Automating the process of taxonomy creation and comparison of taxonomy structures

Vasundhara Chakraborty, Rutgers, the State University of New Jersey, Newark.,NJ, vasuchau@gmail.com

Miklos Vasarhelyi, , Rutgers, the State University of New Jersey, Newark.,NJ, miklosv@andromeda.rutgers.edu

Abstract: The ability to automatically extract information from the footnotes of financial statements simplifies access to critical information concerning public companies. However, extraction can be particularly challenging due to great variations in the filing structure and terminologies used. Hierarchical formalization of text becomes a necessity in such circumstances. This is facilitated by the creation of a valid taxonomy. The objectives of this paper are threefold: (1) to develop a semiautomatic method of taxonomy creation; (2) to compare the structure of the taxonomy created with the XBRL US GAAP taxonomy and (3) to demonstrate how the tool developed as a part of this process can be used for more exploratory research. Pension plan footnotes of 10K statements have been used to demonstrate the use of this method. To create a taxonomy, we first collected 10K statements from SEC EDGAR (Electronic Data Gathering, Analysis and Retrieval) then extracted pension footnotes, and restructured the data. We then applied the Hierarchical clustering algorithm to this data to create the taxonomy structure. Comparison of the taxonomy developed with the XBRL taxonomy reveals some differences. In general the reporting trends of companies reveal a greater level of aggregation. Pension footnote structures of forty five randomly selected companies were compared across ten years. Several instances were found where the company has added new terms or a completely new section to the footnote or a term is missing.

The contribution of this paper are: (i) a method to formalize and partially automate the complex and time-consuming process of taxonomy creation using historical data has been proposed; (ii) a generic parsing tool as a part of the taxonomy creation process is developed; (iii) structural differences between the official XBRL US GAAP taxonomy and taxonomy using historical data are demonstrated (iv) potential use of the parsing and matching tool for other exploratory research in accounting is shown.

I. Introduction and Background

Overview

Credible and timely financial information is very important for market participants. Since all content cannot fit into the accounting framework, there is a need for companies to be able to disclose other relevant information. The form of disclosure needs to be the same across firms and should be such that it allows and encourages investors to process it.

Extracting useful information is particularly important for analysts and investors but it has not been easy to do so. As a result of Sarbanes-Oxley, Securities and Exchange Commission actions, and recent shareholder suits, some firms have decided not to give earnings guidance. Thus, for many investors, receiving good independent analysis about potential investments has become a problem. However a recent requirement for mandatory XBRL filing by companies is a step in this direction. Development of tags is an indispensable part of this process. In this paper, we develop a detailed methodology to generate taxonomies semi-automatically and also explore possibilities of applying this semantic parsing tool for other research.

Motivation

Financial statements submitted to the SEC are a necessary source of company information for a wide variety of users including company management, investors, creditors, governmental oversight agencies, and the IRS. The footnotes expand on the quantitative 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 clarity or conciseness that are followed in the footnotes.

Hierarchical formalization of text is necessary to find relevant information from the vast amount of data that is available. This necessitates the generation of a comprehensive taxonomy. A very promising approach to solving this problem is the use of tagged data, which enables producers and consumers of financial information to switch resources away from costly manual processes, typically involving time-consuming comparison, assembly and re-entry of data.

In February 2005, the SEC announced that it will start accepting voluntary filings in XBRL format. In May 2008 SEC voted for mandatory XBRL filing and suggested that companies begin filing financial statements in an interactive tagged format for fiscal periods ending in late 2008. Although recent XBRL filings allow retrieval of line items from financial statements, extraction of specific items from footnotes is still not common. XBRL filings require the use of tags which come from XBRL taxonomies. As of today these taxonomies are being created manually.

Because filers have considerable flexibility in how financial information is reported under U.S. reporting standards, it is possible that a company may wish to use a non-standard financial statement line item that is not included in the standard list of tags. As a result filing companies create a company-specific element called an extension. For example, a company may use the caption “Estimated Future Benefit Payments” and may look for it in the standard label list. What is available is “Benefit payments”. However, the use of non-standard company specific extensions may introduce errors due to reduced comparability of data. As per recommendations from the SEC, use of company-specific extensions should be very limited. Still, filers sometimes continue to use familiar labels instead of adopting a standard. One solution to this issue is the use of historical data to create a comprehensive taxonomy which can prevent creation of extension taxonomies.

