The International Journal of Digital Accounting Research
Vol. 17, 2017, pp. 33-65
ISSN: 2340-5058

A HYBRID METHOD FOR TAXONOMY CREATION

Vasundhara Chakraborty. Monmouth University. vchakrab@monmouth.edu

Miklos Vasarhelyi. Rutgers University. USA miklosv@andromeda.rutgers.edu

Abstract. Automatic extraction of information from the footnotes of financial statements can be particularly challenging due to a wide variation in filing structure and terminologies. Standardized text and use of data tagging can facilitate this process. This paper: (i) proposes and demonstrates a new hybrid method of taxonomy creation, using historical data; (ii) compares the taxonomy structure using the new method, with that of the existing XBRL US GAAP taxonomy; (iii) shows evidence of structural differences between the official XBRL US GAAP taxonomy and the new hybrid taxonomy and (iv) demonstrates how the tool so developed could be used for more exploratory research. Comparison of this new structure with that of the existing XBRL taxonomy structure reveals that its creation from historical data provides a greater level of aggregation compared to the XBRL US GAAP taxonomy for Pension footnotes.

Keywords: Taxonomy, XBRL, Text analytics, Pension, 10K

DOI: 10.4192/1577-8517-v17_2

Submitted February 2017
Accepted May 2017

1. INTRODUCTION

1.1 Overview

Credible, prompt and precise financial information is an indispensable tool for investors and analysts alike. Important and relevant information is not only limited to the tabular contents of financial statements, but also includes what can be extracted from disclosures. Such disclosures may have a free format, or appear as a combination of both numerical and textual data. Retrieving useful information is a complicated process. As a result of the Sarbanes-Oxley Act, Securities and Exchange Commission (SEC) actions, and recent shareholder lawsuits, some firms have decided against furnishing earnings guidance. As a result, many investors feel handicapped by the absence of any good, independent, analysis concerning potential investments. Mandatory XBRL filing could be a step in the right direction to solving this issue. However, development of data tags is an indispensable part of this process. In this essay a new, hybrid, method for creating taxonomies is proposed, which uses a combination of manual and automatic steps. Also, an analysis of the basic structure of this new taxonomy is conducted, as is a comparison to the existing XBRL taxonomy.

1.2 Motivation and Research Questions

Financial statements submitted to the SEC are an indispensable source of company information for a wide variety of users. Such users include company management, investors, creditors, governmental monitoring agencies, and the IRS. The footnotes expand on the quantitative and qualitative free-format information financial statements, by providing qualitative information that allows for a greater understanding of a company's true financial performance. Unfortunately, there are no standards for the clarity or conciseness of the wording that is used in footnotes.

Standardization of text, along with the maintenance of a hierarchy, is necessary for finding relevant information within the vast amount of data available. Liberal use of synonymous or polysemous terms and phrases inserts an added challenge to achieving the standardization of texts and formalization of different processes. For example, Appendix I shows a list of all the different sub-sections within pension disclosures. It includes the number of synonyms, or different variations, found for each of these sub-sections. This list was created after automatically extracting the

pension footnote sections and tables therein from the 10K statement samples used. The general heading for Pension disclosures has the highest number of variations at twenty-five. The different categories under pension disclosure have a maximum of eight variations. These variations need to be standardized through a formalized process. Appendix II lists the detailed line items within each of these sections, and the number of synonymous terms or variations found. To fulfill this requirement, the generation of a comprehensive taxonomy is vital. An essential ingredient for accomplishing this task is the creation of legitimate data tags and the mapping of line items using those tags. This would enable both producers and consumers of financial information, to save the resources that would usually be used in these traditionally manual processes.

The SEC announced in February 2005, that it would start accepting voluntary filings in XBRL format. This was followed, in May 2008, by its decision to make XBRL filing mandatory. During this time they also required that companies begin filing financial statements in an interactive, tagged, format starting with fiscal periods ending in late 2008. The XBRL data format allows for the retrieval of line items from financial statements. Such retrievable information includes information from balance sheets, income statements, and statements of cash flow. However, extraction of specific line items from footnotes is relatively uncommon. This is simply because data tags for footnotes may not be available. Moreover, the creation of these tags must be done manually using a normative approach. Because filers have considerable flexibility in how financial information is reported under U.S. reporting standards, it is possible that a company may choose to use non-standard, company-specific extensions. Unfortunately, the use of non-standard company specific extensions may introduce errors. This is due to a reduced ability to compare data prepared with a heavy usage of extensions. As per recommendations from the SEC, the use of company-specific extensions should be very limited. Despite this, filers often continue to use nonstandard labels traditionally familiar to them, instead of adapting to the standard tags. Piechocki et al. (2009) have shown that in recent filings, XBRL filers very often extend the U.S. GAAP taxonomy (the generally accepted accounting principles adopted by U.S. Secretaries and Exchange Commission), even though suitable tags were available for these items. This happens when instead of using available tags companies makeup their own leading to repetition. Filers tend to add elements to the base taxonomy, even when the base taxonomy already includes a semantically

equivalent element. Piechocki et al. (2009) speculated that this might be due either to the large size and complexity of the taxonomy, or a perfunctory search through it to look for semantically matching items. As a result, the filers fail to find the appropriate match. In some other cases filers may have used unique terms, which find their way into the reports in the form of newly added extensions.

Bovee et. al.(2002) argued that a poor fit may lead to information loss and a subsequent resistance in the adoption of the taxonomy. In their view although XBRL is likely to improve comparability and consistency of financial reporting, and may facilitate near-continuous reporting, some questions arise regarding correspondence of the proposed taxonomy with the firms' preferred reporting practices. Bovee et al. (2005) proposed that there should be an empirical approach towards the evaluation and improvement of XBRL taxonomies, so that the taxonomy matches with historical data. Bovee et. al. (2002, 2005) opined that there is a need to empirically evaluate taxonomies.

The method used by XBRL US to develop tags has evolved over the years. In preparation for the release of the 2007 version of the XBRL US GAAP taxonomy, experts from six accounting firms came together to decide upon the tags to be used. Audit compliance checklists of these six firms were consulted as a reference point for developing XBRL tags. This team reviewed all the regulations in US GAAP as a part of the taxonomy development process. The preliminary taxonomy was then sent to the FASB for an initial review. A relevant subject matter expert at FASB then reviewed each element that was created. Simultaneously, the SEC did a similar review. They engaged corporate finance specialists and asked them to analyze existing elements and add others if necessary. After reviews by the auditors, the FASB, and the SEC, the complete list of elements was sent to XBRL US for testing. This testing was conducted by mapping real financial statements to the XBRL taxonomy. Line items that could not be mapped were added as additional elements.

As regulations change, firms must report differently. Therefore, taxonomies may also need to be updated to adapt to these changes. In the 2008-2016 period, seven versions of the US GAAP taxonomy were published. Manually processing relevant information and creating tags for each of these changes can be very complicated and time consuming. In this paper a hybrid method is proposed, which uses historical data from company filings for taxonomy creation, and a

combination of manual and automatic steps to create a taxonomy. This will prove that part of the process can be automated, thus raising the first research question:

RQ1: How can taxonomies be created automatically using historical data from financial statements?

Merely creating the taxonomy automatically is not a panacea. It should be comprehensive enough to include all or most of the elements of the text. To determine whether the taxonomy created using this hybrid method is more effective than one created manually, it is imperative to analyze their differences if any exist. From this emanates the second research question:

RQ2: What are the structural differences between the XBRL pension footnote taxonomy available, one created using a manual method, and a new taxonomy created by a hybrid method?

Semantic parsing techniques have been used to extract data from the pension footnotes of 10K filings of firms in the data corpus. Text analysis techniques, such as object filtering, automatic indexing, and cluster analysis (Chen et. al. 1992, 1994; Chuang et. al. 2005), are used to develop a hybrid taxonomy building process.

In this case our hybrid method is used to focus on financial statement footnotes

However, this method does possess the, potential for application to a much wider variety of uses in the development of taxonomies for other types of objects such as controls, risks, or audit methods.

