [40] attempt to address some of the uncertainty around the utilization of WordNet in text classification tasks by characterizing the behavior that can be expected in a variety of situations, and report preliminary results obtained from a comprehensive study where WordNet features, part of speech tags, and term weighting schemes are incorporated into two-category text classification models generated by both a Naïve Bayes text classifier and a SVM text classifier. Watson [41] examined four very different Case-Based Reasoning (CBR) applications: 1. Nearest neighbor, 2. Induction, 3. Fuzzy logic, and 4. Database technology. Cheng, et al [42] developed classification predictive models by substructure pattern recognition and different machine learning methods, including SVM, C4.5 decision tree, k-nearest neighbors and random forest.

3.5 Ranking Models

The term frequency-inverse document frequency (TF-IDF) approach is used to weigh each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories [43]. The original TF-IDF term weights are thought to be attribute values of documents that are treated as indivisible objects in many Information Retrieval (IR) models. From our novel perspective, TF-IDF term weights are treated as the outcome of local relevance decision-making at different document locations [44]. TF-IDF is often used to construct a vector space model, which evaluates the importance of a word in a document. The importance increases proportionally with the number of times that a word appears in a document, compared to the inverse proportion of the same word in the whole collection of documents [45]. Shi, et al [46] presented two feature location approaches based on BM25 and its variant BM25F algorithm. Comparisons were done between the two algorithms and the

Vector Space Model (VSM), and Unigram Model (UM). The result shows that the BM25 and BM25F are consistently better than the other two IR methods. Bilmes, et al [47] introduced factored language models (FLMs) and generalized parallel backoff (GPB). Significantly, FLMs with GPB can produce bigrams with significantly lower perplexity, sometimes lower than highly-optimized baseline trigrams. In a multi-pass speech recognition context, where bigrams are used to create first-pass bigram lattices or N-best lists, these results are highly relevant. Amati, et al [48] introduced and created a framework for deriving probabilistic models of Information Retrieval. Highly effective Information Retrieval Models can be generated using the framework. This framework is based on the models of divergence-from-randomness and it can be applied to arbitrary models of IR, divergence-based, language modelling and probabilistic models included. Clinchant, et al [49] introduced the family of information-based models for ad hoc information retrieval. Combined with notions of highlights in BM25, and more recently in DFR models, it can lead to simpler and better models.

Chapter 4

System Architecture

4.1 Overview

Recall what we presented in the Introduction Chapter 1, IDA is an Intelligent Document Analysis System for Evaluating Corporate Governance Practices Based on SEC Required Filings, which is capable of retrieving SEC required documents of public companies and performing analysis and rating in terms of recommended corporate governance practices.

The IDA System consists of a user friendly interface and a collection of knowledge bases, databases and semantic networks, a rating engine and web portal. For every one of the predefined questions, we designed a question-dependent knowledge base in which rules are used to analyze the possible answer phrases that may exist in a company's proxy filing. A question-dependent database provides the question-related data and information that are needed when retrieving relevant passages. A question-dependent semantic net is also developed for each individual question in which question-related key words and their synonyms are used to retrieve answers from the company document database. A Machine Learning process to automatically increase the synonyms in the semantic net. Finally, the ranking module assigns a final score to the company indicating how well its governance policy compares with established standards and then displays the result.

4.2 Components of IDA System

4.2.1 System Interface

There are five components in the System Interface (shown in Figure 2.1): Question Rule Editor, Synonyms Editor Window, Search Result Window, Question List Window, and Synonym List Window. Below is a brief explanation of each:

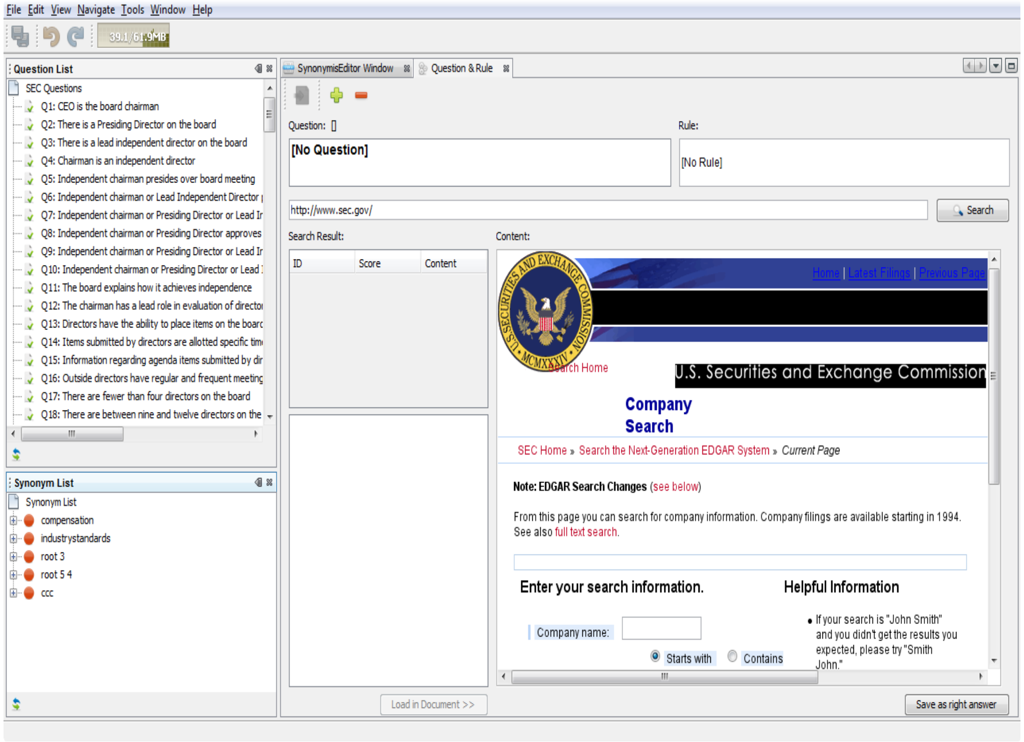
1. Question Rule Editor Window – by using the SEC website link, a company's required financial documents can be retrieved. We then select a specific question from the question list, type the key words in the rule window, perform the search to get the answer to the question.

2. Synonyms Editor Window (shown in Figure 2.2) – a rule can be created by entering the condition part in the provided window. A rule specifies the conditions by which a search is conducted. There are three logical operators to connect conditions: AND, OR and NOT. The system allows any number of synonyms to be defined. Examples such as a AND b NOT c (search documents contains a AND b, but NOT contain c); a OR (b AND c) (search documents contains an OR contains b AND c).

3. Search Result Window – display the score of each related passage in order, display the passage(s) retrieved, and highlights the keyword in the passage(s).

4. Question list Window – Display the predefined questions, which were manually created from the "Corporate Governance Handbook 2005 Developments in Best Practices, Compliance, and Legal Standards", and entered in a local database.

5. Synonym list Window – display a list contains all the synonyms for each rule created for the question.



Question List

SEC Questions

- Q1: CEO is the board chairman
- Q2: There is a Presiding Director on the board
- Q3: There is a lead independent director on the board
- Q4: Chairman is an independent director
- Q5: Independent chairman presides over board meeting
- Q6: Independent chairman or Lead Independent Director
- Q7: Independent chairman or Presiding Director or Lead Ir
- Q8: Independent chairman or Presiding Director approves
- Q9: Independent chairman or Presiding Director or Lead Ir
- Q10: Independent chairman or Presiding Director or Lead
- Q11: The board explains how it achieves independence
- Q12: The chairman has a lead role in evaluation of directo
- Q13: Directors have the ability to place items on the boarc
- Q14: Items submitted by directors are allotted specific tim
- Q15: Information regarding agenda items submitted by dir
- Q16: Outside directors have regular and frequent meeting
- Q17: There are fewer than four directors on the board
- Q18: There are between nine and twelve directors on the

Synonym List

- compensation
- industrystandards
- root 3
- root 5 4
- ccc

Question: []

[No Question]

Rule:

[No Rule]

http://www.sec.gov/

Search Result:

ID	Score	Content

Content:

Home | Latest Filings | Previous Page

Search Home

U.S. Securities and Exchange Commission

Company Search

SEC Home » Search the Next-Generation EDGAR System » *Current Page*

Note: EDGAR Search Changes (see below)

From this page you can search for company information. Company filings are available starting in 1994. See also full text search.

Enter your search information.

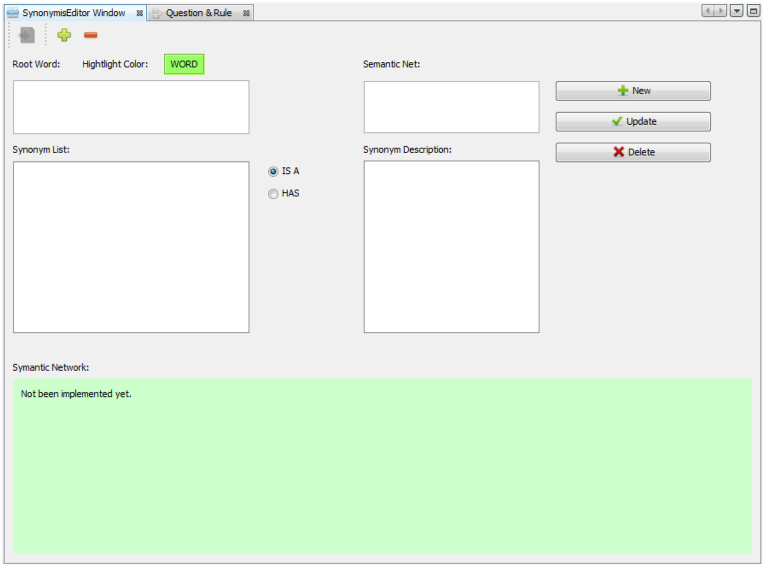
Company name:

● Starts with ○ Contains

Helpful Information

- If your search is "John Smith" and you didn't get the results you expected, please try "Smith John."

Figure 2: Interface



Root Word: Hightlight Color: WORD

Semantic Net:

Synonym List:

● IS A
○ HAS

Synonym Description:

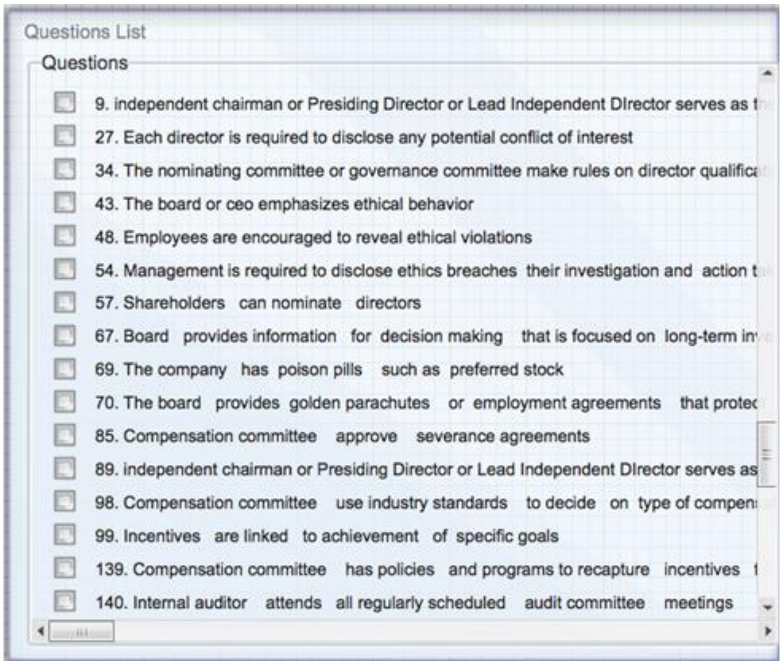
Symantic Network:

Not been implemented yet.

Figure 3: Synonyms Editor

4.2.2 Question List Database

As we discussed in the previous chapter, the 200 plus pre-defined questions were created based on the “Corporate Governance Handbook 2005: Developments in Best Practices, Compliance, and Legal Standards” [4], all these questions are manually stored in a local database, called Question List database (as shown in Figure 4).



□ 9. independent chairman or Presiding Director or Lead Independent DIrector serves as t
□ 27. Each director is required to disclose any potential conflict of interest
□ 34. The nominating committee or governance committee make rules on director qualifica
□ 43. The board or ceo emphasizes ethical behavior
□ 48. Employees are encouraged to reveal ethical violations
□ 54. Management is required to disclose ethics breaches their investigation and action t
□ 57. Shareholders can nominate directors
□ 67. Board provides information for decision making that is focused on long-term in
□ 69. The company has poison pills such as preferred stock
□ 70. The board provides golden parachutes or employment agreements that protec
□ 85. Compensation committee approve severance agreements
□ 89. independent chairman or Presiding Director or Lead Independent DIrector serves as
□ 98. Compensation committee use industry standards to decide on type of compen
□ 99. Incentives are linked to achievement of specific goals
□ 139. Compensation committee has policies and programs to recapture incentives t
□ 140. Internal auditor attends all regularly scheduled audit committee meetings

Figure 4: Question List

4.2.3 Local Knowledge Base

Question-dependent rules are created for each pre-defined question, and stored in the IDA’s local knowledge base.

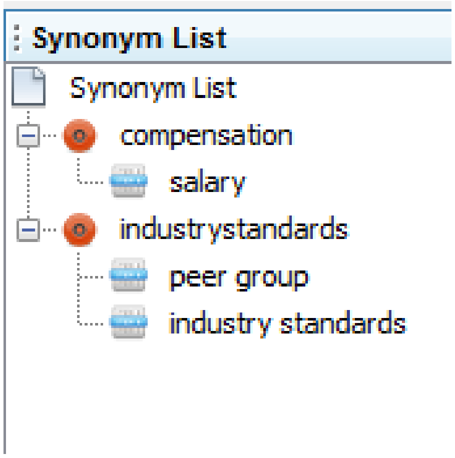
According to Hearst [50] [51], one of the most important part of text mining is Thesaurus Generation. The knowledge-based query construction proposed here uses IDA knowledge rules. The process is activated when IDA is triggered to generate some output, which is used as the input to the query construction process. The rule retrieval function retrieves all the rules that were used in the decision.

Basically, there are three types of knowledge:

- Factual knowledge – facts, think data, such as information about employees, shareholders, company policy, etc., it is stored in the computer memory.
- Sequential knowledge – set of steps in a specific order for each task. The CPU processes instructions sequentially.
- Logical knowledge – expressed in "If … Then" statement, it is rule based knowledge.

4.2.4 Semantic Net

A question-dependent semantic net can be defined to describe a key word in the context of the question. In the IDA system, we implement a synonyms manager (as shown in Figure 5), it is used to manage all synonyms stored in the database. When a new rule is created, synonyms can be added, modified, and deleted when needed. The list was first created and added by the user, it is then expected to conduct an inference reasoning process to discover new synonyms on its own.



Synonym List

- Synonym List
  - compensation
    - salary
  - industrystandards
    - peer group
    - industry standards

Figure 5: Synonym List

4.2.5 Fuzzy Logic

Many decision-making tasks of investors are too complex to be understood quantitatively. However, humans succeed by using knowledge that is imprecise rather than precise. Fuzzy logic refers to a logic system which represents knowledge and reasons in an imprecise or fuzzy manner for reasoning under uncertainty.