Bovee (2005) proposes that there should be an empirical approach towards the evaluation and improvement of XBRL taxonomies so that the taxonomy matches with the historical data. Vasarhelyi (2002) argues that a poor fit may lead to information loss and to subsequent resistance to adoption of the taxonomy. Bovee (2002) points out that although it is expected that XBRL will improve comparability and consistency of financial reporting and may facilitate near-continuous reporting, some questions arise regarding how well the proposed taxonomy corresponds to firms' preferred reporting practices. Bovee (2005, 2002) suggests that there is a need to empirically evaluate taxonomies.

Discussion with a subject matter expert in the taxonomy creation process reveals that 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 to develop XBRL tags. This team reviewed all the regulations in US GAAP as a part of the taxonomy development process. The preliminary taxonomy is then sent to the FASB for an initial review. Each element that was created was then reviewed by a subject matter expert at FASB. Simultaneously, the SEC did a similar review. They engaged corporate finance experts 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. Testing was done by mapping real financial statements to the taxonomy. Line items that could not be mapped were added as additional elements. Once the taxonomy was released in 2007 there was a period of review before the voluntary filing program began. It was found after the voluntary filings that companies used 30% to 40% extension tags. As a result some additional tags were added. A diagrammatic representation of the taxonomy creation process by XBRL US is shown in Fig1.

As regulations change, firms must report differently. Therefore, taxonomies would also need to be updated often. This process could be very complicated and time consuming if done manually. We propose a methodology which could at least partially automate the process of using historical data for taxonomy creation. And this leads us to the first research question where we ask:

RQ1: What method should be used to create a taxonomy automatically using historical data from financial statements?

Merely creating the taxonomy automatically is not a panacea. There is a need to assure that it is comprehensive enough to include all or most of the elements of the text. To determine whether a taxonomy created using this process is more effective than one created manually, we need to first observe whether any differences exist between the two. Thus we ask the second research question:

RQ2: What are the structural differences between the official XBRL pension footnote taxonomy and a taxonomy created by a semiautomatic method?

Once we determine what these differences are, we can measure which of the two taxonomies results in more accurate tagging of line items of pension footnotes. Thus we ask the third research question as follows:

RQ3: Is tagging of pension footnote data more effective using tags produced from the alternate method as compared to the tags from the official XBRL taxonomy?

Semantic parsing techniques have been used to extract data from pension footnotes of 10K filings of Fortune 500 companies.[1] Object filtering, automatic indexing, and cluster analysis are used to develop an automated taxonomy building process.

The rest of the paper is organized as follows: Section II discusses some basic concepts; Section III presents the literature review; Section IV discusses the methodology; Section V shows our results; Section VI explains the applications of the tool; Section VII discusses limitations and conclusion and Section VIII is the references.

II. Basic concepts

Semantic Parsing

Natural Language refers to the language spoken by people. Natural language processing (NLP) refers to applications that deal with natural language in one way or another. Information retrieval, machine translation and language analysis are a part of NLP. Semantic Parsing is a way to perform language analysis. In computer science and linguistics, parsing (more formally called *syntactic analysis*) is the process of analyzing text, made of a sequence of tokens (for example, words) to determine its grammatical structure.

Parsing splits a sequence of characters or values into smaller parts. It can be used for recognizing characters or values that occur in a specific order. In addition to providing a powerful, readable, and maintainable approach to regular expression pattern matching, parsing enables a user to create custom languages for specific purposes. A simple form of parsing is for splitting strings.

The parse function splits the input argument, or string, into a block of multiple strings, breaking each string wherever it encounters a delimiter, such as a space, tab, new line, comma, or semicolon. For example:

"It is going to rain today" would be parsed into individual words as follows:

"It" "is" "going" "to" "rain" "today"

Regular expressions can be used to match strings when they occur in specific sequences. This allows a user to extract any particular string or specific patterns.