The rest of the essay is organized as follows: Section II discusses prior research; Section III discusses the methodology and Section IV discusses the results, conclusions and limitations.

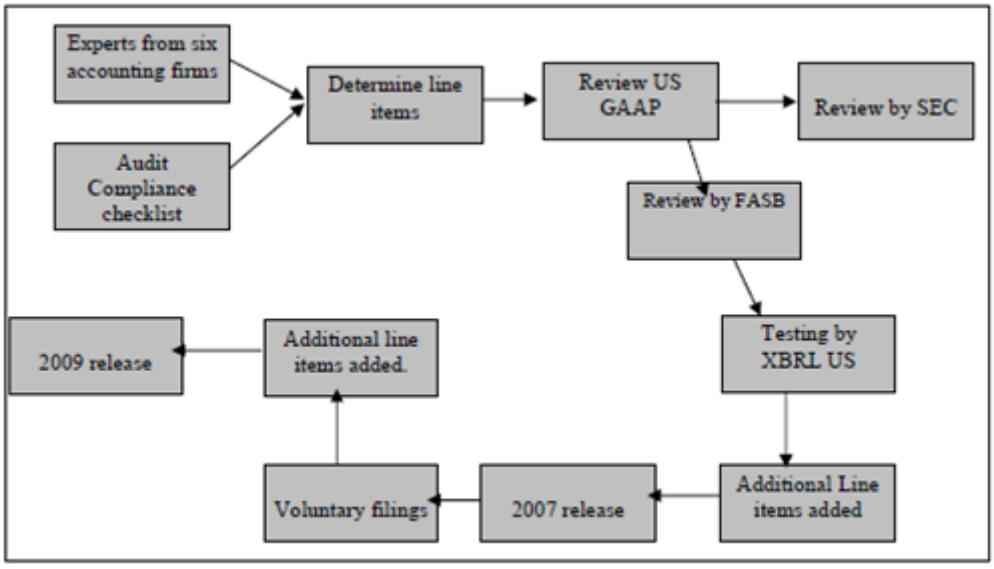

Figure 1. Method used by XBRL US GAAP

2. PRIOR RESEARCH

2.1 Extracting accounting information from financial statements

There is an abundance of electronic data available to users regarding the financial performance of firms. While it is common to perform automatic analysis of the numbers in financial statements, it is generally difficult to autonomously extract meaning from the textual portions of financial reports. Due to an absence of standardization, information can be easily obfuscated in these textual portions. A human expert can decipher the deeper meanings behind the use of synonymous terms, and comprehend why certain words were used, and others not. Machines or computers cannot easily interpret these nuances without the existence of a pre-established pattern or logical framework. Since such standardization does not exist it poses an increased difficulty to extracting textual financial information from statements. EdgarScan[1] was built as one of the earlier efforts using text

[1] It is an interface to the United States Securities and Exchange Commission Electronic Data Gathering, Analysis and Retrieval (SEC EDGAR) Filings. EdgarScan pulls filings from the SEC's servers and parses them automatically to find key financial tables and normalize financials to a common format that is comparable across companies. Using hyperlinks we can go directly to specific sections of the filing, including the financial statements, footnotes, extracted financial data and computed ratios. EdgarScan was created by PwC (2001) and made publicly available on the Internet in response to the need for automatic extraction of accounting numbers from the EDGAR filings.

analysis to provide a solution to some of these problems. It was created by PwC in response to the need for automatic extraction of accounting numbers from the EDGAR filings and was made publicly available on the Internet.

Nelson et al. (2000) created the EDGAR agent to show how intelligent agents can be used to gather financial information. They stress how intelligent agents can be used to bring about organizational changes, and mention their potential for making improvements in the field of auditing. However this system only processes quarterly SEC filings, identifies only a few of the most important accounting numbers, and only interacts with a single online information source (the SEC EDGAR repository). Their agent searches the SEC EDGAR database for the current cash balance of a company, and can calculate financial ratios based on this information. This can be helpful when comparing the performance of a company with industry average information. Their agent can also calculate the current market value of a company using information on its stock price, quick ratio, current ratio, and gross margin over sales.

Over time various other types of tools have been developed which provide access to SEC filings. Three general classes of tools have emerged—third party[2], free, and commercial tools (Gerdes, 2003). Some of the free tools analyzed by Gerdes (2003) include 10K Wizard, EDGAR, EdgarPro, EdgarScan, Freedgar, Edgar-Online, Search-SEC, SEC Info. Whereas some of the commercial tools included Disclosure, Edgar Direct, Global Access, Edgar-Online, Lexis/Nexis, Livedgar, SECnet. These tools were some of the earlier attempts at financial data extraction. In more recent times XBRL's usefulness has been explained in Vasarhelyi et. al. (2011). With the use of XBRL, numerical data can now be downloaded into spreadsheets.

2.2 The need for hierarchical information

The inability of users to promptly locate information within textual accounting documents, has led to more elaborate and time-consuming research efforts. Fisher (2004) argues for consistency in the structure of financial reports. Fisher cites the use of section headings, the sequencing and numbering of sections, and the development of a hierarchal structure as potential means to improving information

[2] The third party providers are secondary sites, which are dependent on the primary providers for their content. Their own capabilities might vary depending on resources. Some may provide real time data but historical data for only 3 weeks is available in a summarized format with no provision for searches. In some other cases it could be functionally equivalent to the parent websites.

retrieval. The importance of storing data hierarchically for easier extraction of information has also been mentioned elsewhere in the literature. It is as important to capture and control the knowledge base underlying accounting decisions, as it is to develop the systems that automate accounting functions. Document structure is significant because of its relationship to understandability, accessibility, and retrieval precision. Fisher (2004) suggests that XML is a good choice for document structure because XML DTD[3] allows for the defining of data. Similarly, Routen and Bench-Capon (1991) argue that hierarchical formalization of text is preferable for the creation of knowledge-based systems. Gangolly (1995) has also recommended the adoption of hierarchical formalizations of meta-level information in FASs. This may also be applicable in the case of financial statements, or other documents containing financial information. Gangolly (1995) suggests hierarchical structuring of accounting standards based on three points: 1) distinguishing between changes in the original standards; 2) maintaining meta-level information and date stamps of such changes; 3) separation of other meta-level information. Such structuring could facilitate the development of knowledge-based systems. Federal Accounting Standards Board (FASB) undertook the Codification project in 2009, which addresses the importance of a uniform and consistent document structure. The development and incorporation of a topical arrangement with corresponding subtopics provides the beginnings of a hierarchical structure (Fisher et. al. 2009). The formal specification of prescribed and ordered sequences of topics, subtopics, and sections, provides a consistency in structure that can facilitate the development and implementation of indexing and searching procedures (Fisher 2004).

2.3 Examples of various methods used

Some of the early methods used include Edgar2xml, and Coding Agent. Leinnemann (2001), introduced text mining techniques in order to implement Edgar2xml. This software agent extracts fundamental company data from text in the SEC's EDGAR database, and outputs this data in a format that is used to support stock market trading decisions. Katriel (1997) described a computer system called Coding Agent, which is capable of assigning category codes to

[3] The purpose of a DTD(Document Type Definition) is to define the legal building blocks of an XML document. It defines the document structure with a list of legal elements. A DTD can be declared inline in an XML document, or as an external reference.

short texts. Semantic parsing and text mining techniques were used in the development of Coding Agent.

Chen et al. (1992) describe a detailed layout of the steps taken to generate a thesaurus automatically and evaluate it for the worm community system (WCS). This was followed by Chen et al. (1994) who created a thesaurus for Drosophilia information. Chuang et al. (2005) described taxonomy generation for text segments using a hierarchical algorithm. In this paper, semantic parsing, together with object filtering and automatic indexing, is used to extract data from 10K statements. A hierarchical agglomerative algorithm is then applied to generate the taxonomy automatically.