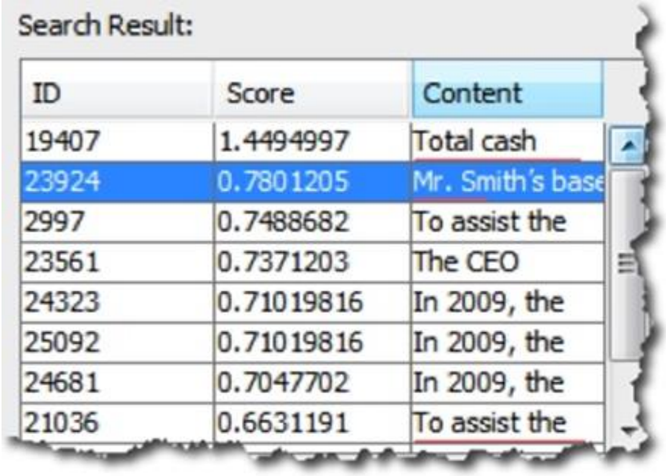
A good strategy in data mining and classification tasks is to use the simplest description of the data that does not compromise accuracy: extract crisp logical rules first, use fuzzy rules if crisp rules are not sufficient, and only if the number of logical rules required for high accuracy of classification is too large use other, more sophisticated tools. In many applications simple crisp logical rules proved to be more accurate and were able to generalize better than many machine and neural learning algorithms [52]. In other applications fuzzification of logical rules gave more accurate results [53]. Crisp logical rules may be converted to a specific form of fuzzy rules (Sect. VIII) and optimized using gradient procedures, providing higher accuracy without a significant increase of the complexity or decrease of comprehensibility of the rule-based system.

4.2.6 Machine Learning

We use Case-based reasoning (CBR) - the machine learning algorithm used to seek answers for new companies based on the solution of similar past company for the same question. The roots of CBR was traced to the work of by Riesbeck and Schank [54]: “A case-based reasoned solves problems by using or adapting solutions to old problems.” Once a target passage has been accepted as the matching answer to the question, the attributes in the target passage are entered into the IDA System’s local semantic net for future use.

4.2.7 Ranking Model

Finally, we can display the score result in the web portal. For example, question *“Compensation committee use industry standards to decide on type of compensation and level of compensation”,* and Intel Corp. is selected as the testing company. The following Figure 6 shows the search result; the highest score is the most relevant answer.

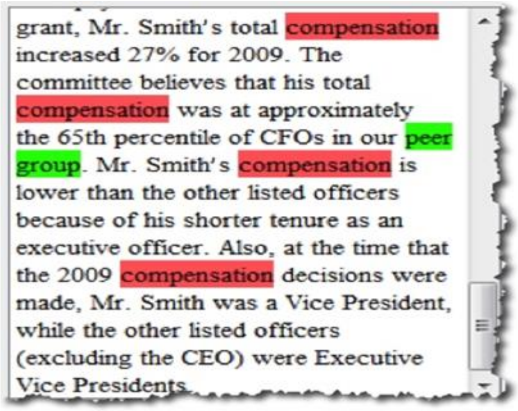


Search Result:

ID	Score	Content
19407	1.4494997	Total cash
23924	0.7801205	Mr. Smith's base
2997	0.7488682	To assist the
23561	0.7371203	The CEO
24323	0.71019816	In 2009, the
25092	0.71019816	In 2009, the
24681	0.7047702	In 2009, the
21036	0.6631191	To assist the

Figure 6: Score Result

The search result for the top score is shown below, and all the key words in the synonym list are highlighted in Figure 7.



grant, Mr. Smith's total compensation increased 27% for 2009. The committee believes that his total compensation was at approximately the 65th percentile of CFOs in our peer group. Mr. Smith's compensation is lower than the other listed officers because of his shorter tenure as an executive officer. Also, at the time that the 2009 compensation decisions were made, Mr. Smith was a Vice President, while the other listed officers (excluding the CEO) were Executive Vice Presidents.

Figure 7: Answer Result

4.3 Workflow

This section presents the basic flowchart of the execution of the IDA System:

1. Information Retrieval – Retrieve filed financial documents of a company under consideration from the SEC website.

2. Answer Extraction – Extract the related passage(s) from the retrieved document using the following techniques: key words parsing, semantic parsing, structure parsing, and organizational parsing.

3. Result Analysis – Using selected ranking module to calculate relevant scores, then the rating module assigns rating scores.

4. Display Result – Display the scores in the result window of the interface for best match.

The previous steps are also presented diagrammatically in the following Figure 8:

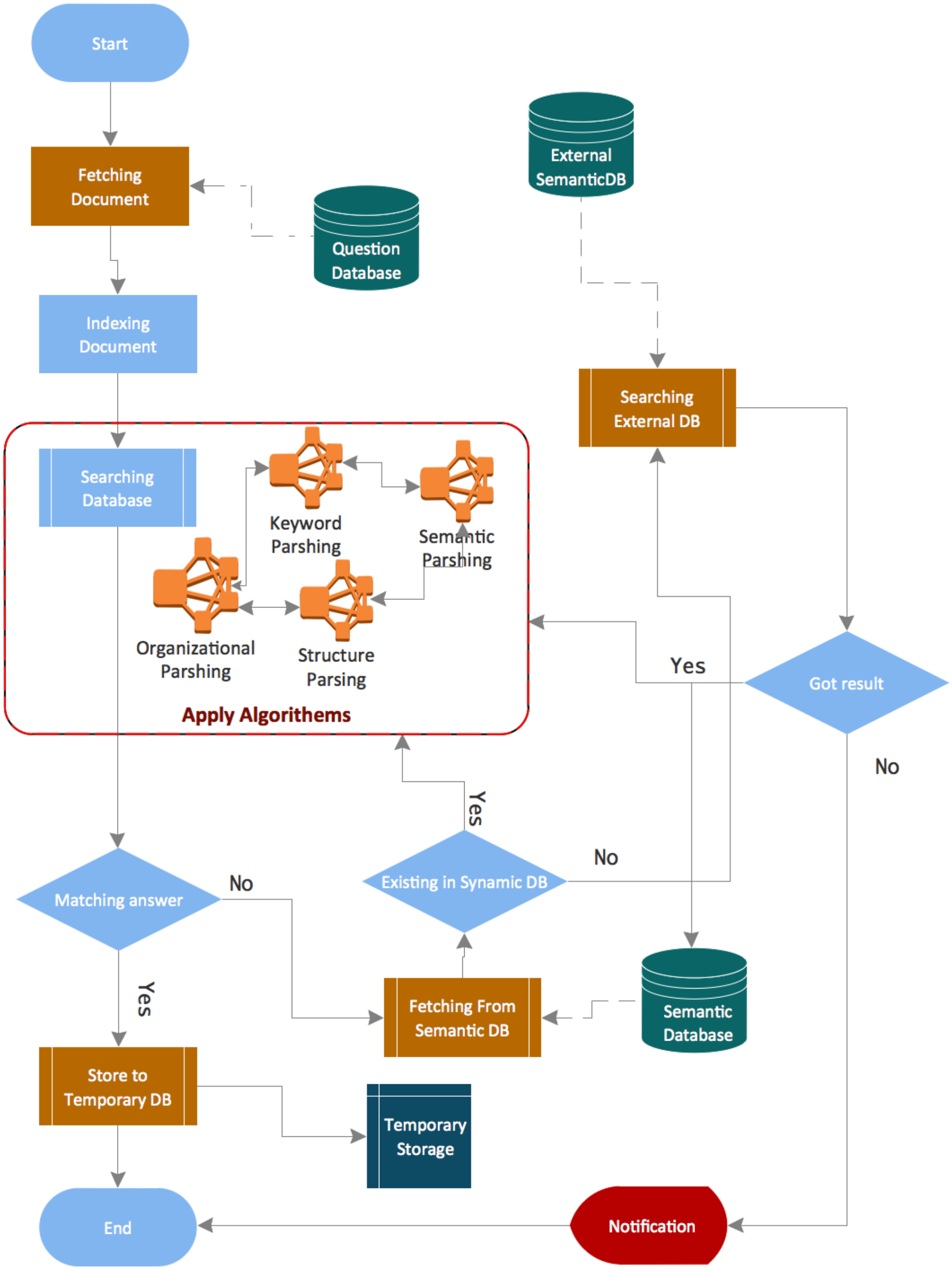


Figure 8: Workflow

The subsections below present Information Retrieval, Answer Extraction, and Result Analysis in details.

4.3.1 Information Retrieval

In order to retrieve the SEC's documents from their EDGAR database to our IDA System, we perform the following steps:

- Get the document's URL from SEC's website
- Use java HttpURLConnection class to get the connection [55]
- Retrieve the document to local computer memory

Once the documents from SEC's website are retrieved, we use Lucene Information Retrieval (IR) engine to perform the search. We implement Lucene in the following steps, as shown in Figure 9:

- Index the stories collection using Lucene Indexer class. This will produce a folder with index files for query search.
- Remove punctuation marks and parse each story as a query using Lucene QueryParser class.
- Search for related stories using Lucene Search class.

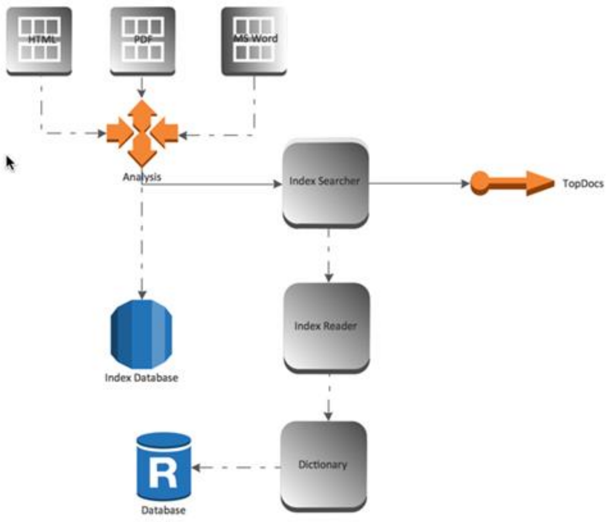


Figure 9: IR Engine

4.3.2 Answer Extraction

The following techniques used in the answer extraction process:

- Use the key words/phrases from a particular question to select rule(s) in the local knowledge bases.
- Use the selected rule(s) to retrieve related passage(s) from the financial document
- If match not found via the key word, the system performs additional steps to perfect the search:
  - Search the synonyms list in the local semantic network.
  - Search the public semantic networks to find similar words/phrases, once match is found, add new synonym to the local semantic network

    for future reuse, perform a machine learning process to automatically increase the synonyms.

    - Perform search in the document utilizing different parsing methods to define new synonyms which are not commonly used in general, add new synonym to the local semantic network for future reuse, again, perform the machine learning process.

- If match not found via the key word, and semantic net, the system uses two more methods to perfect the search:

    - Structure search: checking word context, perform shallow context analysis.

    - Organizational search: perform a novel contextualization approach, estimate the probability of finding the query terms within the passage while considering all term occurrences in the position of the document.

- Question-related measurements computed during the process.

- Analysis performed using the triggered rules during the process.

- Repeat the above steps till the end of the document.

4.3.3 Result Analysis

After possible answers are retrieved, we use the ranking model in Lucene [56] to generate a hit list for each query (story) with all the stories in the collection ranked according to their similarity scores, we then choose the related stories from the top ranked ones using a threshold.

Lucene performs ranked retrieval using a standard tf.idf model [57] [58].

By default, the similarity score of query *q* for document *d* is defined as follows:

$$Score\ (q,d)=\sum_{t\ in\ d} (tf(_{t\ in\ d})\ x\ idf(t)^2\ x\ B_q\ x\ B_d\ x\ L)xC$$

where

$B_q = getBoost\ (t\ field\ in\ q)$

$B_d = getBoost\ (t\ field\ in\ d)$

$L = lengthNorm\ (t\ field\ in\ d)$

$C = coord\ (q, d)\ x\ queryNorm(S)$, *and*

$$S = sumOfSquaredWeight = \sum_{t\ in\ d} (idf\ (t)\ x\ B_q)^2$$

The terms in the above equation can be explained as follows [59]:

- Boost (t field in d) gets the 'boost' for this query clause. Field boosts come in explicitly in the equation and are set at indexing time. The default value of field boost is 1.0.

- lengthNorm (t field in d) is the normalization value of a field, given the number of terms within the field. This value is computed during indexing and stored in the index.

- queryNorm (q) is the normalization value for a query, given the sum of the squared weights of each of the query terms

- coord (q, d) computes a score factor based on the fraction of all query terms that a document contains.

The similarity measure yields a score between 0 and 1, the higher the score, the greater the similarity. The higher score result will display in the IDA System’s result window.

4.4 Summary

In this section, we design the function architecture of the IDA System.   We briefly introduce the whole process of the system, and present each component of the IDA System, and its functions.

Chapter 5

Local Knowledge Base

5.1 Background and Introduction

5.1.1 Background

In general, a knowledge base (KB) is used to store intricate structured and unstructured information used. It is a centralized repository for information, a database of related information about a particular subject. Rule based or knowledge based systems are specialized software that encapsulate human intelligence like knowledge and thereby make intelligent decisions quickly and in repeatable form.

In the 1970s, all the large management information systems store their data in some type of hierarchical or relational database. To distinguish from the term database, the first knowledge based systems had data needs which were the opposite of the database requirements such as flat data, transactions, and large data, etc. An expert system requires structured data. A knowledge base is an object model often called ontology with classes, and instances.

The early expert systems were less complex, and had little need for multiple users. The purpose was just to search for a particular answer, there was no critical demand to store large amounts of data to a permanent memory storage. Researchers found out that the potential benefits of being able to store, analyze, and reuse data and allowing multiple users were exponential.

With the rise of internet documents, hypertext, and multimedia support, knowledge base expert system has evolved to their present form.

Any information model for a knowledge database should be comprised of three parts: text, structure, and query language. It must specify how the text is seen (i.e. character set, synonyms, stop words, hidden portions, etc.), the structuring mechanism (i.e. markup, index structure, type of structuring, etc.), and the query language (i.e. what things can be asked, what the answers are, etc.).

5.1.2 Motivation

The wider capabilities of KBS technology allow for more complex applications, which have a stronger impact on organizational structure than most conventional systems, and which often required a more sophisticated user-system interaction than is the case with conventional systems.

Compared with processes that do not use a knowledge base, such as manual processes, systems with a knowledge base will get the result faster, more efficiently, reliably, and accurately. The rules created can be used and reused again whenever needed.

5.1.3 Challenges

A problem often cited in KBS construction is the knowledge acquisition bottleneck. It turns out it is very difficult to extract the knowledge that an expert has about how to perform a certain task efficiently, in such a way that the knowledge can be formalized in a computer system. The actual realization of a KBS often poses problems as well. The reasoning methods that are used in KBS's are not always fully understood.