[1] http://www.sec.gov/idea/searchidea/webusers.htm

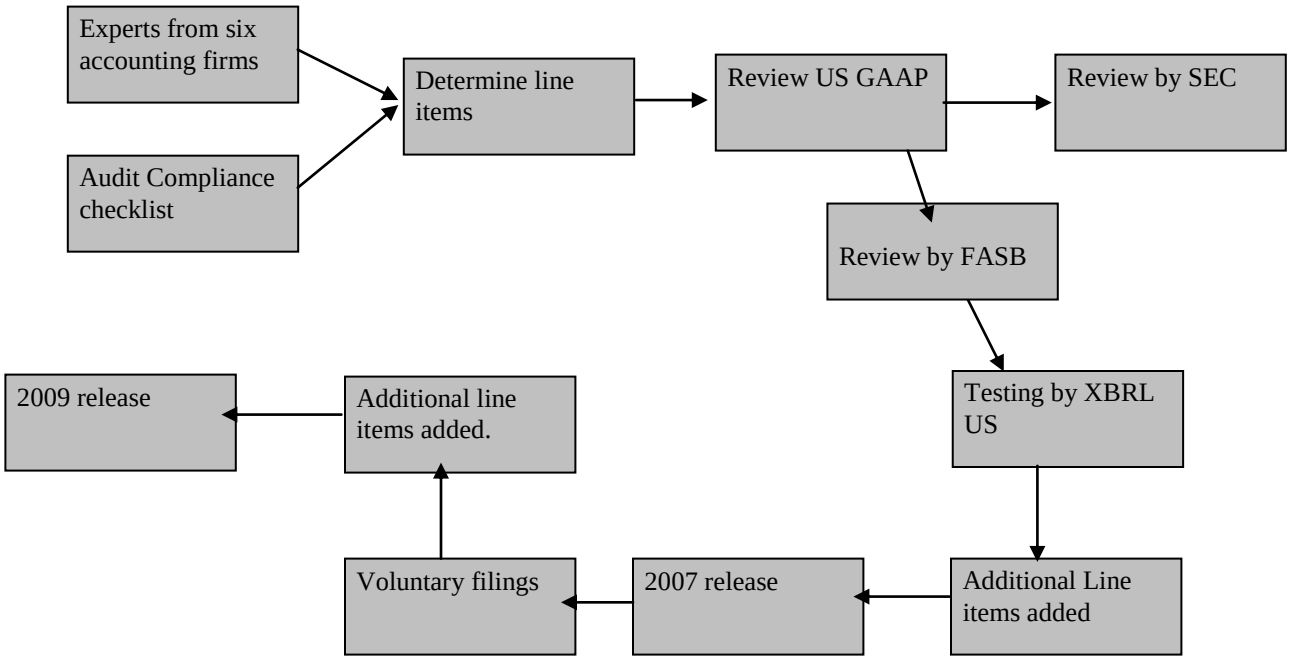

Fig. 1 Taxonomy generation process followed by XBRL US

Object filtering:

Bates (1986) proposed a design model for subject access in online catalogs and stressed the importance of building domain-specific lexicons for online retrieval purposes. A domain- specific, controlled list of keywords helps to identify legitimate search vocabularies and help searchers ``dock'' on to the retrieval system (Chen 1992). Therefore, object filtering is the first step in the preparation of a knowledge base and deals with the preparation of a domain-specific set of keywords. For the purpose of taxonomy creation for pension footnotes, a formal list of domain-specific keywords is not available. As a result, guidelines from SFAS 87 and SFAS 158 were studied to develop a group of relevant terms.

Automatic Indexing:

Object filtering creates a preliminary list of words to be used as a part of the knowledge base. Automatic indexing is the next step where the data corpus is scanned to identify additional words or phrases that are conceptually related. These new terms are then added on to the existing knowledge base. This is a method which has been widely used in the field of information science (Chen 1992). Salton (1989) presents a description of automatic indexing, which includes various steps such as dictionary look-up, stop-wording, word stemming, and term-phrase formation. The algorithm first identifies individual words, and then uses stop words to remove non-semantic bearing words such as 'the,' 'a,' 'on,' 'in,' 'and,' etc.