Garnsey (2006) used semantic parsing techniques to determine the feasibility of applying statistical methods. This was done in order to automatically group related accounting concepts together. Vasarhelyi et al. (1999) expand the traditional financial audit framework, and argue in favor of the scope of evidence collection to cover on-line corporate information (particularly the news), using semantic analysis methods.

One instance where semantic analysis is considered very useful is for the extraction of information from financial statements made available by intermediaries. Bovee et al. (2005) extracted accounting numbers from financial statements available in EDGAR. They match the line-item labels, and associated numbers, to synonyms of tags that exist within the XBRL taxonomy. This helps to convert consolidated balance sheets, income statements, and statements of cash flow into an XBRL-tagged format.

Wu et al. (2000) studied the feasibility of automatic classification of financial accounting concepts through a statistical analysis of term frequencies used within the financial accounting standards. The procedure makes use of a principal-components analysis to reduce the dimensionality of the dataset. It then uses an agglomerative nesting algorithm (AGNES) to derive clusters of concepts.

The classic literature in this context by Salton (1989), presents a blueprint for automatic indexing that typically includes dictionary look-up, stop-wording, word stemming, and term-phrase formation. Crouch (1990), and Crouch et al. (1992) used a complete-link algorithm for the automatic generation of a global thesaurus.

Chuang et al. (2005) describe taxonomy generation for text segments using a hierarchical algorithm.

Kothari et al. (2009) use dictionary methods to calculate the number of positives and negatives in each disclosure text. Feldman et al. (2010) use a classification scheme of words, separating them into positive and negative categories. These categories are then used to measure any tone change in an MD&A section as it relates to prior periodic SEC filings. Kravet et. al. (2011) extract risk disclosures from 10-K forms by searching for sentences that involve predefined, risk-related keywords. Rogers et al. (2011) use both general-purpose and context-specific text dictionaries to quantify optimistic and pessimistic tones within a firm's earnings announcement. Some supervised learning methods include Li (2010), which uses a naive Bayesian classifier to classify the tone and content of forward-looking statements contained in corporate 10-K and 10-Q filings. Huang et. al. (2011) develop a multi-label text classification algorithm to classify risk factors in section 1A of 10-K form into 25 risk types. Humpherys et al. (2011) use linguistic features to distinguish fraudulent from non-fraudulent 10-K reports using readily available classifiers. Cecchini et al. (2010) develop a method that automatically creates an ontology for the text in an MD&A section of a 10-K form. This could then be used for classifying the financial events of firms. Grimmer et. al. (2011) develop a computer-assisted method for discovering insightful conceptualizations through the clustering of input objects.

The use of unsupervised clustering methods to analyze texts is very limited. Yang (2014) introduces unsupervised learning methods into the field of financial accounting. They simultaneously discover and quantify risk types from textual disclosures and propose an unsupervised topic model, which they call sent-LDA. Brent et al. (2011) built a coding scheme by training pre-defined categories used to assign a code to text documents automatically. This was done in order to analyze economic crises through newspaper articles.

Hagenau et al. (2012) designed a text classification approach for processing financial news that is used to automate stock price prediction. In terms of financial analysis Beattie et al. (2004) introduced a comprehensive four-dimensional framework for the content analysis of accounting narratives. It uses a coding system based on four attributes in order to give structure to accounting texts. Gray et al. (2014) provided a taxonomy to guide research in the application of data

mining to fraud detection in financial statement audits. Fisher et al. (2010) and Chakraborty et al. (2014) applied text and data mining to automatically classify academic articles in accounting and improve understanding of the accounting lexicon.

2.4 Using the Hierarchical clustering algorithm

Hierarchical clustering algorithms come in one of two types: i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis). In this paper we use the agglomerative clustering method (Tan et. al. 2005). This algorithm works by grouping data points one by one on the basis of the nearest distance measure of all the pairwise distances between a particular data point and the others. There are many available methods for distance calculation. These include: (i) single-nearest distance or single linkage; (ii) complete-farthest distance or complete linkage; (iii) average-average distance or average linkage; (iv) centroid distance; (v) ward's method, where the sum of squared euclidean distance is minimized. These methods were built into the CLUTO tool, which is used in the application of clustering methods. Several examples of its successful implementation can be found in the related literature. Müller (1999) uses hierarchical clustering to automatically generate a taxonomy in a large collection of documents. Chuang et. al. (2002) use a hierarchical agglomerative clustering algorithm to group similar queries and generate cluster hierarchies using a cluster partition technique.

In this essay, parsing is used together with object filtering and automatic indexing to extract data from 10K statements. Then, a hierarchical agglomerative algorithm is applied to generate the taxonomy structure automatically. This method is appropriate since we conduct an explorative study where we intend to find any emerging structure. This is as opposed to if we were trying to confirm pre-existing beliefs as in a confirmatory study.

3. METHODOLOGY

3.1 The sample

A data corpus of 10K filings submitted by public companies was created. Data to be used pertained to companies that were randomly selected from the Fortune 500 list. Altogether, one hundred and twenty 10K statements from as many different

firms were manually downloaded from the SEC Edgar website. Eighty were used as the training dataset and forty as the test dataset.

3.2 Generating the pension taxonomy

Figure 2 is a broad overview of the steps taken to partially automate the process of pension footnote taxonomy creation. The first stage, data collection and restructuring, involved downloading the data, writing a program, which changes its structure. This is then, used to extract text from the documents, and preform a word count to determine the frequency of occurrence of line items.

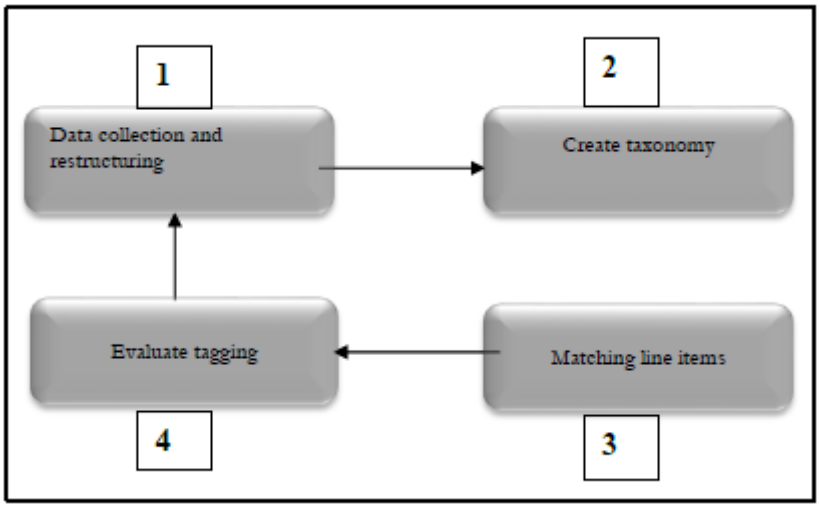

Figure 2. An overview of the proposed method to generate hybrid taxonomy

In the second step, the taxonomy structure was created. In the third step, line items were mapped to the tags created based on this taxonomy. This was done using a program that was created as part of this process. Finally, line items that could not be mapped to any of the tags were analyzed manually. Any new and frequently occurring terms were added to the terms database, the remainder were checked for synonymity to those in the database. The four steps were repeated thrice before arriving at the final results. Fig. 3 is a flow sheet diagram explaining various activities involved within each of these four steps. Steps shown within the dotted line were repetitive. In the explanation below each numeral identifies a step's sequence number. The whole process consists of a combination of

automatic and manual steps. The manual steps are shown in gray and the automatic steps are in blue.