The major challenge of any system related to information retrieval is to help its users to find what they need. Knowledge bases are not relational databases, in which the information is already formatted and meant to be retrieved by a "key". In this case, the information is there, but there is no easy way to extract it. The user must specify what he/she wants, see the results, then reformulate the query, and so on, until he/she is satisfied with the answer. Here are some main issues we have discovered:

- Data integrity – how to deal with incomplete, inexact data
- Data inference
- Concurrently selecting rule methods and appropriate analysis
- No standard language for rule definition
- Mechanisms to exchange data between the rule engine and application
- Mechanisms to store the result to a permanent storage
- Mechanisms to reload a previously saved data into the rule engine
- Ability to deal with multiple versions of the rule base and its effect of past results stored in the database

To differentiate from the previous research, we are focusing on addressing the data integrity, etc. problems.

Through our extensive evaluation, the data shows that by implementing a local knowledge base, we can improve performance of the process by retrieving information faster, more efficiently and obtain more reliable and accurate answers to better serve the needs of corporate governance practices.

5.2 IDA Approach

When we first perform the process of Passage Extraction, the most relevant word(s) in the question are highlighted. We then use the Semantic Role Analysis to match relevant passages for each question. This is a rule-based, language-independent technique for mapping textual fragments to arbitrary labels. According to Hearst [50] [51], other than question answering, information retrieval, and information extraction; one of the most important parts of text mining is Thesaurus Generation. The knowledge-based query construction proposed here uses IDA knowledge rules. The process is activated when IDA is triggered to generate some output, which is used as the input to the query construction process. The rule retrieval function retrieves all the rules that were used in the decision. A parsing process then begins. The system evaluates the similarity of the answer and question and actually calculates the similarity between sentences, which is influenced by many factors [60]. In our approach, we mainly consider four factors: key words parsing, semantic parsing, structure parsing and organizational parsing, which are explained in details in the following sections.

5.2.1 Key Words Parsing

In this step, the control structures used in the IDA are parsed. There are a number of variations of If-Then statements that can be used in the System logic, such as Simple If-Then statement, IF-Then-Else statement, If-Then-Elseif statement, Switch statement, Call statement, and others [61]. In IDA System, only two types of the above statements are analyzed according to the methods described in the following sections.

5.2.1.1 Simple If-Then Statement

The type of statement is divided into two sentences, the If sentence and the Then sentence. The concepts in the If sentence are recognized as condition concepts, and the concepts in the Then statement are recognized as decision concepts. Consider the one question *"The board or CEO emphasizes ethical behavior"*. One of the question-dependent rules is presented below:

```
If (k = "ethical") Then
        $a$ = "relevant passages"
```

Where k is the key word of the question sentence, *a* is the answer sentence.

We perform experiments using the above question, although we know that it is hard to use just the key word without its synonyms to retrieve the correct answer due to the diversity of natural language, we are still able to find some result shown in the Table 1 below.

It is possible to have more than one relevant passage in a single question. Our strategy for passage extraction assumes that the optimal passage(s) in a question should have all the query key words with respect to a question under consideration. Also they should have higher density of query concepts than other fragments of text in the question.

Company	**Passage**	**Relevance**	**Analysis**
Deltek, Inc. 2014	All of our directors have demonstrated business experience and acumen and a commitment of service to our Board of Directors. Our directors also possess significant executive leadership experience, knowledge and expertise derived from their service as executives of various businesses and, in many instances, as members of the boards of directors of other public and/or private companies. In their conduct, we believe our directors have displayed integrity, honesty and adherence to high ethical standards.	True	1 passage found using the key word.
Chevron Corp 2014	We are honest with others and ourselves. We meet the highest ethical standards in all business dealings. We do what we say we will do. We accept responsibility and hold ourselves accountable for our work and our actions.	True	4 passages found, represent the best answer.
3M Co 2014	1. Personal qualities of leadership, character, judgment, and whether the candidate possesses and maintains throughout service on the Board a reputation in the community at large of integrity, trust, respect, competence, and adherence to the highest ethical standards. 2. All of our employees, including our Chief Executive Officer, Chief Financial Officer, and Chief Accounting Officer, are required to abide by 3M's Code of Conduct to ensure that our business is conducted in a consistently legal and ethical manner. 3. The Board also has adopted a Code of Business Conduct and Ethics for directors of the Company. This Code incorporates long-standing principles of conduct the Company and the Board follow to ensure the Company's business and the activities of the Board are conducted with integrity and adherence to the highest ethical standards, and in compliance with the law.	True	3 passages found, all relevant.

Table 1: Keyword Result 1

For example, ‘presiding’, *and* ‘director’ *are the keywords for the example question* “There is a Presiding Director on the board.” *The rules in the knowledge bases contain the keyword(s), the conditions in a rule are connected by three types of logical operators: AND, OR and NOT, these logical operators are parsed in the same order in which they occurred. The logical operator “NOT” has a key role in undesired elements from the retrieval set. Based on this operator parsing, concepts are extracted as operands of the parsed operators for query construction. As an example, consider the previous question* “There is a Presiding Director on the board.” *The question-dependent rules are presented below:*

```
If (k = “presiding” AND k = “director”) Then

        a = “relevant passages”
```

Table 2 below shows the research result for the above question.

Company	**Passage**	**Relevance**	**Analysis**
3M Co 2015	Independent Lead Director The Board has designated one of its members to serve as a Presiding Director, with responsibilities that are similar to those typically performed by an independent chairman ("Lead Director"). Michael L. Eskew was appointed Lead Director by the independent directors effective November 12, 2012, succeeding Dr. Vance Coffman who had served as Lead Director since 2006. Michael Eskew is a highly experienced director, currently serving on the boards of The Allstate Corporation, International Business Machines Corporation, and Eli Lilly and Company, and was the former Chairman and CEO of United Parcel Service, Inc.	True	5 passages found, 1 relevant
Alcoa Inc. 2015	Judith M. Gueron is our current Lead Director. Ms. Gueron, whose term as a director expires at the annual shareholders meeting on May 1, 2015, will not be standing for re-election. In accordance with the directors’ retirement policy in the Company’s Corporate Governance Guidelines, Ms. Gueron has announced her intention to retire from the Board effective May 1, 2015 as she would reach age 75 during a new three-year term.	True	Passage found with Lead Director
Caterpillar Inc. 2014	Our Chief Executive Officer also serves as the Chairman of the Board and we have an independent director who is elected by the Board to serve as the Presiding Director, with broad	True	7 passage found, 1 relevant

	authority and responsibility over Board governance and operations. Eleven of our twelve director nominees are independent. See "Board Composition and Leadership Structure" on page 5 for more information.		

Table 2: Keyword Result 2

5.2.1.2 Key Word Similarity

Key word similarity is the similarity degree between question and answer. It is measured by the number of same words contained in both question and answer. It is calculated as following:

$$KeySim(q,a) = \frac{\mathrm{Same}(\mathrm{q},a)}{Key(q) + Key(a) - Same(q,a)}$$

Where *KeySim(q, a)* represents the number of the exact same keywords in the question q and answer a. Key(q) indicates the number of keywords in question q, and Key(*a*) indicates the number of keywords in answer *a*. If the same keyword appears more than once, it is counted only once.

5.2.2 Semantic Parsing

We noticed from many searches that using exact keyword(s) might not lead to the correct answer. Most of the time, synonyms of the keyword(s) will retrieve the correct passages.

5.2.2.1 If-Then-Elseif Statement

The Elseif clause is treated similar to a simple If-Then statement, with Elseif similar to If.

Since there is no template provided by the SEC and the documents are written by different people and submitted by different companies, the IDA System is basically conducting text mining in a free text domain. Take a look at the question, “Does the compensation has policies and programs to recapture incentives from management in case of malfeasance?”, the keyword *“recapture”* may be expressed by different terms, such as *“recoup”, “deduct”, “take out”,* and *“stuck off”.* In order to retrieve the correct answer, a semantic network needs to be developed to store a synonym list for the term *“recapture”* for this question, which will be presented in detail in Chapter 6. The rule set for this question can be developed as following:

```
If (c = “recapture” AND c = “incentives”) Then
    d = “relevant passages”
Elseif (c in (synonym list of “recapture”)) Then
    d = “relevant passages”
```

Where c = “” is the condition for If clause, when it’s true, retrieve the relevant passages. Elseif c is in the synonym list of *“recapture”* ， retrieve the relevant passages.

From the examples in Table 3, we found out that the exact key words from the question don’t appear in every report of different companies. New synonyms need to be added to a local semantic net (detail explained in Chapter 5) to improve the performance of the system each time a search is done. Our approach works more efficiently than the manual approach, and previously discussed natural language approaches.

5.2.2.2 Semantic Similarity

The calculation of semantic similarity is based on the word semantic calculation. Qun and Sujian [62] introduced the similarity calculation based on HowNet, Xianfeng and Pengfei [60] used the same method. In our approach, we also use this method to calculate semantic similarity.

Given the questions q and the answer a, q contains a set of keywords k_{11}, k_{12}, … k_{1i}, also contains a set of keywords k_{21}, k_{22}, … k_{2j}. The similarity between k_{1i} $(1 \leq i \leq n)$ and k_{2j} $(1 \leq j \leq m)$ can be expressed as *Sim*(k_{1i}, k_{2j}). The following calculation represents the semantic similarity between the question q and the answer a:

$$SemSim(q,a) = \frac{1}{2}(\frac{1}{n}\sum_{i=1}^{n} \max\{Sim(k1i, k2j) | 1 \leq j \leq m\} + \frac{1}{m}\sum_{i=1}^{m} \max\{Sim(k1i, k2j) | 1 \leq i \leq n\})$$

Company	**Passage**	**Relevance**	**Analysis**
Alcoa, Inc. 2015	Recovery of Incentive Compensation If the Board learns of any misconduct by an executive officer that contributed to the Company having to restate all or a portion of its financial statements, it shall take such action as it deems necessary to remedy the misconduct, prevent its recurrence and, if appropriate, based on all relevant facts and circumstances, take remedial action against the wrongdoer in a manner it deems appropriate. In determining what remedies to pursue, the Board shall take into account all relevant factors, including whether the restatement was the result of negligent, intentional or gross misconduct. The Board will, to the full extent permitted by governing law, in all appropriate cases, require reimbursement of any bonus or incentive compensation awarded to an executive officer or effect the cancellation of unvested restricted or deferred stock awards previously granted to the executive officer.	True	- *Recapture*: recovery, reimbursement - *Incentives*: incentive, compensation, bonus
ADT Corp 2016	Pay Recoupment Policy The Company's pay recoupment policy provides that, in addition to any other remedies available to it and subject to applicable law, the Company may recover any incentive compensation (whether in the form of cash or equity) paid by the Company to any Executive Officer that resulted from any financial result or operating metric that was impacted by the Executive Officer's fraudulent or illegal	True	- *Recapture*: Recoupment, recover - *Incentives:* incentive, compensation

	conduct. Our Board of Directors has the sole discretion to make any and all determinations under this policy. The Compensation Committee periodically reviews this policy to determine whether any changes are warranted.		
Boeing 2014	Clawback Policy We will require reimbursement of any incentive payments to an executive officer if the Board determines that the executive engaged in intentional misconduct that caused or substantially caused the need for a substantial restatement of financial results and a lower payment would have been made to the executive based on the restated financial results. This policy is described in our Corporate Governance Principles.	True	- *Recapture:* Clawback, reimbursement - *Incentives:* incentive

Table 3: Semantic Result

5.2.3 Structure Parsing

It is highly possible the system will not be able to recognize highly complex nominal phrases during the free text processing [63]. Assuming that a specific question sentence is associated with certain verbs or verb groups which trigger the answer passage, then it will be very difficult to find the appropriate fillers without knowing the correct clause structure. We propose a divide and rule parsing strategy:

1. First, only the verbs or verb groups and the topologic structure of the sentence according to the linguistic field.

2. Next, general phrasal grammars are applied to the contents of the different fields.

Checking word context provides another way to perform word classification [64]. Performing shallow context analysis means that not only the verbs or verb groups are considered in the analysis, other words in the sentence are also being analyzed during the processing.

The divide and rule parsing relies on a configured preprocessing strategy in order to achieve the desired simplicity and performance. Word segmentation processes mapped characters into larger units called tokens and identifies their types. Currently, we use generic classes for semantically ambiguous tokens, for example, abbreviations such as “CEO”. The variety of token classes simplifies the analyzing process of the consecutive section of the question. Each token identified as a potential word form is submitted to the morphological analysis. Each token recognized as a valid word form is associated with the list of its possible readings, characterized by stem, inflection information and part of sentence. Let us take a look at the following example in Table 4 using the question, *“The board explains how it achieves independence.”*

In this case, if we only consider the verb in the question during the parsing process, it is almost impossible to retrieve the correct answer. Noticed that *“director”, “independence”,* and *“standard/guideline/policy”* almost appears in every answer. We then use structure similarity method to calculate the similarity degree between the question and the answer.

Company	**Passage**	**Relevance**	**Analysis**
3M Co. 2015	The Board has adopted a formal set of Director Independence Guidelines with respect to the determination of director independence, which either conform to or are more exacting than the independence requirements of the NYSE listing standards, and the full text of which is available on our Web site at www.3M.com, under Investor Relations — Corporate Governance. In accordance with these Guidelines, a director or nominee for director must be determined to have no material relationship with the Company other than as a director. The Guidelines specify the criteria by which the independence of our directors will be determined, including strict guidelines for directors and their immediate family members with respect to past employment or affiliation with the Company or its independent registered public accounting firm. The Guidelines also prohibit Audit and Compensation Committee members from having any direct or indirect financial relationship with the Company, and restrict both commercial and not-for-profit relationships of all directors with the Company. Directors may not be given personal loans or extensions of credit by the Company, and all directors are required to deal at arm's length with the Company and its subsidiaries, and to disclose any circumstance that might be perceived as a conflict of interest.	True	
Alcoa Inc. 2015	In its Corporate Governance Guidelines, the Board recognizes that independence depends not only on directors' individual relationships, but also on the directors' overall attitude. Providing objective, independent judgment is at the core of the Board's oversight function. Under the Company's Director Independence Standards, which conform to the corporate governance listing standards of the New York Stock Exchange, a director is not considered "independent" unless the Board affirmatively determines that the director has no material relationship with the Company or any subsidiary in the consolidated group. The Director Independence Standards comprise a list of all categories of material relationships affecting the determination of a director's independence. Any relationship that falls below a threshold set forth in the Director	True	

	Independence Standards, or is not otherwise listed in the Director Independence Standards, and is not required to be disclosed under Item 404(a) of SEC Regulation S-K, is deemed to be an immaterial relationship.		
Intel Corp 2015	The NASDAQ rules have objective tests and a subjective test for determining who is an “independent director.” In addition to the Board-level standards for director independence, the directors who serve on the Audit Committee each satisfy standards established by the U.S. Securities and Exchange Commission (SEC), as no member of the Audit Committee accepts directly or indirectly any consulting, advisory, or other compensatory fee from the company other than their director compensation, or otherwise has an affiliate relationship with the company. Similarly, the members of the Compensation Committee each qualify as independent under NASDAQ standards. Under these standards, the Board considered that none of the members of the Compensation Committee accept directly or indirectly any consulting, advisory, or other compensatory fee from the company other than their director compensation, and that none have any affiliate relationships with the company or other relationships that would impair the director’s judgment as a member of the Compensation Committee.	True	

Table 4: Structure Result

5.2.3.2 Structure Similarity

This method is to mark the similarity of sentences from the position of the keywords or its synonyms. It reflects the similarity degree of the same words or synonyms of question q and answer *a* in the position, which can be measured by the number of the adjacent sequence inverse of the same words or synonyms. The following calculation represents the method:

$$StrSim(q,a) = 1 - \frac{Rev(q,a)}{MaxRev(q,a)}$$

Where *MaxRev*(*q*, *a*) represents the maximum reverse of natural sequence that the question q and answer *a* has the same number of keyword, *Rev(q, a)* represents the reverse of natural sequence that is constituted by keywords of the question *q* in the position of the answer *a*.