Cluster analysis:

Cluster Analysis, also called data segmentation, deals with grouping or segmenting a collection of objects (also called observations, individuals, cases, or data rows) into subsets or "clusters". Objects within a cluster are closer to each other than objects assigned to different clusters. The 'closeness' depends on the degree of similarity (or dissimilarity) between the individual objects being clustered. For example consider that there are one thousand documents related to sports and we don't know which document belongs to which sport (eg. soccer, tennis and cricket). Cluster analysis could be used in such a case to find similarities between documents based on the words or phrases being used and group them together into different clusters for each kind of sport.

There are different types of clustering algorithms. In this paper hierarchical clustering has been used. In this type of clustering the data is partitioned in a series of steps. Hierarchical Clustering includes agglomerative methods, which is a bottom up approach, and divisive methods, which is a top-down approach. It can be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis. An example of such a dendrogram is given in Fig. 2

Example of Agglomerative hierarchical clustering

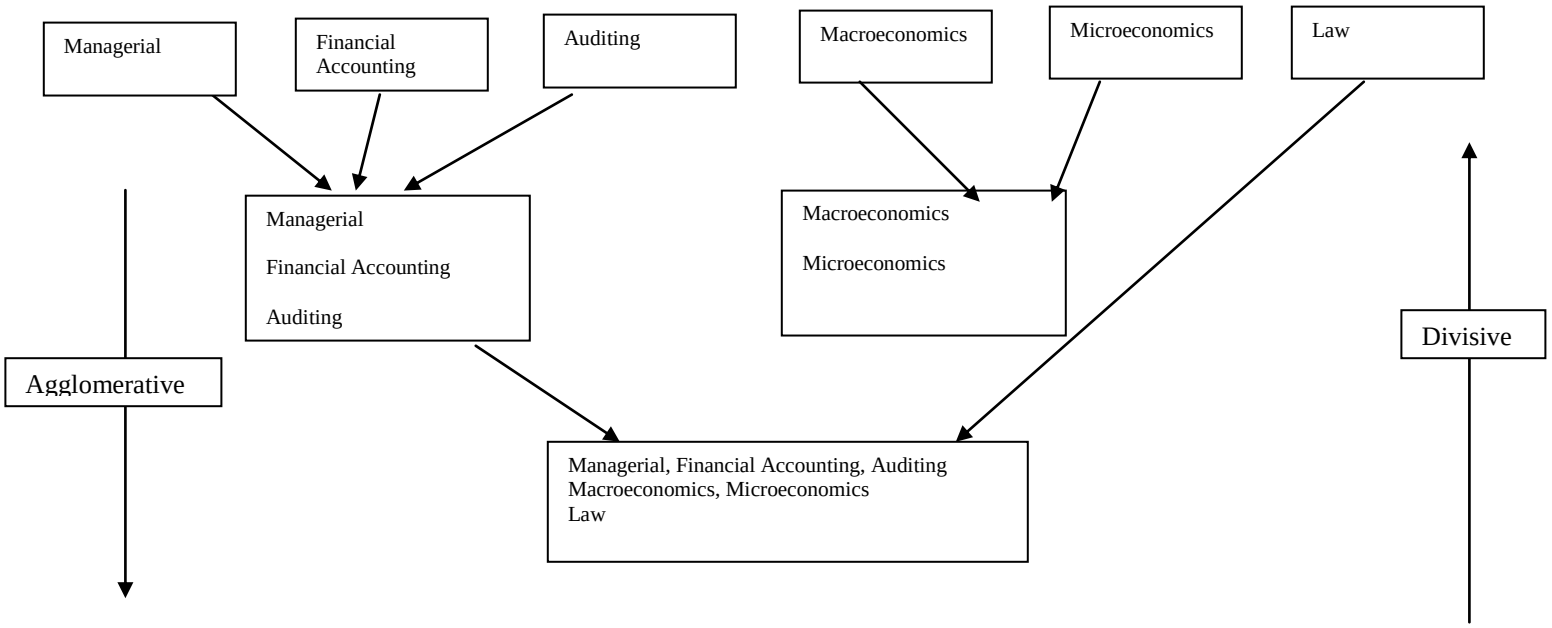

Fig. 2 Hierarchical clustering dendrogram

Consider that we want to cluster various documents using the agglomerative method. In the first step all the documents are separate leaf nodes. First all the documents in the area of managerial accounting, financial accounting, auditing, microeconomics, macroeconomics and law are clustered into respective groups. In the next step all the documents in the domain of managerial accounting, financial accounting and auditing are clustered into a single group being under the common umbrella of accounting. Similarly documents in the field of microeconomics and macroeconomics are clustered together being in the field of economics. Law is a separate cluster by itself. Finally all the documents are clustered together under the social science domain.

Knowledge Base:

Generally speaking, a knowledge base is a centralized repository of information. It is a database which contains information related to a common subject. An example of a commonly used knowledge base is a library.

A knowledge base is used to build information, organize contents and optimize data retrieval for users of the information. It is not a static collection of information, but a dynamic resource that may itself have the capability to learn and expand itself.

In this paper, the knowledge base consists of pension footnote data stored in an Access database. To build this knowledge base, we began by creating a list of terms expected to be disclosed in the pension footnote as per SFAS 87 and SFAS 158 guidelines. We then added all the line items parsed out from the pension footnotes. The knowledge base evolved further as terms were added with the parsing of each new 10K. The addition of synonyms of terms and phrases expanded the knowledge base further.