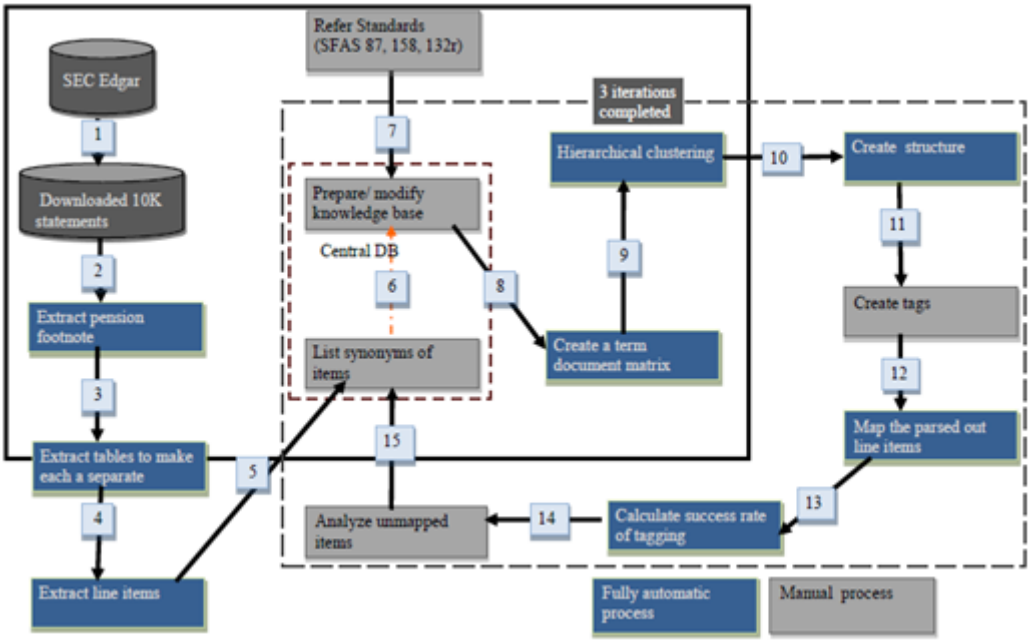

Figure 3. Details of the procedure for taxonomy creation

Some of the steps mentioned above are explained in detail below:

Creation of the knowledge base

Object filtering: This process involves the creation of domain-specific keywords and their synonyms. Based on the specifications of SFAS 158, a list of items was prepared. Each of these should ideally be included in the pension footnote taxonomy.

Automatic Indexing*:* The 10K statements were filtered using the list that was prepared during object filtering. An automatic indexing routine then processed the remaining text. The automatic indexing routine is a computer code written for this purpose. The following steps are part of the automatic indexing procedure:

(1)Word identification: The remaining words (after object filtering) in each document were identified.

(2) Stop-wording: To eliminate unwanted words a combination of two stop-word

lists[4] were used. Word stemming was not used as this process can completely change the context of a word, which in many cases can be crucial.

Developing the Parsing tool: This is a very crucial step since variations found in the reporting styles of pension disclosures (in the data corpus), make the application of readily available parsing tools like CMU Toolkit or IBM Textminer almost impossible. The variation in the usage of terminologies to express similar content in financial reports makes the comparison of data difficult and the use of generic tools for parsing ineffective. Hence, to satisfy the very specific needs of this project a new tool was created. This tool consists of a group of software programs each of which is designated with a specific function. The parsing tool was created by writing a tailored program, which operates in Visual Studio.

Restructuring the data format: Each pension disclosure extracted from the statements can be considered a separate document. Each of these documents is similar in content and structure. Therefore, applying a clustering algorithm on this sample would result in them all being grouped under one category. In order to rectify this issue, it was necessary to restructure the data. First, the pension disclosure was extracted from the 10K statements. Next, each of the tables were extracted and stored separately. This resulted in the formation of several documents originating from each single pension disclosure. This set of newly formed documents could then be clustered effectively.

Creation of the document-term matrix: The creation of the document-term matrix is an integral part of the process since the document-term matrix provides statistics on frequency of term usage. This information is necessary for the application of the hierarchical clustering algorithm. A word count is used to measure the number of times a word or phrase occurs in a document. This information is then used to calculate each term's total frequency among the collected documents. Terms appearing at least twenty times were included in the list of synonyms and subsequently added to the knowledge base. This cutoff was decided based on the length and number of documents that are used in the sample. Ultimately however, the final decision is subjective. In this case because the terms used by companies were so varied, the frequency of occurrence cutoff could not

[4] The first stop word list contains five hundred and seventy one words and was built by Salton and Buckley for the experimental SMART information retrieval system at Cornell University. The second stop word list was obtained from the Onix Text Retrieval Toolkit. The function of a stop word list is to eliminate frequently occurring words that do not have any semantic bearing.

be set high. Doing so would possibly eliminate other essentially similar terms from being included in the matrix due to slight textual differences. On the other hand, not having any cut off number would also be dangerous. This is because then the list would theoretically include all terms within the dataset. Such a result would defeat the purpose of this exercise. After manually reviewing the results for a few terms, it was decided that twenty could be an acceptable number. This allows for the inclusion of terms, which could potentially result in the emergence of a pattern, but not all terms. The document-term matrix resulting from the training dataset has eighty rows (for the number of companies) and one thousand two hundred and seventy five columns.

Creating a hierarchy of terms and the subsequent creation of taxonomy: Application of an agglomerative algorithm on the document-matrix leads to the creation of hierarchical groups. This acts as the basic structure of the taxonomy. Hence, automation can facilitate the creation of a hierarchical structure and selection of terms, but does not generate hypercubes[5] or domains. This was taken into consideration when comparing the new taxonomy with the official XBRL taxonomy.

Term matching in the test set data: With the automatically created structure as reference, data tags were constructed manually. The tagging process began by first identifying all the elements, their names and their corresponding labels. The line items from the test dataset of forty 10K statements were then extracted and matched with the tags. A program was tailor made for the specific purpose of term matching. Any unmatched terms were automatically flagged.

Analyzing the unmapped items: Automatic matching of line items against existing tags was not always completely successful. Some unmatched items remained. These were stored in an Access database for future reference. Several reasons exist that could explain why the terms remained unmapped. These include the absence of a matching tag, the absence of a matching synonym, or inefficiency on the part of the parsing program. Analysis of the unmapped items stored in the database, led to the identification of new synonyms for line items and made the

[5] A hypercube represents a set of dimensions. Hypercubes are abstract elements in the substitution Group of hypercube Item that participate in has-hypercube relations and hypercube-dimension relations. For more information please check: http://www.xbrl.org/Specification/XDT-REC-2006-09-18+Corrected-Errata-2009-09-07.htm

tagging process comprehensive. These new terms or phrases were incorporated into the knowledge base. This marked the completion of one cycle and the beginning of the next. This cycle was repeated thrice before arriving at the final result. The decision to restrict the number of repetitions to a particular figure was arbitrary, intuitive, and dependent the analyst's judgment.

3.3 Generic use of the Tool

The main goal of this project was to design and develop a method for taxonomy generation. As a part of this process a tool for data extraction had to be created. If this were to be reconfigured as a generic tool, it could be used for data extraction from other disclosures as well. To help achieve this goal a central database was created which consisted of 12 tables. Storage of information in such tables increases accessibility of data. Specific table design features, such as adding unique identifiers, helped make the tool more generic. In this case these tables represented features related to the pension footnote. For other instances this database would need to be redesigned to facilitate application of the tool within it is new context.

4. RESULTS

Results were obtained by performing an evaluation of the methods success rate for tagging items. A visual representation of the phases shows how the clustering works. Finally an analysis of the results is performed in order to better understand the clusters that were obtained.