5.2.4 Organizational Parsing

We also present a novel contextualization approach for passage retrieval. The approach leverages the fundamental principle underlying the positional language model (PLM) [65] that was introduced for the document retrieval task. The key idea behind PLM is that any term in a document is represented as a density function which expresses the probability of finding the term at any position in the document. Similarly, our model estimates the probability of finding the query terms within the passage while considering all term occurrences in the position of the document. The method we devise lets any occurrence of a query term in the document affect the passage score regardless of whether the occurrence is actually in the passage. This effect depends on the distance between the query term occurrences and the passage. Thus, the whole document, or more precisely, all query term occurrences in the document, provide context for the passage. Evaluation performed with the INEX focused-retrieval benchmarks shows that our approach substantially improves over previously proposed passage retrieval methods, including those that use various contextualization techniques [66].

Take a look at the question, "*The company has poison pills such as preferred stock.*" The keyword is poison pills, but the question is what a poison pill is? In the financial world, a poison pill is a corporate strategy to prevent hostile takeovers. When a company has a poison pill, the acquirer is less attracted to the stock of the target company. There are two types of poison pills:

1. Flip-in: it allows existing shareholders to buy more shares at a discount price.
2. Flip-over: it allows stockholders to buy the acquirer's shares at a discounted price after the merger.

Through research, we determined that more than 90% of the companies with poison pills do not have poison pill or its synonyms in their report. The detailed plan is located in the Employee Stock Purchase Plan section of the report. The rest of the companies without a poison pill just simply state that there is no poison pill. The result and analysis is presented as following Table 5:

Company	**Passage**	**Relevance**	**Analysis**
Citrix Systems, Inc. 2015	*2015 Employee Stock Purchase Plan (ESPP)* Our Board believes it is in our best interest and in the best interest of our shareholders that the 2015 ESPP be approved. Shareholders are requested in this proposal to approve the 2015 ESPP. Eligible employees have been granted options to purchase common stock under the 2015 ESPP beginning on January 16, 2015, with such options conditioned on approval of the 2015 ESPP by our shareholders, and no additional offering of options to purchase stock under the 2005 ESPP will be made. The 2015 ESPP allows all full-time and certain part-time employees to purchase shares of Citrix common stock at a discount to fair market value. The 2015 ESPP is an	True	No keyword No synonym

	important component of the benefits package that we offer to our employees. We believe that it is a key factor in retaining existing employees, recruiting and retaining new employees and aligning and increasing the interest of all employees in our success.		
ABM 2016	*Approve Amendment of the 2004 Employee Stock Purchase Plan (ESPP)* Under the ESPP, shares will be purchased at a price equal to 95% of the fair market value of one share of Common Stock on the day on which the shares are purchased. On December 31, 2015, the closing price for the Common Stock of the Company was $28.47 per share. The number of shares of Common Stock that a participant purchase in each offering period is determined by dividing the total amount of payroll deductions withheld from the participant’s compensation during the offering period by the purchase price. Shares purchased pursuant to the ESPP are subject to a minimum holding period of six months following purchase before sale of the shares shall be permitted.	True	No keyword No synonyms
Hewlett Packard Enterprise 2015	Stockholder Rights No “poison pill” (stockholders’ rights plan)	True	Exact keyword
Intel Corporation 2015	The Board’s policy is to obtain stockholder approval before adopting any “poison pill.” If the Board later repeals this policy and adopts a poison pill without prior stockholder approval, the Board will submit the poison pill to an advisory vote by Intel’s stockholders within 12 months. If Intel’s stockholders do not approve the Board’s action, the Board may elect to terminate, retain, or modify the poison pill in the exercise of its fiduciary responsibilities. The Board has adopted a policy that the company will not issue shares of preferred stock to prevent an unsolicited merger or acquisition.	True	Exact keyword

Table 5: Organization Result

5.3 Summary

In this chapter, we represent a rule-based local knowledge base. Within the local knowledge base, we implement four different parsing methods to analyze the financial document to retrieve the answer passages. With our approach, the IDA System provides a more effective, accurate, and faster process in answer extraction.

Chapter 6

Semantic Net

6.1 Background and Introduction

6.1.1 Background

A semantic net presents knowledge in patterns between objects. It can be represented by a graphic network with a node linking to other nodes with relations, where a node is an object, or concept, or other kind of entity. In a semantic network, these nodes are represented by terms and their relationships to other terms in the network. For example, to place the concept 'website' in a semantic net, one might begin by saying that it contains information, runs on computers, and serves users. The link types are 'contains', 'runs on', and 'serves', the nodes are 'website', 'information', 'computers', and 'users' (Figure 10).

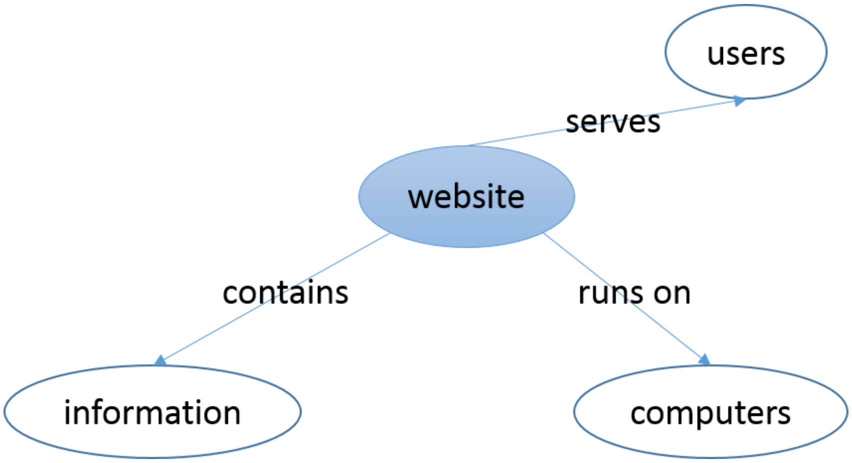


Figure 10: Semantic

Some semantic links manifest inheritance [67]. For example, if the network connects the node 'teacher' to the node 'person' with the link 'is a', then one can infer that the properties of 'teacher' are inherited from those of 'person'. Inheritance conveys transitivity. If a person is a female, then by transitivity a teacher is a female. The basic idea is:

> **If** two words occur frequently in the same context, such as same page, paragraph, sentence, or part of speech.
>
> **Then** there must be some semantic relation between them

Originally, semantic networks stem from the "existential graphs" introduced by Charles Peirce in 1896 to express logical sentences as graphical node-and-link diagrams [68]. Later on, similar notations have been introduced, such as conceptual graphs [69], all differing slightly in syntax and semantics. Despite these differences, all the semantic network formalisms concentrate on expressing the taxonomic structure of categories of objects and the relations between them.

6.1.2 Motivation

Semantic networks have a lot to offer to data integration tasks. They enable the identification of similarities between text attributes from difference sources, and improve the data integration accuracy and efficiency. Semantic representation of text sources is more direct and the discovery of semantic mapping between the various sources and the medicated schema [70] is more straightforward.

6.2 Our Approach

This is a rule-based, language-independent technique for mapping textual fragments to arbitrary labels. According to Hearst [50] [51], other than question answering, information retrieval, and information extraction; one of the most important parts of text mining is Thesaurus Generation. The knowledge-based query construction proposed here uses IDA knowledges rules. The process is activated when IDA is triggered to generate some output, which is used as the input to the query construction process. The rule retrieval function retrieves all the rules that were used in the decision.

Basically, there are three type of knowledge:

- Factual knowledge – facts, think data, such as information about employees, shareholders, company policy, etc., it is stored in the computer memory.
- Sequential knowledge – set of steps in a specific order for each task. The CPU processes instructions sequentially.
- Logical knowledge – expressed in "If … Then" statement, it is rule based knowledge.

As an example, consider the previous question "*Does the compensation committee have policies and programs to recapture incentives from management in case of malfeasance?*" One of the question-dependent rules is presented below:

IF *there exists a term "recapture" AND a term "incentives"*

Then *retrieve relevant passages*

Since there is no template provided by the SEC and the documents are written by different people and submitted by different companies, the IDA System is basically conducting text mining in a free text domain. A semantic network needs to be developed specifically for the above question in order to match the key word “recapture,” which may be expressed by different terms, such as “recoup”, “deduct”, “take out", and “stuck off”. See the following semantic network developed for this question. Similarly, the term "incentives" is also linked to other synonym terms such as “compensation”, “bonus”, and “salary” (Figure 11).

The above semantic net is represented as a directed, labeled graph with nodes N = $\{n_0, n_1 \ldots n_i\}$ and links $\{(n_j, n_k, l) \mid n_j, n_k \quad \text{N}; l \quad label\}$ where $label = \{is\text{-}a\}$.

The networks also embed a set of synonyms, relations, and concepts relevant only to a particular question. If needed, we can use the data in the database to compute question related measurements. Using the question “*Compensation committee use industry standards to decide on type of compensation and level of compensation*” as an example, the average salary level for a CEO in different industries may be different depending on the market value of the company. By employing a set of pre calculated data stored in the database, the IDA System is able to find an appropriate compensation level for each CEO in a particular industry. For example, the CEO’s average salary in a grocery company of a market value in the range of 100 million is different than the CEO's average salary in an IT industry of a market value in the range of 1000 billion. With this question-dependent database, compensation standards in different industries can be easily located and retrieved for the purpose of passage analysis and retrieval.

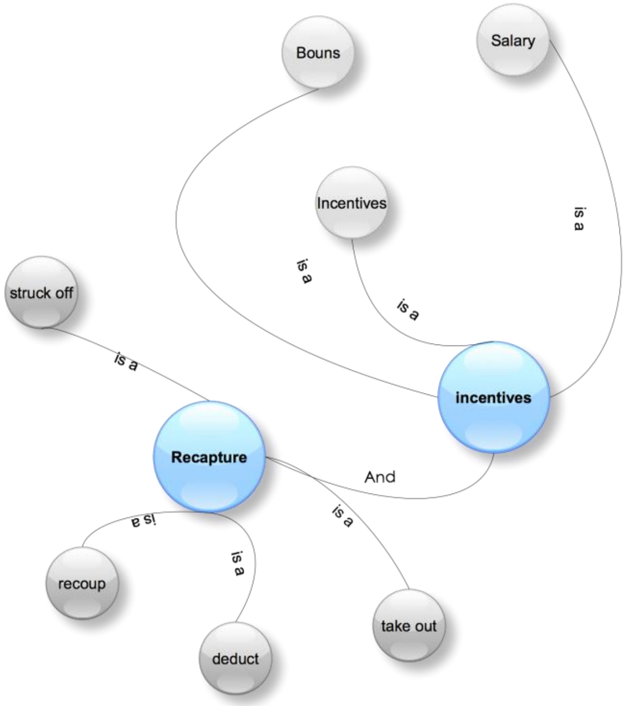


Figure 11: Semantic Net

6.3 Summary

In this chapter, we present the rule-based, question-dependent local semantic net of the IDA system. The implementation of the semantic net improves the performance of the IDA system. It allows the system to perform searches more efficiently, and accurately.

Chapter 7

Fuzzy Logic

7.1 Background and Introduction

7.1.1 Background

Fuzzy logic is a computing approach based on “degrees of truth” rather than the usual modern computer approach based on “true or false” (1 or 0) Boolean logic.

The idea of fuzzy logic was first introduced by Dr. Lotfi Zadeh [71] in 1960s, Dr. Zadeh was working on the problem of a computer's understanding of natural language, which cannot be easily translated into the absolute terms of 0 and 1. German mathematician Dieter Klaua also introduced a fuzzy set theory with a graded membership predicate and a graded equality relation in 1965 [72]. At the same time, Salii [73] defined more general kinds of structures called L-relations, which he studied in an abstract algebraic context.

There are many mathematical constructions similar to or more general than fuzzy sets. Since fuzzy sets were introduced in 1965, a lot of new mathematical constructions and theories treating imprecision, inexactness, ambiguity, and uncertainty have been developed. Some of these constructions and theories are extensions of fuzzy set theory, while others try to mathematically model imprecision and uncertainty in a different way [74] [75] [76].

Fuzzy logic that includes 0 and 1 are extreme cases of truth (or “the state of matters” or “fact”) but also includes the various states of truth in between so that, for example, the result of a comparison between two things could be not “tall” or “short” but “.38 of tallness.”

Fuzzy logic seems closer to the way our brains work. We aggregate data and form a number of partial truths which we aggregate further into higher truths which in turn, when certain thresholds are exceeded, cause certain further results such as motor reaction. A similar kind of process is used in an artificial computer neural network and expert systems.

It may help to see fuzzy logic as the way reasoning really works and binary or Boolean logic as simply a special case of it.

7.1.2 Motivation

A good strategy in data mining and classification tasks is to use the simplest description of the data that does not compromise accuracy: extract crisp logical rules first, use fuzzy rules if crisp rules are not sufficient, and only if the number of logical rules required for high accuracy of classification is too large use other, more sophisticated tools. In many applications simple crisp logical rules proved to be more accurate and were able to generalize better than many machine and neural learning algorithms [52]. In other applications fuzzification of logical rules gave more accurate results [53]. Crisp logical rules may be converted to a specific form of fuzzy rules (Sect. VIII) and optimized using gradient procedures, providing higher accuracy without significant increase of the complexity or decrease of comprehensibility of the rule-based system.

Fuzzy logic allows for a set membership value to range (inclusively) between 0 and 1, and anything in between representing linguistic and imprecise term like “slightly”, “quite”, and “very” [77]. Specifically, it allows partial membership in a set. It is related to fuzzy sets and possibility theory. Fuzzy logic is a form of multi-valued logic derived from fuzzy set theory to deal with reasoning that is both robust and approximate rather than brittle and exact. In contrast with “crisp logic”, where binary sets are either true or false, fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Furthermore, when linguistic variables are used, the degrees may be managed by specific functions.

7.2 Our Approach

Many decision-making tasks of investors are too complex to be understood quantitatively. However, humans succeed by using knowledge that is imprecise rather than precise. Fuzzy logic refers to a logic system which represents knowledge and reasons in an imprecise or fuzzy manner for reasoning under uncertainty.

The basic fuzzy reasoning process in our IDA is summarized below:

- FUZZIFICATION: convert numeric data to literate words using fuzzy membership functions, and determine the degree of truth for the word. It calculates the degree to which the input data match the condition of the fuzzy rules.
- INFERENCE: the truth value of the condition of each rule is computed using AND, NOT, or OR, and applied to the conclusion part of each rule. The result is one fuzzy subset to be assigned to the output variable for each rule.