Document-term Matrix:

A document-term matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

Fig. 3 is an example of a document-term matrix. The first column "File" is the name of the document. The column headings from the second column onward show the words that occur in each document. Each cell in the matrix indicates how many times a certain word occurred in a document. For example the word pension occurs 6 times in the AT&T document.

File	Pension	Retirement	Service Cost	Benefits
AT&T	6	2	2	1
B&N	3	0	2	2
Dell	4	0	2	2
Costco	0	3	2	1
J&J	3	0	1	2
Ford	0	4	2	2
GM	7	1	3	2
GE	4	1	2	3

Fig.3. Example of Document-term Matrix

Markup Language and use of tags:

A markup language is a collection of codes used to indicate the structure and format of the data presented. Hyper Text Markup Language (HTML) and XML (eXtensible Markup Language) are examples of two very popular markup languages. HTML is designed to display data, focusing on appearance. XML, on the other hand, is designed to describe data, focusing on content and context. eXtensible Business Reporting Language (XBRL) is an extension of XML. XBRL is a derivative language of XML which is used for the electronic communication of business and financial data used for business reporting around the world. It started in 1999 as an investigation to use eXtensible Markup

Language (XML, http://www.W3C.org.) for the electronic reporting of financial information and is technically based on the eXtensible Markup Language (XML).

Using tags simplifies the data sharing process. Different applications can access data from the same source irrespective of the environment.

9.PENSION, POSTRETIREMENT AND OTHER EMPLOYEE BENEFIT PLANS
Note 21—Employee Benefit Plans
pension and Postretirement Plans
Note 6—Retirement Plans
Note 11. Employee Benefit Plans • *Defined Benefit pension Plans and Postretirement Plans* (as a separate section under Employee Benefit plans)
w. Retirement-Related Benefits
EMPLOYEE BENEFIT PLANS
Defined Contribution pension Plans
Defined Benefit pension Plans
Albertson's defined benefit Plan
Shaw's defined benefit plan
14. Employee Benefit Plans • *Defined Benefit pension Plan* • *Defined Contribution Plans* • *Postretirement Benefits*

Fig.4 Examples of the different ways in which a pension heading can be represented

This kind of tagging can be very useful when companies tend to use different terminologies to represent the same terms. Fig. 5 shows some examples of different terms (referred to as synonyms) used by companies to represent the same line items. Fig. 4 shows examples of synonyms used by different companies to define the pension footnote headings. Appendix I shows a list of line items for which several synonyms have been found to exist in the 10K filings that were studied.