4.1 Performance evaluation

	No of occurrences	Number correctly identified	Success Rate (%)	No of occurrences	Number correctly identified	Success Rate (%)	No of occurrences	Number correctly identified	Success Rate (%)	No of occurrences	Number correctly identified	Success Rate (%)
Training dataset	78	78	100	1890	1834	97	21000	20580	98	22968	22492	97.92
Test dataset	40	40	100	1060	996	94	15400	14784	96	16500	15820	95.8

Table 1. Performance evaluation for data tagging

It was necessary to establish the comprehensiveness of the generated taxonomy. This is a measure of how well the pension disclosure data can be mapped to the data tags created using the generated taxonomy. The first step was to calculate the success rate for tagging different data items. "Success rate" is defined as the number of items correctly mapped to thier corresponding tags, when such tags exist. This is compared to the total number of mapped items (Bovee et. al. 2005). Evaluation of the parsing tool was conducted using both the test dataset and the training dataset. As shown in Table 1, in the training dataset, each of the seventy-eight existing Pension headers were correctly identified by the parsing tool. Therefore, the success rate is 100%. This was manually evaluated. Similarly, for the test dataset, all of the forty Pension headers were correctly identified, indicating a success rate of 100%

Next, the same process was applied on the detailed paragraph in the disclosure. Out of 1,890 categories present in the training dataset, 1,834 were correctly identified. This yields a success rate of 97%. In the test dataset out of the 1,060 existing instances, 996 were correctly identified giving a success rate of 94%. Within each category there were several line items, which needed to be identified and extracted using the parsing tool. Of the 21,000 line items, 20,580 were correctly identified, yielding a success rate of 98%. The overall success rate is 97.92% for the training dataset, and 95.8% for the test dataset. Manual checks were carried out to establish the success rates at different levels, by determining whether a term was correctly tagged when it actually existed.

4.2 Visualization of clusters

Applying a hierarchical clustering algorithm to our document matrix, generating clusters of words or phrases used in the pension disclosure. These clusters form the basic reference point for creating the Pension taxonomy. Figure 4 shows, in a matrix format, how these word clusters were formed. Shades of different intensity were used to graphically represent the predominance/infrequency of different terms present in the word pool. White or light grey shades indicate values near zero, while darker shades indicate larger values. In the figure, black horizontal dividers separate the different clusters. The X-axis shows a partial list of the phrases from the word pool. The Y-axis is divided into several clusters, using horizontal bars.

Interpretation of this matrix can be explained by examining one of the clusters, highlighted in fig. 4. The phrases like "actuarial losses," "benefit obligation end of year," "benefit obligation beginning of year," etc. are shown in dark shades, implying that the aforementioned phrases are predominant. They also fell within the brackets of two horizontal lines forming a cluster called "Change in Benefit Obligation". This visual was useful as a reference while developing the taxonomy structure and was handy when finalizing the Pension taxonomy structure.

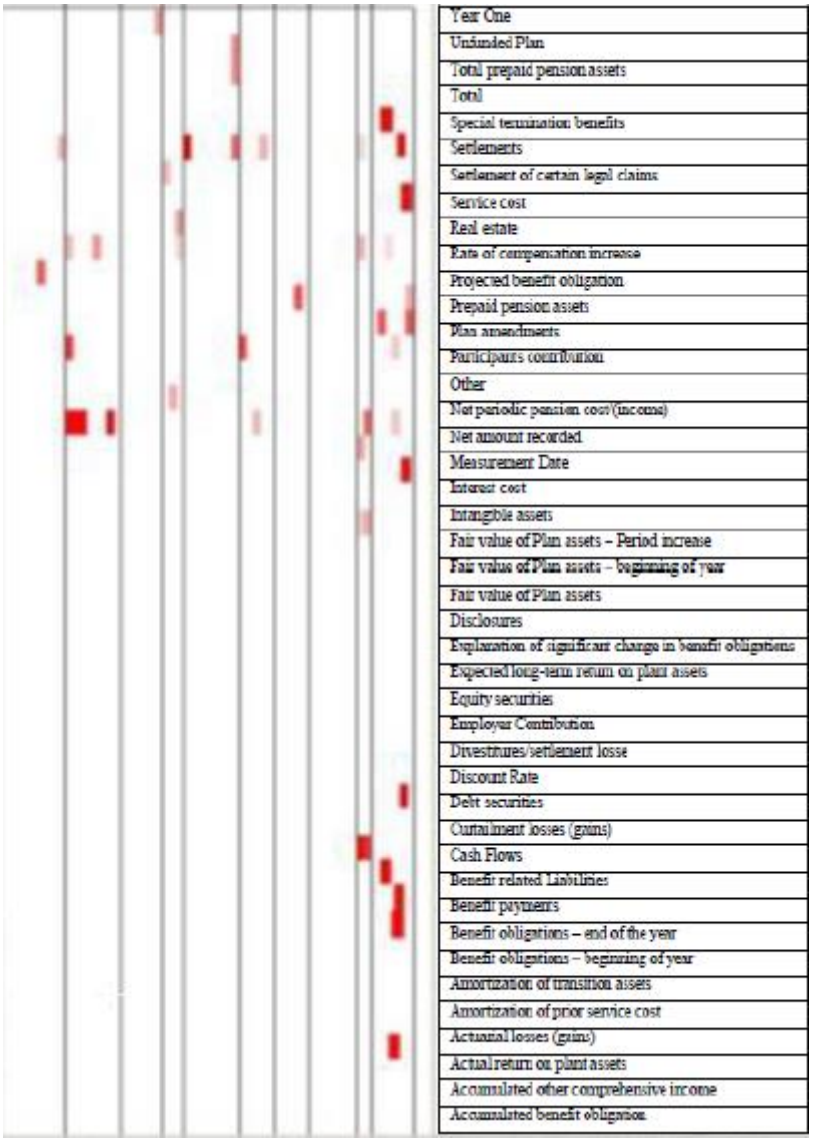

Figure 4. Concentration of phrases in different clusters

4.3 Analysis of results: Comparison of taxonomy structure

After the creation of the new Pension taxonomy structure, it was compared to the existing structure of the official XBRL Pension taxonomy. This is important for two reasons: (i) a comparison helps reassure whether this new hybrid method works and is a worthy replacement for the present normative method; (ii) it aids in the understanding of any differences which may exist between the two methods.

Comparison between the two taxonomies presented here is conducted as a general comparison of their hierarchical structure as opposed to a measurement of any mathematical distance between them. Tables 2 and 3 show some examples of the differences between the structures of the two taxonomies. They present a listing of the different line items used for pension disclosure reporting, details about which of these line items were found in the hybrid taxonomy, and which ones can be found in the XBRL taxonomy. The occurrence of a line item in a particular taxonomy is indicated by marking "Yes" under its respective column. All the major categories were found in both taxonomies.

The "change in benefit obligations" section shown in Table 2 is fairly constant amongst firms. It is therefore part of the Hybrid taxonomy, but it can also be found in the official XBRL taxonomy.

"Change in fair value of plan assets", shown in Table 3, is a category where some line items occur only in the Hybrid taxonomy, while other line items appear in both, but are more detailed in the XBRL taxonomy. "Funded benefit obligation", "unrecognized net loss", "unrecognized prior service cost", "net amount recorded" are some line items found exclusively in the Hybrid version and do not appear in the XBRL taxonomy. It can also be concluded that firms generally aggregate the "actual return on plan assets" and "plan, purchases, sales and settlements" despite these being disaggregated under the XBRL taxonomy.

Under the category "Amounts recognized in the Consolidated Statement of Financial Position" the line item "prepaid pension assets" was further disaggregated in the XBRL taxonomy, but appeared as a single consolidated term in the Hybrid version.

CATEGORIES			XBRL	Hybrid
Change in benefit obligations			Yes	Yes
	Benefit obligations, beginning of year		Yes	Yes
	Service cost		Yes	Yes
	Interest cost		Yes	Yes
	Plan amendments		Yes	Yes
	Participants' contribution		Yes	Yes
	Actuarial losses (gains)		Yes	Yes
	Gross Prescription Drug Subsidy Receipts Received		Yes	No
	Business Combinations and Acquisitions, Benefit Obligation		Yes	Yes
	Divestitures		Yes	No
	Foreign Currency Exchange Rate Changes		Yes	Yes
	Benefit payments		Yes	Yes
	Special termination benefits		Yes	Yes
	Settlements		Yes	Yes
	Curtailment losses (gains)		Yes	Yes
	Settlement/curtailment/acquisitions/dispositions		Yes	Yes
	Benefit obligations, end of year		Yes	Yes

Table 2. Comparison of Taxonomy Structure for Category "Change in benefit obligations"

"Information for pension plans with an accumulated benefit obligation" and "Weighted-average asset allocation of the pension and postretirement plans" are two major categories existing in the Hybrid taxonomy created from the historical data. The "Weighted average asset allocation" does not appear as a standalone category in the official XBRL taxonomy. Some of the required terms within this category such as "Equity Securities", "Debt Securities", "Real Estate" and "Other", occur in the XBRL taxonomy but in a different position and under a different major category than in our hybrid taxonomy. The major category "Information on plan assets" appears in both types of taxonomies. Under this major category there is a sub-category for "Weighted average asset allocation".