The output of each rule is scaled by the rule condition's computed degree of truth.

- COMPOSITION: all of the fuzzy subsets assigned to the output variable are combined together to form a single fuzzy. The operation SUM takes the point wise sum over all of the fuzzy subsets.

- DEFUZZIFICATION: convert the fuzzy output set to a numeric value. IDA uses the MAXIMUM method. It selects the maximum value of the fuzzy sets as the crisp value for the output variable.

For example, the question "*Director must accumulate meaningful amount of stock*". Let us assume we need to convert a specific stock amount value into linguistic terms such as low, average, and high. We define the following fuzzy member functions: $Q_{low}(x)$, $Q_{avg}(x)$, and $Q_{high}(x)$ where Q is the quantity of directors' stock.

With a tool like fuzzy reasoning, IDA is able to handle any specific number and to convert it into vague and imprecise linguistic term. It is a very critical step since many rules in the knowledge base and cases in the case base are often expressed in vague and imprecise language. In the Figure 12 below, low, average, and high represent a percentage scale of common stock shares owned by all executive officers and directors as a group over total outstanding common stock in a company.

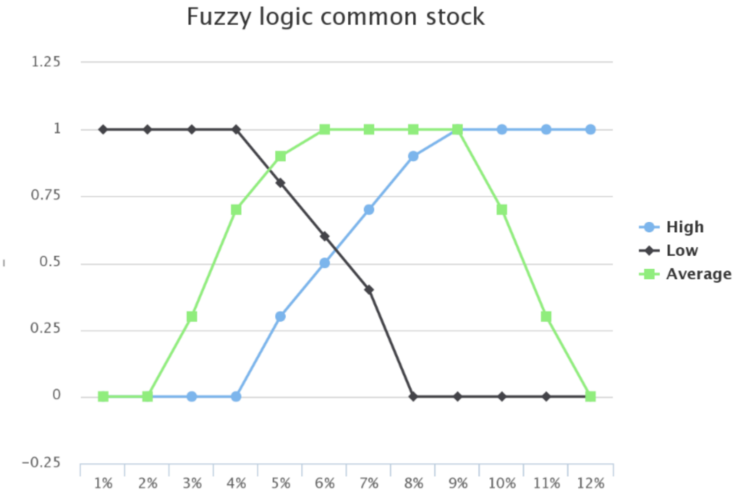


Figure 12: Fuzzy Logic Common Stock

7.2.1 Case Study 1

We test the question “*Director must accumulate meaningful amount of stock.*” using fuzzy reasoning. We retrieve relevant information from companies’ SEC documents, such as stock holding tables (Index A) for the following companies:

- Facebook: At the close of business on March 24, 2014, we had 1,991,090,546 shares of Class A common stock outstanding and 572,806,544 shares of Class B common stock outstanding and entitled to vote. Holders of our Class A common stock are entitled to one vote for each share held as of the above record date. Holders of our Class B common stock are entitled to ten votes for each share held as of the above record date. Based on the stock table for Facebook, in fiscal year 2014, all executive officers and directors as a group of Facebook owns 18,481,389 shares of class A common stock, and 484,943,817 shares of class B common stock, which is about 9.11% of the total class A common stock, about 84.66% of class B common stock.

25.99% for both. When converted to meaningful word, it is **very high** compared to the average based on the experiments.

- 3M CO: Each share of our common stock outstanding as of the close of business on March 13, 2015, the record date, is entitled to one vote at the Annual Meeting on each matter properly brought before the meeting. As of that date, there were 635,316,881 shares of common stock issued and outstanding. The stock table for 3M includes all 3M stock-based holdings, as of February 28, 2015, of the directors and the Named Executive Officers set forth in the Summary Compensation Table, and the directors and executive officers as a group. For 3M CO, all directors and executive officers as a group owns 3,139,993 shares, which is **0.49%** of total outstanding in fiscal year 2014. It is **low**.

- ALCOA: In fiscal year 2014, all directors and executive officers as a group of Alcoa owns 5,776,639 shares of common stock, **0.47%** of total shares. It is **low.**

- American Express: All current directors, nominees and executive officers owns **0.96%** of total outstanding common stock in fiscal year 2014. It is about the **average.**

It is not possible to give a true false answer in the case. No crisp logic can be applied here. Fuzzy reasoning is represented in the process. The following Figure 13 shows four companies' director stock percentage, and Table 6 represents the numeric data in literal words.

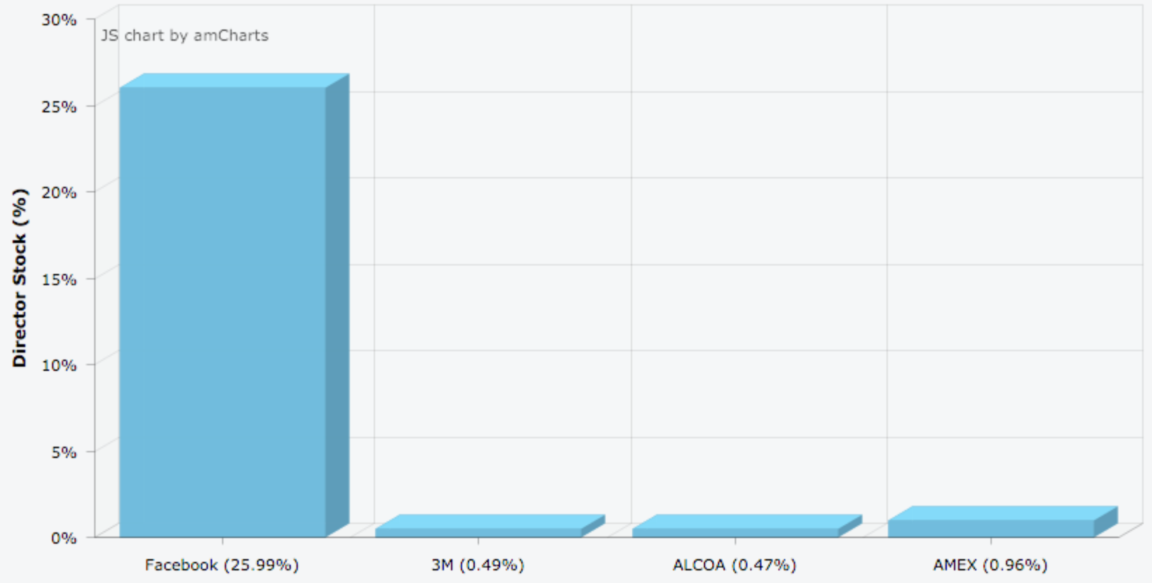


Figure 13: Director Stock Comparison Chart

Company	**Numeric Data**	**Literal Word**
Facebook	25.99%	Very High
3M	0.49%	Low
ALCOA	0.47%	Low
AMEX	0.96%	Average

Table 6: Director Stock Comparison

7.2.2 Case Study 2

We also studied another question: *"Meaningful limits on director compensation."* What is a "meaningful limit" on director compensation? According to the Internal Revenue Code, under section 162(m) [78], many equity plans will include per-person grant limits to enable compensation to be treated as performance-based compensation, which is reasonable, and has a meaningful limit. Companies that set such limits should ensure that they are set low enough to be "meaningful", yet high enough so that they will not be exceeded inadvertently. As the company grows each year, director compensation is likely to increase. The company should consider such growth in setting limits of director compensation and submitting to shareholders. The limits should be revisited on an annual basis to ensure it remains meaningful and will not be exceeded. Also, the limit should be in the procedures at the time director compensation is set to ensure that it will not exceed the limit approved by shareholders.

Director compensation that is most likely to be viewed as unreasonably high (e.g., director compensation that is above the 90th percentile, and significantly more than the median director compensation, of its peer group), without a compelling rationale for such pay. Companies should evaluate director compensation, especially in light of that paid to peers' directors. Companies should also explain in the annual proxy statement any special circumstances or attributes of the board that would justify special pay.

How to decide a limit? There are different ways:

1. a fixed dollar amount limit

2. a percentile limit relative to peer companies or an index/industry median

3. a limit relative to historic director compensation

A meaningful limit that applies to both equity and cash compensation could, for example, provide that the maximum number of shares of common stock subject to awards that are granted during a single fiscal year to any nonemployee director, taken together with any cash fees paid to such nonemployee director during the fiscal year, shall not exceed a certain percentage in total value. The total value should be based on an objective formula, such as the grant date fair value of such awards for financial reporting purposes.

7.2.2.1 Facebook

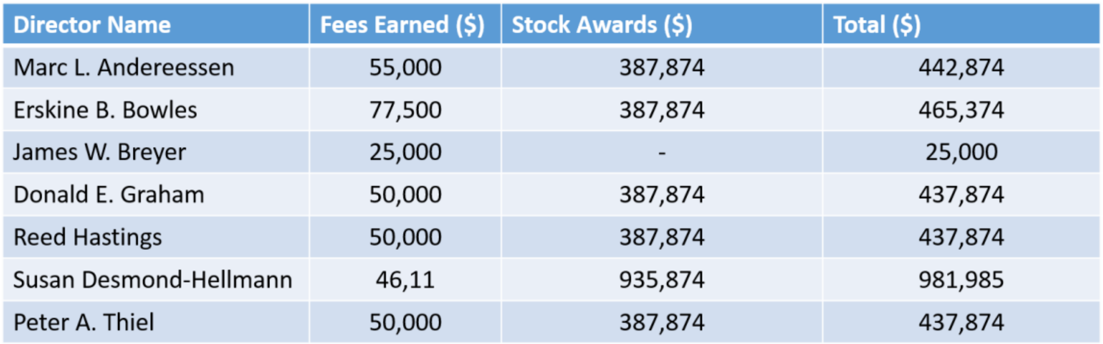
In June 2014, a shareholder plaintiff filed a derivative action against the directors and several officers of Facebook. Complaint, *Espinoza v. Zuckerberg*, C.A. No. 9745 (Del. Ch. filed June 6, 2014). In that case, the plaintiff alleges that the Facebook board of directors unjustly enriched themselves, breached their fiduciary duties, and wasted the company's assets by awarding themselves excessive compensation under Facebook's 2012 Equity Incentive Plan. In 2013, Facebook's nonemployee directors received, under the plan, an average of $461,000 in equity compensation, which the plaintiff claims were excessive relative to the company's peers. The plaintiff also contends that the plan's individual limit (of 2.5 million shares) is not meaningful because, at Facebook's then-current stock price, the board could theoretically grant each director up to approximately $145 million worth of equity.

From the search result, we found the following facts:

Non-Employee Director Compensation Arrangements

Each non-employee member of our board of directors receives an annual retainer fee of $50,000. Prior to October 1, 2013, the chair of our audit committee received an additional annual retainer fee of $20,000. Effective October 1, 2013, members of our audit committee (other than the chair) receive an annual retainer fee of $20,000, and the chair of our audit committee receives an annual retainer fee of $50,000.

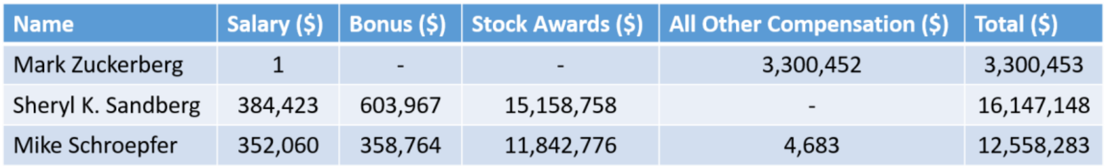
On September 13, 2013, our board of directors approved an annual grant of 7,742 RSUs to each non-employee director, which is equal to $300,000 divided by the average daily closing price of Facebook Class A common stock in August 2013. These RSUs vest on May 15, 2014 so long as the non-employee director is a member of our board of directors on such date. The following Table 7 represents each director' total



Director Name	Fees Earned ($)	Stock Awards ($)	Total ($)
Marc L. Andereessen	55,000	387,874	442,874
Erskine B. Bowles	77,500	387,874	465,374
James W. Breyer	25,000	-	25,000
Donald E. Graham	50,000	387,874	437,874
Reed Hastings	50,000	387,874	437,874
Susan Desmond-Hellmann	46,11	935,874	981,985
Peter A. Thiel	50,000	387,874	437,874

Table 7: Facebook 2013 Directors Compensation

The following Table 8 presents, for each of the named executive officers, information regarding outstanding stock options and RSUs held as of December 31, 2013.



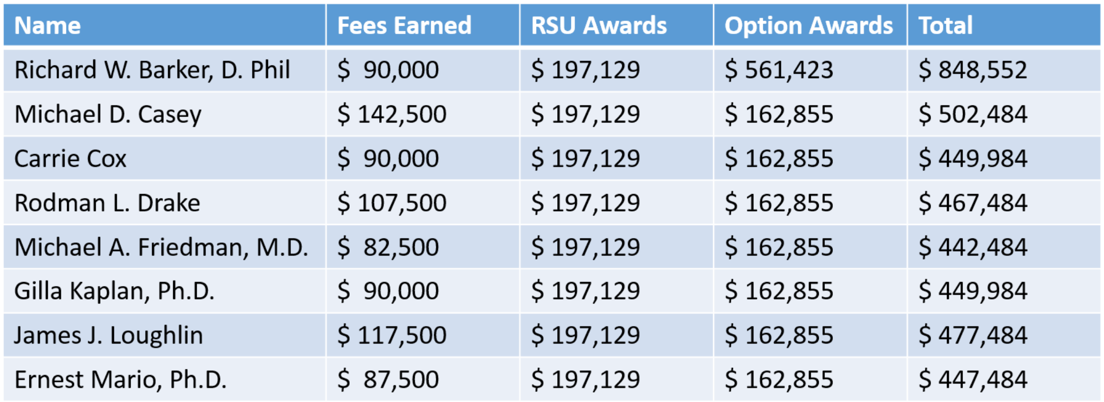
Name	Salary ($)	Bonus ($)	Stock Awards ($)	All Other Compensation ($)	Total ($)
Mark Zuckerberg	1	-	-	3,300,452	3,300,453
Sheryl K. Sandberg	384,423	603,967	15,158,758	-	16,147,148
Mike Schroepfer	352,060	358,764	11,842,776	4,683	12,558,283

Table 8: Facebook 2013 Executive Officers Compensation

7.2.2.2 Celgene Corporation

In October 2014, a derivative action was filed against the directors and officers of Celgene Corporation. The complaint asserts claim for breach of fiduciary duty, waste, and unjust enrichment stemming from allegedly excessive director compensation in 2012 and 2013. In 2012, the Celgene nonemployee directors received, under the company's stock plan, average total compensation of $502,484. In 2013, that average was $833,119. The plaintiff claims these awards were excessive relative to the company's peers, who averaged approximately $320,000 and $350,000 in per-director compensation in 2012 and 2013. The plaintiff also attacks the stock plan's 1.5-million-share individual limit as "illusory" and not "meaningful" because, at the company's then-current stock price, the board could theoretically grant each director up to about $145 million worth of equity.