CATEGORIES			XBRL	Hybrid
Change in fair value of plan assets			Yes	Yes
	Fair value of plan assets, beginning of year		Yes	Yes
	Actual return on plan assets		Yes	Yes
		Actual Return on Plan assets still held	Yes	No
		Actual Return on Plan assets sold during period	Yes	No
		Actual Return on Plan assets, Total	Yes	No
	Employer contributions		Yes	Yes
	Participants' contribution		Yes	Yes
	Fair value of plan assets, Period Increase (Decrease)		Yes	Yes
	Foreign Currency Exchange Rate Changes		Yes	Yes
	Transfers between Measurement levels		Yes	No
	Benefit payments		Yes	Yes
	Plan, Purchases, Sales and Settlements		Yes	No
		Business Combinations and Acquisitions, Plan assets	Yes	No
		Divestitures, Plan assets	Yes	No
		Settlements, Plan assets	Yes	No
		Purchases, Sales and Settlements, Total	Yes	No
		Settlement/curtailment/acquisitions/dispositions	No	Yes
	Fair value of plan assets, Period Increase (Decrease)		Yes	Yes
	Fair value of plan assets, end of year		Yes	Yes
	Funded (unfunded) benefit obligation		No	Yes
	Unrecognized net loss		No	Yes
	Unrecognized prior service cost		No	Yes
	Net amount recorded		No	Yes

Table 3. Comparison of taxonomy structure for category "Change in fair value of plan assets"

Implementation of FAS 132r (a) calls for more granular reporting by firms, but clear specifications on this aspect have not been provided. Analyzing historical data to assess what kind of granular reporting firms have been opting for, could expedite the process of expanding the XBRL taxonomy. In keeping with expectations, some new terms like "US stocks" and "International stocks" provide differentiation. Others like "Long duration bonds" and "Alternative investments"

have been found and are only included in the new Hybrid taxonomy. Another major category "Information for Pension plans with an accumulated benefit obligation in excess of Plan assets" is present in both the taxonomies. However, line items within this category such as "accumulated postretirement benefits obligation" (APBO) and "Accumulated Benefit Obligation (ABO) Less Fair Value Of Plan Assets" can be found in the historical data even though they are no longer in use after the implementation of FAS 158.

Major categories present in both the Hybrid and XBRL taxonomies include "Funded Status of the Plan", "Unfunded Plan", "Accumulated Benefit Obligation", "Amounts Amortized from Accumulated Other Comprehensive Income (Loss) in next Fiscal year", "Pension plans with a benefit obligation in excess of plan assets", "Explanation of Significant change in Benefit Obligations or Plan assets not apparent from other disclosures", "Measurement Date", and "Pension plans with a Accumulated benefit obligation in excess of plan assets".

5. CONCLUSIONS, LIMITATIONS AND FUTURE RESEARCH

The objective of this study is twofold: (1) To develop a hybrid method for taxonomy creation and (2) to compare the structure of the taxonomy created using the hybrid method against the XBRL US GAAP taxonomy. The term 'hybrid' is used in reference to the combination of manual and automatic processes that are present within the methodology of this study. A prototype tool was designed and developed to extract and restructure information from the pension footnotes of 10k statements. A hierarchical clustering algorithm was applied to the data for the development of a taxonomy structure. The parsing module was evaluated. It functioned well, with an overall success rate of 97%, for the training data set and 95% for the test data set. Comparison of the structure of the hybrid taxonomy with that of the XBRL taxonomy, revealed some differences between the two. In general it was found that companies tend to aggregate some of the data, whereas a more disaggregated structure is followed in the XBRL taxonomy.

The parsing tool developed as a part of this process, could potentially be used for other research. Its usefulness has been demonstrated by comparing data from randomly selected firms, which appear on the Fortune 500 list over a ten-year span. It was revealed that in some cases the companies added new terms or even a

completely new category into their filings. Future research may be carried out that explores some of these trends in pension footnote reporting.

Future research could also make use of the hybrid method in this paper. Here it is used to focus on financial statement footnotes. However, it can also be used in the development of taxonomies of other types of accounting and audit objects. These may include the taxonomy of controls, risks, or audit methods.

One of the limitations of this study is the use of 10K filings from only Fortune 500 companies in the generation of the taxonomy. As a result the taxonomy might be a representation of the trends observed only in these companies, rather than representing a more varied and larger cross section of companies. Also, arguments could be advanced against using data directly from the filings. Such practice may lead to the creation of a taxonomy which is a reflection of the way companies report, rather than a true representation of the reporting standards as they are intended to be. To address this issue the knowledge base was built using both empirical data as well as terms appearing in the FAS statements. This is a very extensive process involving manually going through the log of unmapped items and creating tags based on that evaluation process. Future research maybe carried out that addresses some of these issues by using a larger data corpus. Also, using filings of a variety companies which encompass a range of sizes, as well as increasing the sample size could prove to be far more effective. Furthermore, the proposed method is a combination of automatic and manual processes. Human intervention could not be ruled out completely. Taking up the challenge of completely replacing human intervention, and designing an entirely automated process should also be a prime motivation for future research.

In order to successfully design and progressively create taxonomies, it is necessary to revisit some of the steps that were involved and are shown in Fig. 3. The gray colored boxes indicate manual processes. “Refer Standards” is a step where a human expert has to read through SFAS 87,158,132R(a), understand which line items are the relevant and required, and then include them in the list of terms to be used for building the taxonomy. It is probably impossible to completely automate this step, because human intelligence and domain expertise is required to assess which line items should be a part of the knowledge base. However, if an adroit system could be built with the capability of parsing text from the SFAS guidelines, then this step could be completely automated by

extracting the specific line items and saving them in a relational database for future reference,.

If future research were to be conducted into fully automating the process of tag creation, then an expert system would have to be built. However, such a process would be extremely difficult as well as time consuming. The chances and magnitude of any benefits being gleaned could be outweighed by the costs involved.

Starting the creation of the knowledge base by beginning with a list of terms acquired from GAAP or other SFAS guidelines could help prevent the omission of any required items. Questions could be raised regarding the use of only one hundred and twenty of the 10K filings from a list of the Fortune 500. This study consists of automatic as well as manual steps. Manual steps are especially involved when analyzing unmatched terms that couldn't be tagged, or while building the knowledge base, amongst other things. If the number of 10K statements used is increased many fold, it would increase the work required for the manual portion of the process. This could make the entire process more time-consuming and can complicate things further. The scope of this study is to propose and demonstrate a new Hybrid method for taxonomy. Creation of a comprehensive taxonomy for practical use would, however, require including the complete data available. Also it would require consideration as to the inclusion of data over a period of several years. It can be said that due to several changes in the FAS requirements related to pension disclosure, it would suffice to just use data released after the changes rather than using data from before changes. This eventually limits the amount of data range that is practically available for use. Using data over a very long period could lead to the inclusion of some terms that might be outdated and no longer in use. If data over a prolonged period is to be used, then excluding archaic terms may require a very complicated filtering system to be built. Moreover, while designing such a filter, one can never be sure about which kind of outdated terms could be expected to pop back up. Hence, this becomes an increasingly difficult problem. After taking all these factors into consideration, the data corpus was not expanded to include data over a longer period of time.

Apart from concerns about the size of the data corpus, it is important to acknowledge that the text analytic methods used here could be refined further.

Regular expression matching has been used for data extraction. However, to further improve the information extraction process, techniques such as Latent Semantic Analysis (LSA), as well as weighing of the terms and factors, could be used. Currently Principal Component Analysis (PCA) has been applied to the knowledge base for the selection of terms and phrases while building this taxonomy. However, methods like Singular Value Decomposition (SVD), using LSA and weighing of terms, could potentially improve the process further and should to be considered in future research. One criticism could be why a mathematical distance function was not used to compare the taxonomies. This would be a very complicated process, and deserves a separate study. Such work is outside the scope of this paper since the primary objective of this paper is to suggest a method for a hybrid taxonomy creation process.

6. REFERENCES

BEATTIE, V.; MCINNES, B.; & FEARNLEY, S., (2004): "A methodology for analyzing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes", Accounting Forum, vol. 28, n. 3: 205–236.

BOVEE, M.; KOGAN, A.; SRIVASTAVA, R.P.; VASARHELYI, M.A.; NELSON, K.M. (2005): "Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL)", Journal of Information Systems, vol. 19, n. 1: 19-41.

BOVEE, M.; ETTREDGE, M.; SRIVASTAVA, R.P.; VASARHELYI, M.A. (2002): "Does the Year 2000 XBRL Taxonomy Accommodate Current Business Financial Reporting Practice?", Journal of Information Systems, vol. 16, n. 2: 165-182.

CECCHINI, M.; AYTUG, H.; KOEHLER, G.J.; PATHAK, P. (2010): "Making words work. Using financial text as a predictor of financial events", Decison Support Systems, vol. 50, n. 1: 164-175.

CHAKRABORTY, V.; CHIU, V.; VASARHELYI, M. (2014): "Automatic classification of accounting literature", International Journal of Accounting Information Systems, vol. 15: 122–148

CHEN, H.; LYNCH, K.J. (1992): "Automatic Construction of Networks of Concepts Characterizing Document Databases". IEEE Transactions on Systems, Man, and Cybernetics, vol. 22, n. 5: 885-902.

CHEN, H.; SCHATZ, B.; MARTINEZ, J.; NG, T.D. (1994): "Generating a Domain-Specific Thesaurus Automatically: An Experiment on FlyBase" (Working Paper MAI-WPS 94-02): Center for Management of Information College of Business and Public Administration, University of Arizona.

CHUANG, S.; CHIEN, L. (2005): "Taxonomy Generation for Text Segments: A Practical Web-Based Approach", ACM Transactions on Information Systems, vol. 23, n. 4: 363–396.

CHUANG, S.; CHIEN, L. (2002): "Towards automatic generation of query taxonomy: a hierarchical query clustering approach", ICDM 2003. Proceedings. 2002 IEEE International Conference on Data Mining.

CLUTO, http://glaros.dtc.umn.edu/gkhome/views/cluto

CROUCH, C. J. (1990): "An Approach to the Automatic Construction of Global Thesauri", Information Processing and Management, vol. 26, n. 5: 629-640.

FASB pension accounting Summary of Statement No. 132 (revised 2003), http://www.fasb.org/st/summary/stsum132r.shtml

FASB Statement No. 87 Employers' Accounting for pensions, http://www.fasb.org/st/summary/stsum87.shtml

NYSSCPA, http://www.nysscpa.org/cpajournal/2005/1005/essentials/p28.htm

FRBSF, http://www.frbsf.org/publications/economics/letter/2003/el2003-19.html

FELDMAN, R.; GOVINDARAJ, S.; LIVNAT, J.; SEGAL, B. (2010): "Management's tone change, post earnings announcement drift and accruals", Rev. Accounting Stud., vol. 15, n. 4:915–953.

FISHER, I. E. (2004): "On the structure of financial accounting standards to support digital representation, storage and retrieval", Journal of Emerging Technologies in Accounting, vol. 1: 23 – 40.

FISHER, I. E.; GARNSEY, M.R. (2006): "The semantics of change as revealed through an examination of financial accounting standards amendments", Journal of Emerging Technologies in Accounting, vol. 3, n. 1: 41–59.

FISHER, I.E.; MCEWEN, R.A. (2009): "On a logical structure for the authoritative accounting literature: A discussion of the FASB's codification project". Issues in Innovation, vol. 3, n. 1: 32–56.

FISHER I.E.; GARNSEY, M.R. (2010): "The Role of Text Analytics and Information Retrieval in the Accounting Domain", Journal of Emerging Technologies in Accounting, vol. 7: 1-24.

FISHER, I. E.; GARNSEY, M.R. (2010): "Improving information retrieval from accounting documents: A prototype digital thesaurus for employee benefits", Paper presented at the 2010 AAA Midyear Meeting of the Information Systems and the Strategic and Emerging Technologies Sections (January), Clearwater, FL.

FRAZIER, K. B.; INGRAM, R. W.; TENNYSON, B. M. (1984): "A methodology for the analysis of narrative accounting disclosures", Journal of Accounting Research, vol. 22, n. 1: 318–331.

EDGAR ONLINE, http://www.edgar-online.com/

GANGOLLY, J. (1995): "Some thoughts on the engineering of Financial Accounting standards", Artificial Intelligence in Accounting and Auditing

GANGOLLY, J.; TAM, K. (2002): "On lexical acquisition for the financial reporting domain: Preliminary results of the analysis of year 2000 EDGAR filings", In Proceedings of the Eleventh Annual Research Workshop on: Artificial Intelligence and Emerging Technologies in Accounting, Auditing and Tax. San Antonio, TX: AAA.

GARNSEY, M. R. (2006): "Automatic Classification of Financial Accounting Concepts", Journal of Emerging Technologies in Accounting, vol. 3: 21-39.

GERDES, J. (2003): "EDGAR-Analyzer: Automating the analysis of corporate data contained in the SEC's EDGAR database", Decision Support Systems, vol. 35, n. 1: 7-29.

GOEL, S. (2008): "Qualitative information in annual reports and the detection of corporate fraud: A natural language processing perspective", Doctoral dissertation, University at Albany–SUNY.

GOEL, S.; GANGOLLY, J. (2009): "Can linguistic predictors detect fraudulent financial filings?", Paper presented at the Eighteenth Annual Research Workshop

on Strategic and Emerging Technologies Section of the AAA ,August 2009. New York, NY.

GRANT, C. H.; CONLON, S. J. (2006): "EDGAR extraction system: An automated approach to analyze employee stock option disclosures", Journal of Information Systems, vol. 20, n. 2: 119–142.

GRAY, G. L.; DEBRECENY, R. S. (2014): "A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits", International Journal of Accounting Information Systems, vol. 15, n.4: 357–380.

GRIMMER, J.; KING, G. (2011): http://www.pnas.org/content/108/7/2643.full.pdf

HAN, J.; KAMBER, M. (2006): "Data Mining: Concepts and Techniques", 2nd ed; The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March. ISBN 1-55860-901-6

HANAGEU, M.; LIEBMANN, M.; NEUMANN, D. (2012): "Automated news reading: Stock price prediction based on financial news using context-specific features", System Science (HICSS), 2012 45th Hawaii International Conference on IEEE.

HUANG KE-WEI; LI, Z. L. (2011): "A Multi-Label Text Classification Algorithm for Labeling Risk Factors in SEC Form 10-K. ACM Trans" On Management Information Systems, vol. 6, http://doi.acm.org/10.1145/1290002.129 0003

HUMPHERYS, S.; MOFFIT, K.; BURNS, M.; BURGOON, J., FELIX, W. (2011): "Identification of fraudulent financial statements using linguistic credibility analysis", Decision Support Systems, vol. 50, n. 3:585–94.

INGRAM, R. W.; FRAZIER, K. B. (1980): "Environmental performance and corporate disclosure", Journal of Accounting Research, vol. 18, n. 2: 614–622.

INGRAM, R. W.; FRAZIER, K. B. (1983): "Narrative disclosure in annual reports", Journal of Business Research, vol. 11, n. 1: 49–60.