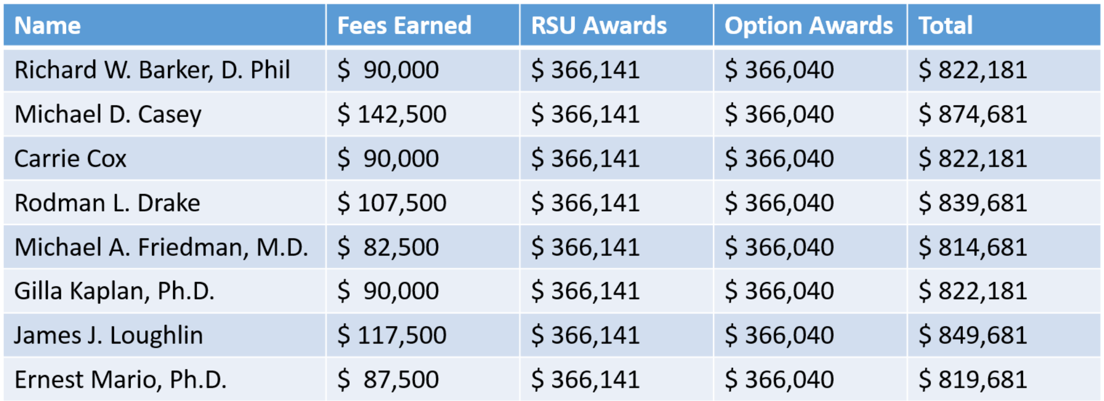
As described more fully below, the following Table 9 summarizes the annual cash compensation for the Non-Employee Directors serving as members of our Board of Directors during fiscal 2012.



Name	Fees Earned	RSU Awards	Option Awards	Total
Richard W. Barker, D. Phil	$ 90,000	$ 197,129	$ 561,423	$ 848,552
Michael D. Casey	$ 142,500	$ 197,129	$ 162,855	$ 502,484
Carrie Cox	$ 90,000	$ 197,129	$ 162,855	$ 449,984
Rodman L. Drake	$ 107,500	$ 197,129	$ 162,855	$ 467,484
Michael A. Friedman, M.D.	$ 82,500	$ 197,129	$ 162,855	$ 442,484
Gilla Kaplan, Ph.D.	$ 90,000	$ 197,129	$ 162,855	$ 449,984
James J. Loughlin	$ 117,500	$ 197,129	$ 162,855	$ 477,484
Ernest Mario, Ph.D.	$ 87,500	$ 197,129	$ 162,855	$ 447,484

Table 9: Celgene 2012 Directors Compensation

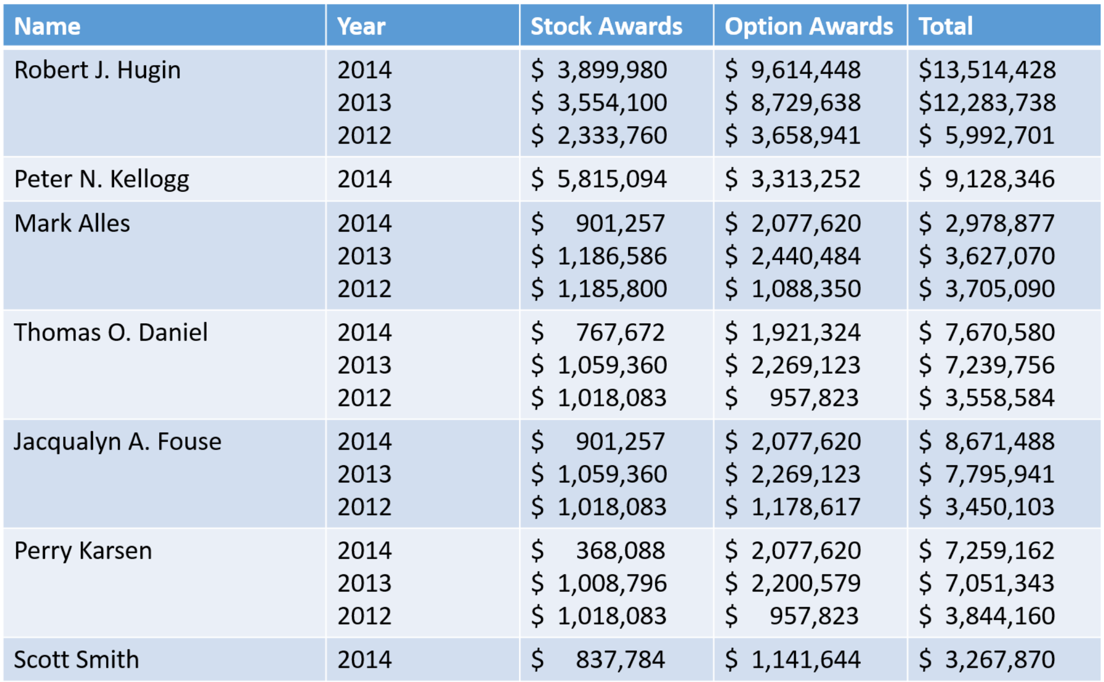
As described more fully below, the following Table 10 summarizes the annual compensation for the Non-Employee Directors serving as members of our Board of Directors during fiscal 2013.



Name	Fees Earned	RSU Awards	Option Awards	Total
Richard W. Barker, D. Phil	$ 90,000	$ 366,141	$ 366,040	$ 822,181
Michael D. Casey	$ 142,500	$ 366,141	$ 366,040	$ 874,681
Carrie Cox	$ 90,000	$ 366,141	$ 366,040	$ 822,181
Rodman L. Drake	$ 107,500	$ 366,141	$ 366,040	$ 839,681
Michael A. Friedman, M.D.	$ 82,500	$ 366,141	$ 366,040	$ 814,681
Gilla Kaplan, Ph.D.	$ 90,000	$ 366,141	$ 366,040	$ 822,181
James J. Loughlin	$ 117,500	$ 366,141	$ 366,040	$ 849,681
Ernest Mario, Ph.D.	$ 87,500	$ 366,141	$ 366,040	$ 819,681

Table 10: Celgene 2013 Directors Compensation

The following Table 11 sets forth information regarding compensation earned by our NEOs for the fiscal years ended December 31, 2014, 2013, and 2012.



Name	Year	Stock Awards	Option Awards	Total
Robert J. Hugin	2014 2013 2012	$ 3,899,980 $ 3,554,100 $ 2,333,760	$ 9,614,448 $ 8,729,638 $ 3,658,941	$13,514,428 $12,283,738 $ 5,992,701
Peter N. Kellogg	2014	$ 5,815,094	$ 3,313,252	$ 9,128,346
Mark Alles	2014 2013 2012	$ 901,257 $ 1,186,586 $ 1,185,800	$ 2,077,620 $ 2,440,484 $ 1,088,350	$ 2,978,877 $ 3,627,070 $ 3,705,090
Thomas O. Daniel	2014 2013 2012	$ 767,672 $ 1,059,360 $ 1,018,083	$ 1,921,324 $ 2,269,123 $ 957,823	$ 7,670,580 $ 7,239,756 $ 3,558,584
Jacqualyn A. Fouse	2014 2013 2012	$ 901,257 $ 1,059,360 $ 1,018,083	$ 2,077,620 $ 2,269,123 $ 1,178,617	$ 8,671,488 $ 7,795,941 $ 3,450,103
Perry Karsen	2014 2013 2012	$ 368,088 $ 1,008,796 $ 1,018,083	$ 2,077,620 $ 2,200,579 $ 957,823	$ 7,259,162 $ 7,051,343 $ 3,844,160
Scott Smith	2014	$ 837,784	$ 1,141,644	$ 3,267,870

Table 11: Celgene 2012 – 2014 NEOs Compensation

7.2.2.3 Intel

NON-EMPLOYEE DIRECTOR AWARDS

Each year, each non-employee director may be granted awards for a number of shares established by the Board, but the number of shares subject to such awards may not exceed 100,000 shares each fiscal year. This limit is subject to adjustment to reflect stock splits and similar changes in Intel’s capitalization. Subject to limits in the plan terms, the Board has the discretion to determine the form and terms of awards to non-employee directors. Our current practice is to grant each non-employee director a mix of RSUs and OSUs each January with target value of approximately $220,000.

7.3 Summary

In this chapter, we equipped IDA System with fuzzy reasoning mechanism that employs fuzzy logic and reasoning to measure partial truth values of matched rules and data. With fuzzy inference, it allows for set membership values, and conversion of numeric data to literate words. It is a very critical step since many rules in the knowledge base and cases in the case base are often expressed in vague and imprecise language.

Chapter 8

Machine Learning

8.1 Overview

In the case that no matching answer can be found after all the above steps, the IDA System will then search the semantic net with the relevant word(s) to find the matching synonyms. Once the match is found, the IDA System will go through the above steps in 3.2 – 3.3 to seek the answer to the question. These steps will repeat until the matching answer is found. Once the answer is found, the relevant passages we used from the semantic net will be saved to the local knowledge base for the question, so it can be reused as a relevant passage in the future search for this particular question.

In such cases, we use Case-based reasoning (CBR) - the machine learning algorithm used to seek answers for new companies based on the solution for similar past companies for the same question. The roots of CBR was traced to the work of by Riesbeck and Schank [54]:

"A case-based reasoned solves problems by using or adapting solutions to old problems."

The CBR-cycle comprises 4 steps [79] [80]:

- Retrieve: Given a target question, pick a company, retrieve from memory cases relevant to seek the answers. A case consists of keywords of questions, from the financial document of a company, and how the answer to the questions were derived.

- Reuse: Map the solution from the previous case to the target problem. This may involve adapting the solution as needed to fit the new situation.
- Revise: Having mapped the previous solution to the target situation, test the new solution and, if necessary, revise. For example, new, and similar words need to be added to the semantic net.
- Retain: After the solution has been successfully adapted to the target problem, store the resulting experience as a new case in memory.

There is a set of principles which guide the action of the CBR-cycle:

- retrieve relevant passages from synonyms database to use it in seeking an answer to the current question from a company's document
- attempt to reuse the relevant passage retrieved from a previous case

The following section will show how the application uses this set of principles to solve the problem.

8.1.1 Similarity

We select the following question for this experiment: "*The board provides golden parachutes or employment agreements that protect executive officers*". The keywords highlighted in the question are: "employment agreements", "golden parachutes", and "executive officers". After multiple searches were done for different companies, we found out when the answer has "employment agreements", it might or might not have “golden parachutes”. In general, they have no relation to another; only in the financial industry, they can be treated as similar factors for seeking an answer to the question.

We also found that there are multiple similar attributes in most of the answers whether or not it contains both “employment agreements” and “golden parachutes”, or it just contains one; attributes such as, “executive officers”, “change in control benefits”, “termination”, and “severance arrangements” etc., or with their synonyms. Answers examples like:

- For Intel, “Intel does not provide employment agreements, severance arrangements, or change in control benefits to executive officers”
- For Chevron, “No golden parachutes or golden coffins for NEOs (Chevron’s named executive officers)”.
- For Caterpillar, “Caterpillar does not have any pre-existing severance agreements or packages (such as golden parachutes) under which payments are to be made to any NEO upon a termination of employment or change in control”

Based on the above experiment, the IDA System develops a process to determine the similarities of all the attributes in the question and in the target passage retrieved from companies’ financial documents. This measure may be multiplied by some weight factors. Then the sum of the similarity of all attributes is calculated to provide a measure of the similarity of the key word(s) from the question to the relevant passages in the synonyms database. This can be presented by the following equation:

$$\text{Similarity}(S, T) = \sum_{i=1}^{n} f(S_i) \text{ x } w1_i \,/\, \sum_{i=1}^{n} f(T_i) \text{ x } w2_i$$

Where *S* is the source case; *T* is target case; *n* the number of attributes in each case; *i* an individual attribute (key word) from 1 to *n*; *f* a similarity function for attribute *i* in cases *S* and *T*; and *w1* is the similarity weighting of attribute *i* in the source case, *w2* is the similarity weighting of attribute *i* in the target case.

Similarities fall within a range of zero to one, one is an exact match, and zero is completely dissimilar.

Once a target passage has been accepted as the matching answer to the question, the attributes in the target passage are entered into the IDA System's local knowledge base for future use, thus, completing the CBR-cycle.

8.2 Analogical Learning and Reasoning

Analogical learning is concerned with the improved performance of a system over time without much human intervention. Over the past several years, analogical reasoning and learning have grown to become a central research subject in artificial intelligence. Researchers believe that analogical problem solving (case-based reasoning) provides a promising approach for the acquisition and effective use of knowledge. It is a heuristic approach in which reasoning can be guided, solutions to unfamiliar problems can be constructed, and abstract knowledge can be generalized. In the domain of text extraction and analysis, there is no way of predicting every possible word/term used in the required SEC filings by thousands of corporations.

More often than not, the IDA may encounter a word that was not provided initially in the knowledge bases. Traditional text and analysis systems are brittle, requiring human intervention to handle data which has not been forethought, or which is at the edge of the system domain. They work appropriately only in the narrow areas of knowledge provided to knowledge engineers in advance, and require substantial human assistance to compensate for even slight variations in representations or descriptions. In response to these difficulties, IDA employs analogical learning and is able to benefit from the experience and knowledge accumulated in its previous cycles of text processing and analysis. In what follows, we discuss two important ways of evaluating similarity: semantic and structural similarity evaluation.

8.3 Semantic Network Search and Match

We developed a semantic network for each question under consideration in order to handle question-dependent text retrieval and analysis. In order to describe the process of analogical learning and retrieval in a concise way, we define the following terms:

Case Base {C} : = $<C_1><C_2>.. <C_i>...<C_n>$

Case C_i : = $<D_i><W_i>$

Word W_i : = $<w_{i1}><w_{i2}>..<w_{ij}> .. <w_{im}>$

Description D_i : = $<d_{i1}><d_{i2}>..<d_{ij}>... <d_{is}>$

Where w_{ij} and d_{ij} represent the j^{th} word and the j^{th} descriptor of a case C_i respectively.

Let C_{new} be a new case. The basic algorithm used in IDA can then be described as follows:

- Interpret D_{new} of C_{new} and classify it in terms of the descriptors D_j, where $D_j \in C_j$, $C_j \in \{C\}$.
- Compute the similarity score, $S(D_{new},D_j)$, find a case with the highest $S(D_{new}, D_h)$.
- Select and modify a word $w_h \in D_h$ to w_{new} with respect to D_{new}.
- Determine the suitability of w_{new}.
- If satisfied, construct a complete case C_{new} which consists of D_{new} provided by the user and w_{new} constructed by the system.
- Integrate C_{new} into $\{C\}$ and insert w_{new} into the key word list.
- If not satisfied, choose an alternative case. Try again.

The overall organization of index hierarchies in IDA ensures that conceptually similar words can be located quickly and made available. In addition to the conceptual evaluation process, the similarity evaluation function also employs the feature matching process in the cases of conceptual identity and conceptual irrelevancy. The similarity match scores, in the computation of feature matching, are determined by a combination of the number of features in common and the relative importance, expressed by weights, of those features. More formally, the feature matching calculation can be described as follows.