JACOBS, P. S.; RAU, L. F. (1990): **"**SCISOR: extracting information from on-line news", Communications of the ACM 33, November: 88 - 97

KATRIEL, R.; RAFSKY, L. (1997): "Weight-Rate: A neoclassical approach to text categorization" ACM SIGIR '97 Conference

KRAVET, T.; MUSLU, V. (2011): "Textual risk disclosures and investors' risk perceptions", Rev. Accounting Stud., vol. 18, n. 4:1088–1122.

KOTHARI, S. P.; LI, X.; SHORT, J.E. (2009): "The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis", The Accounting Review, vol. 84, n. 5:1639–1670.

LEINNEMANN, C.; SCHLOTTMANN, F.; SEESE, D.; STUEMPERT, T. (2001): "Automatic Extraction and Analysis of Financial, Data from the EDGAR database", South African Journal of Information Management, vol. 3, n. 2 (preliminary version published in the Proceedings of Web Applications 2000 Johannesburg http://generalupdate.rau.ac.za/infosci/conf/thursday/Leinnemann. htm

MARIA, N.; SILVA, M.J. (2000): "Theme-based retrieval of Web News", Proceedings of the Third International Workshop on the Web and Databases, (WebDB2000)

MUI, C.; MCCARTHY, W. E. (1987): "FSA: Applying AI techniques to the familiarization phase of financial decision making", IEEE Expert, vol. 2, n. 3: 33–41.

MULLER, A.; DORRE, J.; GERSTL, P.; SEIFFERT, R. (1999): "The TaxGen framework: automating the generation of a taxonomy for a large document collection", Proceedings of the 32nd Annual Hawaii International Conference on System Sciences, January

NELSON, K M.; KOGAN, A.; SRIVASTAVA, R. P.; VASARHELYI, M. A.; LU, H. (2000): "Virtual auditing agents: The EDGAR agent challenge", Decision Support Systems, vol. 28, n. 3: 241-253.

O'LEARY, D.; EIS, T. (1991): "A knowledge-based system for cash management: With implications for structuring DSS and with extensions for system learning and knowledge acquisition", In Advances in Working Capital Management, vol. 2, edited by Kim, Y. H., 197–209. Stamford. CT: JAI Press.

O'LEARY, D.; KANDELIN, N. (1992): ACCOUNTANT: "A domain-dependent accounting language processing system", In Expert Systems in Finance, edited by

O'Leary, D. E., and P. R. Watkins, 253–267. Amsterdam, The Netherlands: Elsevier Science Publishers.

PIECHOCKI, M.; FELDEN, C.; GRANING, A.; DEBRECENY, R. (2009): "Design and standardization of XBRL solutions for governance and transparency", International Journal of disclosure and governance, vol. 6, n. 3: 224-240.

ROGERS, J.; VAN BUSKIRK, A.; ZECHMAN, S. (2011): "Disclosure tone and shareholder litigation", The Accounting Review, vol. 86, n. 6: 2155–2183.

ROUTEN, T.; BENCH-CAPON, T.J.M. (1991): "Hierarchical formalizations", International Journal of Man-Machine Studies, vol. 35: 69-93

SALTON, G. (1989) .Automatic Text Processing. Addison-Wesley Publishing Company, Inc., Reading, MA

SCHNEIDER, J. W. (2005): "Verification of bibliometric methods' applicability for thesaurus construction", ACM SIGIR Forum 39: 1.

TAN, P.N.; STEINBACH, M.; KUMAR, V. (2005): "Introduction to data mining", Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

TENNYSON, B. M.; INGRAM, R. W.; DUGAN, M. T. (1990): "Assessing the information content of narrative disclosures in explaining bankruptcy", Journal of Business Finance & Accounting, vol. 17, n. 3: 390–410.

VASARHELYI, M.; PENG, J. (1999): "Qualitative corporate dashboards for corporate monitoring", Information Systems Audit and Control Journal, vol. 5:45-48

VASARHELYI, M.; CHAN, D.; KRAHEL, D. (2011): http://eycarat. faculty.ku.edu/myssi/_pdf/4-Miklos-Chan-Alles-XBRL-Consequences.pdf 2011

WAYMIRE, G. (1985): "Additional Evidence on the accuracy of analyst forecasts before and after voluntary management earnings forecasts", The Accounting Review, vol. 61, n. 1: 129-142.

WU, Y.F.; GANGOLLY (2000): On the Automatic Classification of Accounting concepts: Preliminary Results of the Statistical Analysis of Term-Document Frequencies.

XBRL, Xbrl org, http://xbrl.org/FRTApproved/http://xbrl.us/preparersguide

YANG, B.; ANINDYA, D. (2014): "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures", Management Science, vol. 60, n. 6:1371-1391, http://dx.doi.org/10.1287/mnsc.2014.1930

ZHENG, Z. (2001): International News Connection: A Real-time Online News Filtering and Classification System. ACM SIGIR'01 Workshop on Mathematical/Formal Methods in IR.

Appendix I

Section	No. of Synonyms
Pension header	25
Amounts recognized in the Consolidated Statement of Financial Position captions include or amounts recorded in our consolidated balance sheets	5
Benefit payments	3
Change in benefit obligations	7
Change in fair value of plan assets	5
Cost/(income) of pension plans	4
Funded status	3
Information for pension plans with an accumulated benefit obligation	1
Information for pension plans with an accumulated benefit obligation in excess of plan assets	5
Weighted average actuarial assumptions	8
Weighted average assumptions used to determine projected benefit obligations	4
Weighted-Average Asset allocation	6
Weighted-average asset allocation of the pension and postretirement plans	3
Investment Policies and Strategic Narrative Description	3
Assets, Target Allocations	2
Unfunded Plan	4
Accumulated other comprehensive income, before tax	5
Amounts Amortized from Accumulated Other Comprehensive Income (Loss) in next Fiscal year	3
Pension plans with a benefit obligation in excess of plan assets	5
Alternative Methods to Amortize Prior Service Amounts	2
Alternative Methods to Amortize net gains and losses	3
Method to Determine Vested Benefit Obligation	1
Special Termination Benefits	5
Plan Amendment	6
Settlement and Curtailments	7
Measurement Date	2
Pension plans with a accumulated benefit obligation in excess of plan assets	7
Additional Disclosures about Plan assets	3
Type of Employer and Related Party Securities Included in Plan assets	5
Amount of Employer and Related Party Securities Included in Plan assets	6
Number of shares of Equity Securities Issued by Employer and Related Parties Included in Plan assets	8

List of sections and number of synonyms for each section

Appendix II

Section Details	No. of Synonyms
Accumulated benefit obligation	3
Accumulated other comprehensive income	5
Actual return on plan assets	4
Actuarial losses (gains)	3
Additional Disclosures about Plan Assets	1
Aggregate Accumulated benefit obligation	5
Aggregate Benefit Obligation	3
Aggregate Fair value of Plan assets	4
Aggregate Projected Benefit Obligation	6
Amortization of actuarial (gain) loss	3
Amortization of Gains(Losses)	2
Amortization of net gains(losses)	2
Amortization of net Prior service cost(credit)	2
amortization of net prior service cost(credit) before tax	3
Amortization of net Transition Asset(Obligation)	3
Amortization of prior service cost	4
Amortization of transition assets	4
Amount of Employer and Related Party Securities Included in Plan assets	3
Benefit obligations, beginning of year	3
Benefit obligations, end of year	5
Benefit payments	3
Benefit related liabilities	2
Business Combinations and Acquisitions, Plan assets	4
Business Combinations and Acquisitions, Benefit Obligation	2
Cash Flows	3
Cost of other defined benefit plans	1
Cost of providing Special termination benefits	5
Curtailment losses (gains)	3
Debt securities	1
Derivatives Use	4
Description of Event Resulting in Special or Contractual Termination benefits recognized during period	2
Discount rate	2
Diversification	2
Divestitures	3

Details of items under each section