Let N be a new case with m features,

$$N = \{n_1\ n_2\ ...\ n_m\}$$

O be an old case with s features,

$$O = \{o_1\ o_2 \ldots\ o_s\}$$

CF denote a common feature set

$$CF = \{c_1\ c_2\ ..\ c_k\} \text{ where } k \leq \min(m,s)$$

$$CF = \{ C \in CF \mid c \in N \wedge c \in O \}$$

Thus, a similarity score of a new case N is given as:

$$S(N,O) = \quad 1 \leq i \leq k$$

Where λ_i is a weight assigned to the ith feature of CF.

The basic idea behind feature matching is the prototypically which assumes the cases in the same class would approximately share the same features. A survey of the cognitive literature suggests that people never use all data in memory when faced with a problem. Rather, people recognize, extract, and compile from the data a limited set of features and use it to recall relevant experience. The feature matching process is a computational implementation of this aspect of human analogical problem solving. In the case of conceptual identity, the feature matching method can further discriminate and rank the conceptually identical cases at more specific levels.

As an illustration, consider the following words*, “executive officers”*, “*board of directors”*, “*CEO”*, “*CFO”, “management”* and “*Chairman of the board”*. These terms would not be classified as synonyms by a traditional semantic network or a dictionary. With the similarities shared in a financial domain in general and SEC filings in particular, we group them together to form a local semantic network. Once a new synonym is found, IDA inserts it into appropriate a key word list and therefore, automatically amends its ability to analyze and process SEC filings in the future. Thus, future searches would not be necessary and processing time is reduced. Over time, IDA will demonstrate its improved IQ and learn from its experiences.

8.4 Conceptual Similarity Evaluation and Match

Conceptual similarity in this domain refers to words/phrases with similar conceptual meanings, but may not be considered relevant when looked at separately. The conceptual similarity can be discovered by analyzing the structure and placement of words. Two words may look quite different in a traditional way; they may share conceptual similarity in a particular context. To implement this approach, IDA parses and stores the major components of passages into a template consisting of several slots. IDA then conducts a similarity match for the words in corresponding slots. If the majority of the slots turn out to be similar except one, we may conclude the unrecognized word is a synonyms wit relation to its corresponding counterpart.

To illustrate this approach, consider the following paragraphs extracted from several companies' fillings:

- *"Intel does not provide employment agreements, severance arrangements, or change in control benefits to executive officers"*
- *"Caterpillar does not have any pre-existing severance agreements or packages (such as golden parachutes) under which payments are to be made to any management staff upon a termination of employment or change in control"*

- “*Chevron does not provide golden parachutes or golden coffins for CEO upon a termination”.*

By way of slot comparisons and analysis, the term golden parachutes is similar to severance arrangement. Once identified and recognized, the new word “golden parachutes” is inserted into the key word list associated with its semantic network designed for this particular question. Conceptual similarity can also be identified by the placement of passages and words in similar documents or sections. Once a target passage has been accepted as the matching answer to the question, the attributes in the target passage are entered into the IDA System’s local knowledge base for future use.

8.5 Summary

In summary, case-based reasoning can adapt old solutions to meet new demands, using old cases to explain new situations and using old cases to critique new solutions which will improve the performance of the IDA System.

Chapter 9

Implementation and Evaluation

9.1 Implementation

To evaluate the performance of our proposed system, we developed a local knowledge base and searching engine testbed with one DELL PowerEdge T420 server that was configured with 2X Intel Xeon E5-2400 processors with 128GB ECC RAM and a 1.5TB hard drive. We installed Ubuntu server 14 on server to achieve max performance.

The local knowledge base is implemented on MySQL database community server v5.0, it is an open-source relational database management system with high performance that can handle the massive database operation we need.

When we perform the search, if keywords cannot be found in a searched document, the system will try to get synonyms from a semantic net which is discussed in Chapter 6, if no synonyms are found in local the knowledge base as well, a request will be sent to a public semantic net try to get synonyms from outside of the local environment. The search engine then uses the synonyms from the external database to perform the search, and also saves the new synonyms to the semantic net. Once the result is retrieved, the same result will be saved to the local knowledge base for future use.

The local knowledge base will keep growing each time we perform a search, and each time new rules are found. The system also automatically copies the synonyms from the external database to the local semantic net to improve the system performance, and the system eventually can work offline to a certain degree once all synonyms have been copied to the local semantic net, we can then run the system offline without connecting to internet for semantic net synchronization.

The IDA System was developed using JavaFx based on Java 7.0 to make sure it could run OS independently. The system was also written as a Java API, JavaFX application code which can reference APIs from any Java library.

JavaFX 2.2 and later releases are fully integrated with the Java SE 7 Runtime Environment (JRE) and the Java Development Kit (JDK). Because the JDK is available for all major desktop platforms (Windows, Mac OS X, and Linux), JavaFX applications compiled to JDK 7 and later also run on all the major desktop platforms. Shown in Figure 14 [81].

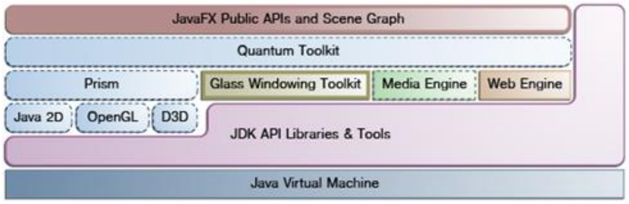


Figure 14: JavaFX Architecture

The JavaFx application for the IDA System is a user friendly interface, it is very easy for the end user to retrieve documents, analyze information and display reporting. It contains the following modules:

- Document import: Documents can be imported form a website directly or imported from a file system.

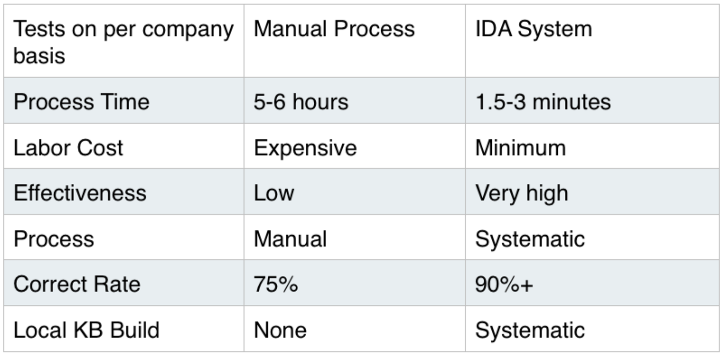
- Search and Analysis module: Built Lucent search engine allows easy user indexing, searching, and ranks answers to any pre-defined questions.

- Machine learning: Case-based reasoning is used here for machine learning mechanism, if the analysis module didn't find any answers for the question, new synonyms will be retrieved from the global knowledge database (campus built) based on CBR methodology.

- Reporting: All analysis data will be displayed to the user via the reporting module for easy understanding. Results will be saved to the database for the purpose of further accuracy analysis.

9.2 Performance Evaluation

In this section, we present the evaluation setup of the local knowledge base. We also perform and display the evaluation results of all three methods, and a comparison evaluation result between is also performed. Compared to the manual process approach, the IDA System reduces the time needed from 5-6 hours to a few minutes per company for all the questions. Table 12 shows the evaluated metrics.



Tests on per company basis	Manual Process	IDA System
Process Time	5-6 hours	1.5-3 minutes
Labor Cost	Expensive	Minimum
Effectiveness	Low	Very high
Process	Manual	Systematic
Correct Rate	75%	90%+
Local KB Build	None	Systematic

Table 12: Evaluation Metrics

The performance of the IDA System is striking, given its simplicity and the time needed to analyze a company's proxy statement. By employing the knowledge-based approach, the IDA System offers a solution to the problem of text explosion by replacing or supplementing the human user with automated answer extraction processes. Instead of conducting a large-scale, brute-force search in unstructured text, the IDA System applies a controlled, knowledge-guided text mining approach that avoids the overabundance problem. Thus, document retrieval is guided, question decomposition and analysis is a priori fixed and answer extraction can be performed automatically, or with a little assistance from humans. By using domain-restricted and question-specific approaches, problems of syntactic processing (parsing), semantic analysis, and contextual analysis commonly seen in natural language processing can be managed effectively.

Chapter 10

Concluding Remarks

In summary, we proposed, developed and tested an Intelligent Document Analysis System to help individual and institutional investors to evaluate the corporate governance practices of individual companies. The work reported in this paper has discussed several important issues in intelligent text mining, machine learning, fuzzy reasoning and information retrieval, has proposed and implemented solutions to these questions, and has demonstrated the usefulness and feasibility of these solutions through the design of the IDA System.

Reference

[1] U.S. Securities and Exchange Commission, EDGAR, http://www.sec.gov/edgar/aboutedgar.htm, 2010.

[2] U. S. Securities, E. Commission et al., "The investor's advocate: How the sec protects investors, maintains market integrity, and facilitates capital formation," Retrieved February, vol. 7, p. 2008, 2008.

[3] U.S. Securities and Exchange Commission, "2014 Report and Certification of Internal Supervisory Controls", 2014. http://www.sec.gov/about/sec-report-certification-internal-supervisory-controls-2014.pdf

[4] Dr. Carolyn Kay Brancato, and Christian A. Plath, "Corporate Governance Handbook 2005: Developments in Best Practices, and Legal Standards", the Conference Board, 2005. https://www.conference-board.org/publications/publicationdetail.cfm?publicationid=983

[5] Madeleine Bates, Ralph M. Weischedel, "Challenges in Natural Language Processing", Cambridge University Press, November 23, 2006.

[6] N. Tomuro, "Interrogative Reformulation Patterns and Acquisition of Question Paraphrases, Proceeding of the Second International Workshop on Paraphrasing, July (2003), pp. 33-40.

[7] S. Wanpeng, F. Min and G. Naijie, "Question Similarity Calculation for FAQ Answering", Third International Conference on Semantics, Knowledge and Grid, (2007), pp. 298-231.

[8] D. Huizhong, C. Yunbo and L. Chinyew, "Searching Questions by Identifying Question Topic and Question Focus", Proceeding of ACL, (2008), pp. 156-164.

[9] L. Yandong, B. Jiang and E. Agichtein, "Predicting Information Seeker Satisfaction in Community Question Answering", Proceeding of SIGIR, New York, USA, (2008), pp. 483-490.

[10] A. Agarwal, H. Raghavan and K. Subbian, "Learning to Rank for Robust Question Answering", Proceedings of the 21st ACM international conference on Information and knowledge management, New York, USA, (2012), pp. 833–842.

[11] B. Li, I. King and M. R. Lyu, "Question Routing in Community Question Answering: Putting Category in Its Place", Proceedings of the 20th ACM Conference on Information and Knowledge Management, Glasgow, Scotland, October (2011), pp. 2041-2044.

[12] Nico Schlaefer, Jennifer Chu-Carroll, Eric Nyberg, James Fan, Wlodek Zadrozny, and David Ferrucci, "Statistical Source Expansion for Question Answering", in proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011.

[13] John R. Frank, Max Kleiman-Weiner, Daniel A. Roberts, Feng Niu, Ce Zhang, and Christopher Re, "Building an Entity-Centric Stream Filtering Test Collection for TREC 2012", in proceedings of the Text Retrieval Conference (TREC), 2012.

[14] Jonathan L. Elsas, Jaime Arguello, Jamie Callan, and Jaime G. Carbonell, "Retrieval and Feedback Models for Blog Feed Search", in proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[15] Yang Xu, Gareth J.F. Jones, and Bin Wang, "Query Dependent Pseudo-Relevance Feedback Based on Wikipedia", in proceedings of the 32th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009.

[16] Evgeniy Gabrilovich and Shaul Markovitch, "Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis", in proceedings of the 20th International Joint Conference on Artifical Intelligence, 2007.

[17] Ofer Egozi, Evgeniy Gabrilovich, and Shaul Markovitch, "Concept-Based Feature Generation and Selection for Information Retrieval", in proceedings of the 23rd National Conference on Artificial Intelligence, 2008.

[18] Michael L. Wick, and Andrew McCallum, "Query-Aware MCMC", in Advances in Neural Information Processing Systems, 2011.

[19] Jeffrey Dalton, and Laura Dietz, “Constructing Query-Specific Knowledge Bases”, in proceedings of the ACM International Conference on Automated Knowledge Base Construction Workshop, 2013.

[20] Ricardo Baeza-yates, Gonzalo Navarro, “Integrating Contents and Structure in Text Retrieval”, SIGMOD Record, Volume 25 Issue 1, March 1996.

[21] Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto, “Modern information retrieval”, Vol. 463. New York: ACM press, 1999.

[22] Deerwester, Scott, et al. "Indexing by latent semantic analysis", Journal of the American society for information science 41.6 (1990): 391.

[23] D. Saravanan, “Text Information Retrieval Using Data Mining Clustering Technique,” in proceeding of International Journal of Applied Engineering Research, 2015.

[24] J. Mannar Mannan, and Dr. M Sundarambal, “Ontology Integration for Query Expansion and Semantic Filtering Using Word net for Information Retrieval,” in proceeding of International Journal of Applied Engineering Research, 2015.

[25] M. Thangaraj, and G. Sujatha, “An Architectural Design for Effective Information Retrieval in Semantic Web,” in proceeding of International Journal of Expert Systems with Applications, 2014.

[26] Anna Stavrianou, Periklis Andritsos, and Nicolas Nicoloyannis, “Overview and Semantic Issues of Text Mining,” in proceeding of ACM SIGMOD Record 36(3), 23-24, 2007.

[27] Jimin Liu, and Tat-Seng Chua, “Building Semantic Perceptron Net for Topic Spotting,” in proceeding of the 39th Annual Meeting on Association for Computational Linguistics, 2001.

[28] Weigang Wang, and Roy Rada, “Structured Hypertext with Domain Semantics,” in proceeding of ACM Transactions on Information Systems, 1998.

[29] Shahrul Azman Noah, Lailatulqadri Zakaria, and Arifah Che Alhadi, “Extracting and Modeling the Semantic Information Content of Web Documents to Support Semantic Document Retrieval,” in proceeding of the Sixth Asia-Pacific Conference on Conceptual Modeling, Volume 96, 2009.

[30] Sasa Nesic, Fabio Crestani, Mehdi Jazayeri, and Dragan Gasevic, “Concept-Based Semantic Annotation, Indexing and Retrieval of Office-Like Document Units,” in proceeding of RIAO ’10 Adaptivity, Personalization and Fusion of Heterogeneous Information, 2010.

[31] George Tsatsaronis, Iraklis Varlamis, and Kjetil Norvag, “SemaFor: Semantic Document Indexing Using Semantic Forests,” in proceeding of the 21st ACM International Conference on Information and Knowledge Management, 2012.

[32] Andreas Becks, Stefan Sklorz, and Matthias Jarke, “A Modular Approach for Exploring the Semantic Structure of Technical Document Collections,” in proceeding of the Working Conference on Advanced Visual Interfaces, 2000.

[33] Edy Portmann, Michael Alexander Kaufmann, and Cefric Graf, “A Distributed, Semiotic-Inductive, and Human-Oriented Approach to Web-Scale Knowledge Retrieval,” in proceeding of the International Workshop on Web-Scale Knowledge Representation, Retrieval and Reasoning, 2012.

[34] Aasia Khanum, Muid Mufti, M. Younus Javed, and M. Zubair Shafiq, “Fuzzy Case-Based Reasoning for Facial Expression Recognition,” in proceeding of Fuzzy Sets and Systems, An International Journal in Information Science and Engineering, Volume 160, Issue 2, 2009.

[35] German Florez, Susan M. Bridges, and Rayford B. Vaughn, “An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection,” in proceeding of Fuzzy Information Processing Society, Annual Meeting of the North American, 2002.

[36] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, and Qiang Yang, et al, “Top 10 Algorithm in Data Mining,” in proceeding of Knowledge and Information Systems, Volume 14, Issue 1, 2008.

[37] Hossam M. Ashtawy, and Nihar R. Mahapatra, "A Comparative Assessment of Ranking Accuracies of Conventional and Machine-Learning-Based Scoring Funcions for Protein-Ligand Binding Affinity Prediction," in proceeding of IEEE/ACM Transaction on Computational Biology and Bioinformatics (TCBB), 2012.

[38] Nan Li, and Desheng Dash Wu, "Using Text Mining and Sentiment Analysis for Online Forums Hotspot Detection and Forecast," Decision Support Systems, Volume 48, Issue 2, 2010. Available at: http://dx.doi.org/10.1016/j.dss.2009.09.003.

[39] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in proceeding of the 10th European Conference on Machine Learning Chemnitz, 1998.

[40] Trevor Mansuy, and Robert J. Hilderman, "A Characterization of WordNet Features in Boolean Models for Text Classification," in proceeding of the fifth Australasian Conference on Data Mining and Analystics, 2006.

[41] I. Watson, "Case-Based Reasoning Is a Methodology Not a Technology," in proceeding of Elsevier Science B.V., Volume 12, Issues 5-6, 1999.

[42] Feixiong Cheng, Jie Shen, Yue Yu, Weihua Li, Guixia Liu, Philip W. Lee, and Yun Tang, “In Silico Prediction of Tetrahymena Phriformis Toxicity for Diverse Industrial Chemicals with Substructure Pattern Recognition And Machine Learning Methods,” in proceeding of Elsevier B.V., Volume 82, Issue 11, 2011.

[43] Dr. Ghayda A. Al-Talib, and Hind S. Hassan, “A Study on Analysis of SMS Classification Using TF-IDF Weighting,” in proceeding of International Journal of Computer Networks and Communications Security, 2013.

[44] Wu, H.C., Luk, R. P., Wong, K. F., and Kwok, K. L., “Interpreting TF-IDF term weights as making relevance decision,” in proceeding of ACM Transactions on Information Systems, 2008.

[45] Ugo Erra, Sabrina Senatore, Fernando Minnella, and Giuseppe Caggianese, “Approximate TF-IDF based on topic extraction from massive message stream using the GPU,” in proceeding of Information Sciences Journal, 2015.

[46] Zhendong Shi, Jacky Keung, and Qinbao Song, “An Empirical Study of BM25 and BM25F Based Feature Location Techniques,” in proceeding of the International Workshop on Innovative Software Development Methodologies and Practices, 2014.

[47] Jeff A. Bilmes, and Katrin Kirchhoff, “Factored Language Models And Generalized Parallel Backoff,” in proceeding of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Volume 2, 2003.

[48] Gianni Amati, and Cornelis Joost Van Rijsbergen, “Probablistic Models of Information Retrieval Based on Measuring the Divergence from Randomness,” in proceeding of Transactions on Information Systems (TOIS), Volume 20, Issue 4, 2002.

[49] Stephane Clinchant, and Eric Gaussier, “Information-Based Models For Ad Hoc IR,” in proceeding of the 33rd International ACM SIGIR Conference on Research And Development In Information Retrieval, 2010.

[50] Hearst, M. Untangling Text Data Mining, and Proceedings of ACL’99: the 37th Annual Meeting of the Association of Computational Linguistics (1999)

[51] Marti Hearst, October 17, 2003, “What is Text Mining?” http://people.ischool.berkeley.edu/~hearst/text-mining.html

[52] Duch, Wlodzislaw, et al. "Hybrid Neural-global Minimization Method of Logical Rule Extraction." JACIII 3.5 (1999): 348-356.

[53] Duch, Włodzisław, et al. "Fuzzy and crisp logical rule extraction methods in application to medical data." Fuzzy systems in Medicine. Physica-Verlag HD, 2000. 593-616.

[54] C.K. Riesbeck, R. Schank, Inside Case-based Reasoning, Erlbaum, Northvale, NJ, 1989.

[55] The Java Tutorials, HttpURLConnetion. http://download.oracle.com/javase/tutorial/networking/urls/connecting.html

[56] Apache Lucene: A high-performance, full-featured Information Retrieval (IR) engine, http://lucene.apache.org

[57] Salton, G. and M.J. McGill (1983), “Introduction to modem information retrieval”, McGraw-Hill, ISBN 0070544840.

[58] Salton, Gerard and Buckley, C. (1988), “Term-weighting approaches in automatic text retrieval”, Information Processing & Management, doi: 10.1016/0306-4573(88) 90021-0.

[59] Cutting, Bialecki et al, 2006, “Narrative Structure Detection and Evaluation”.

[60] Xianfeng, Yang, and Liu Pengfei. "Question Recommendation and Answer Extraction in Question Answering Community." International Journal of Database Theory and Application 9.1 (2016): 35-44.

[61] Afzal, Muhammad, et al. "Knowledge-based query construction using the CDSS knowledge base for efficient evidence retrieval." Sensors 15.9 (2015): 21294-21314.

[62] Li, Sujian, et al. "Semantic computation in a Chinese question-answering system." Journal of Computer Science and Technology 17.6 (2002): 933-939.

[63] Neumann, Günter, Christian Braun, and Jakub Piskorski. "A divide-and-conquer strategy for shallow parsing of German free texts." Proceedings of the sixth conference on Applied natural language processing. Association for Computational Linguistics, 2000.

[64] Reinberger, Marie-Laure, and Walter Daelemans. "Is shallow parsing useful for unsupervised learning of semantic clusters?", International Conference on Intelligent Text Processing and Computational Linguistics. Springer Berlin Heidelberg, 2003.

[65] Yuanhua Lv, and ChengXiang Zhai, “Positional Language Models for Information Retrieval”, in proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009.

[66] David Carmel, Anna Shtok, and Oren Kurland, “Position-Based Contextualization for Passage Retrieval”, in proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 2013.

[67] Rada, Roy. "Hypertext writing and document reuse: the role of a semantic net." Electronic Publishing 3.3 (1990): 125-140.

[68] Stuart J. Russel, and Peter Norvig, “Artificial Intelligence: a Modern Approach”, Prentice Hall, Englewood Cliffs, 1995.

[69] John F. Sowa, “Knowledge Representation: Logical, Philosophical, and Computational Foundation”, Pacific Grove, CA, 2000.

[70] Alon Halevy, Anand Rajaraman, and Joann Ordille, “Data Integration: The Teenage Years”, in Proceedings of the 32^{nd} VLDB, Korea, pp. 9-16, 2006.

[71] L.A. Zadeh. “Fuzzy Sets”, University of California, Information and Control 8, 338-353, 1965

[72] Klaua D., Über einen Ansatz zur mehrwertigen Mengenlehre. Monatsb. Deutsch. Akad. Wiss. Berlin 7, 859–876

[73] Salii, V.N. (1965). "Binary L-relations". Izv. Vysh. Uchebn. Zaved. Matematika (in Russian) **44** (1): 133–145.

[74] Burgin, Mark. "Neoclassical analysis: fuzzy continuity and convergence." Fuzzy Sets and Systems 75.3 (1995): 291-299.

[75] Wang, Xuzhu, and Etienne E. Kerre. "Reasonable properties for the ordering of fuzzy quantities (I)." Fuzzy sets and systems 118.3 (2001): 375-385.

[76] Deschrijver, Glad, and Etienne E. Kerre. "On the relationship between some extensions of fuzzy set theory." Fuzzy sets and systems 133.2 (2003): 227-235.

[77] Baldwin, J.F. "Fuzzy logic and fuzzy reasoning," in Fuzzy Reasoning and Its Applications, E.H. Mamdani and B.R. Gaines (eds.), London: Academic Press, 1981.

[78] 26 U.S. Code 162 – Trade or business expenses, https://www.law.cornell.edu/uscode/text/26/162

[79] Agnar Aamodt and Enric Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *Artificial Intelligence Communications* 7 (1994): 1, 39-52.

[80]

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.375.812&rep=rep1&type=pdf

[81] Cindy Castillo, “JavaFX Architecture”, Oracle, 2013 https://docs.oracle.com/javafx/2/architecture/jfxpub-architecture.htm

Curriculum Vita (Ying Zheng)

Towson University, Towson [redacted]

7800 York Road [redacted]

Towson University, MD [redacted]

[redacted] [redacted]

EDUCATION

D.Sc.	**Towson University, Towson MD**	**September 2009 – Present**
	Information Technology	
M.S.	**University of Northern Virginia, Herndon VA**	**June 2001**
	Computer Science	
B.A.	**Towson University, Towson MD**	**December 1998**
	Mathematics, Minor in Business Administration	

TEACHING EXPERIENCES

Adjunct Lecturer **Department of Computer Science, Towson University**

Teach lecture to up to 30 students of CS/IT/IS majors; solely responsible to develop lecture notes, handout materials, and lab exercises; give lecture, and facilitate discussions with students; manage and respond to course related emails, provide tutoring to students during regular office hours and upon request; prepare and grade assignments, quizzes, and exams; maintain attendance records, and grading records for the following courses:

- **Building Java Programs (COSC236)** **2015 – 2016**
    - Introduction to structured problem solving, algorithm development and computer programming with a modern high-level structured programming language.
- **Metropolitan IT Infrastructure (ITEC201)** **2015 – 2016**
    - Technological aspects that drive the Greater Baltimore area and its surroundings by placing them in a social and economic context. Students will be able to evaluate how these technologies affect our metropolitan area's status and development by comparing our systems to the ones of other cities.
- **Visual Basic Programming (CIS265)** **2014 – 2016**
    - Concepts, tools and techniques of software development using an event-driven language that supports a graphical user interface and an object-oriented environment.
- **General Computer Science C++ Programming (COSC175)** **2010 – 2016**
    - Computer systems overview, algorithm development, data representation, software design and testing methodologies, and brief overview of advanced topics.
- **Intro to Business Programming (CIS212)** **2011 – 2014**
    - A study of computer programming for business applications using a language such as COBOL. Students will design, implement, test and

document programs in application areas such as payroll, accounting, inventory and file maintenance.

- **Computer & Creativity (COSC109) 2009 – 2014**
    - Creative activities involving symbolic manipulation and computer graphics; animation, dynamic storytelling, computer music, visual effects, Web publishing, computer games, artwork and multimedia.
- **Operating Systems (COSC439) 2010**
    - Operating systems as resource managers with emphasis on file processor, memory and device management and processes. Design and implementation of a simulated multiprogramming operating system.
- **Web Analysis & Design (CIS475) 2010**
    - Conceptual design of the Web page interface, HTML, usability testing, implementation, and management.

RESEARCH EXPERIENCES

Research Interests

- Machine Learning, Intelligent Decision Support Systems, Data Mining, Case-based Reasoning, and big data, including smart grid modeling and security.

Publications

- **Journal**
    - **Zheng, Ying,** and Zhou, Harry, "IDA: An Intelligent Document Analysis System for Evaluating Corporate Governance Practices

Based on SEC Required Filings," in proceedings of International Journal of Software Innovation, Volume 3, No. 2, 2015.

- **Conference Proceedings**
  - Zhou, Harry, and **Zheng, Ying**, "Learning from Examples: A Genetic Algorithm Approach Tested in the Domain of Finite Automata Construction," in proceedings of the International Conference on Artificial Intelligence and Software Engineering (AISE), 2014.
  - **Zheng, Ying,** and Zhou, Harry, "An Intelligent Text Mining System Applied to SEC Documents," in proceedings of the ACIS International Conference on Computer and Information Science, 2012.

WORKING EXPERIENCES

AMS Surety Holdings Corp **Columbia, Maryland** **2008 – 2009**

- **Sr. Financial/Program Analyst/Project Lead**
  - Determine system requirements and capabilities needed for a fast developing company. This includes reviewing software vendors, working with an in-house IT unit to incorporate new system into operating environment, making recommendations and facilitating implementation.
  - Evaluate the current systems and developing a strategy and framework for saleable systems.
  - Lead data conversion from QuickBooks to Dynamic GP utilizing VB, MS Excel, MS Access, SQL etc.

  - Train and troubleshoot Accounting staffs, underwrite staff with new General Ledger and new Reporting System.

<u>Total Wine & More</u> **Potomac, Maryland** **2007 – 2008**

- **Sr. Financial/Program Analyst/Project Lead**
  - Project lead to develop automated report generating system.
  - Design and implement report templates for financial statements.
  - Support business & financial database applications using VB, MS Excel, MS Access, SQL etc.
  - Perform financial analysis, various financial reports such as forecast analysis, trend analysis, and budget report etc.
  - Interprets data for the purpose of determining past financial performance and project a financial probability.

<u>Ajilon Consulting</u> **Towson, Maryland** **2005 – 2007**

- **Financial/Program Analyst**
  - Perform month-end close process and prepare financial statements for multiple districts.
  - Maintain general ledger, perform account reconciliations and analysis.
  - Generate financial reports and reconcile reports utilizing MS Access, and MS Excel.
  - Upload accounts journal entries using PeopleSoft.
  - Perform sales/margin analysis, client financial analysis, and expense analysis month end.

- Response for maintaining Access database associated with all financial reports.

<u>PMA, Inc.</u> **Washington, DC** **2000 – 2005**

- **Business Manager (06/2001 – 04/2005)**
  - Responsible for preparing detailed financial reports for eight FAA (Federal Aviation Association) programs.
  - Interact with FAA Program Managers on all financial request associated with their programs.
  - Prepare quarterly performance analysis such as sales analysis, expense analysis, and labor and material analysis.
  - Utilize five-year forecast, actual to budget, and current year to prior years to help analyzing performance.
  - Offer recommendations to improve FAA programs' performance.
  - Estimate costs associated with FAA programs.
  - Generate and test financial reports utilizing Oracle, and SQL.
  - Act as a program lead to chair monthly meetings, gather information for status reports, and reconcile financials for all FAA programs.
- **Financial Analyst (10/2000 – 05/2001)**
  - Support FAA Program Managers by providing cost analysis, circuit cost estimates, and financial forecasts for FAA detailed program reports.
  - Receive Outstanding Employee Award from FAA.

- Promote to Supervisor in 02/2001, and become responsible for training team members.

Johnson Actuarial **Washington, DC** **1999 – 2000**

- **Pension Administrator**
  - Responsible for handling pension plans for multi-size organizations.
  - Prepare quarter, semi-annual, and annual financial reports for client companies.
  - Prepare annual tax forms for client companies' pension plan.
  - Responsible for trouble shooting files and print-sharing problems for MS Network users.

SERVICES

Towson University

- **President** **2015 – 2016**
  - Doctoral Students Computer Science Association (DSCSA)