Different ways of representing the same line items

Line item	Synonym1	Synonym2
Cost/(income) of pension plans	Components of the net periodic benefit cost	Cost of the Company's deferred benefits and postretirement benefit plans
Weighted average assumptions used to determine projected benefit obligations	Weighted-average assumptions used to determine benefit obligations	Weighted-average assumptions used to determine net periodic benefit cost
Weighted-Average Asset allocation	defined benefit pension plan asset allocation	weighted-average target allocations and actual asset allocations

Fig. 5 Examples of synonyms of section headers in pension footnote paragraphs

Extracting accounting information from financial statements

There is abundant electronic data regarding companies' financial performance available to users today. While automatic analysis of financial figures is common, it has been difficult to extract meaning from the textual portions of financial reports automatically. Terms may not be used consistently, or there may be errors or inconsistencies which make automation and analysis difficult. Some examples in the literature which try to address these problem include EdgarScan[2] which extracts financial tables from SEC filings using semantic parsing technology. EDGAR agent (Nelson

[2] 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.

2000) is capable of retrieving quarterly filings (10-Qs) from the SEC EDGAR. However, it can only process quarterly SEC filings, identify only a few of the most important accounting numbers, and interact only with a single online information source (the SEC EDGAR repository). Edgar analyzer (Gerdes 2003) is not successful in retrieving particular pieces of information, instead extracts whole paragraphs when keywords appear.

The need for hierarchical information

The importance of storing data hierarchically for easier extraction of information has been mentioned in the literature. It is as important to capture and control the knowledge base underlying accounting decision making 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 the XML DTD[3] allows defining data. Similarly, Routen and Bench-Capon (1991) argue that hierarchical formalization of text is preferable for the creation of knowledge base systems. Gangolly (1995) suggests hierarchical structuring of accounting standards based on three points: 1) distinguishing changes in standards from the original. 2) maintaining meta-level information and date stamps of such changes. 3) separation of other meta-level information.

III. Prior research

Examples of various methods used:

Different approaches to automatically extract data from financial statements have been mentioned in the literature. Leinnemann (2001) introduced text mining techniques to implement Edgar2xml, a software agent which extracts fundamental company data from the EDGAR database of the United States SEC and outputs this data in a format which is useful to support stock market trading decisions Katriel (1997) describes a computer system called Coding Agent which is capable of assigning category codes to short text. Semantic parsing and text mining techniques have been used for the Coding Agent.

Chen (1992) describes the steps involved in generating a thesaurus automatically. Chen (1995) then uses this concept to automatically generate a thesaurus and evaluate it for the worm community system(WCS). Chen (1994) created a thesaurus for Drosophilia information. Chuang (2005) describes taxonomy generation for text segments using a hierarchical algorithm. In this paper we use semantic parsing together with object filtering and automatic indexing to extract data from 10K statements and apply a hierarchical agglomerative algorithm to generate the taxonomy automatically.

[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.

Garnsey (2006) uses semantic parsing techniques to determine the feasibility of using statistical methods to automatically group related accounting concepts together. Peng and Vasarhelyi (1999) expand the traditional financial audit framework and argue in favor of the scope of evidence collection to cover on-line corporate information (in particular news), using semantic analysis methods.

.

One instance where semantic analysis is considered very useful is extraction of information from financial statements available from intermediaries. Bovee (2005) implements intelligent parsing to extract accounting numbers from financial statements available from the SEC EDGAR repository in project FRAANK. It matches the line-item labels to synonyms of tags in XBRL taxonomy to actual numbers using synonyms and helps to convert consolidated balance sheets, income statements, and statements of cash flows into XBRL-tagged format.

Wu & Gangolly (2000) study the feasibility of automatic classification of financial accounting concepts by statistical analysis of term frequencies in the financial accounting standards. The procedure uses principal-components analysis to reduce the dimensionality of the dataset, and then uses agglomerative nesting algorithm (AGNES) to derive clusters of concepts.

Salton (1989) presents a blueprint for automatic indexing, which typically includes dictionary look-up, stop-wording, word stemming, and term-phrase formation. Crouch (1988), and Crouch & Yang (1999) use a complete-link algorithm for automatic generation of a global thesaurus. Chuang & Chien (2005) describe taxonomy generation for text segments using a hierarchical algorithm.

Using the Hierarchical clustering algorithm

Among clustering methods, the hierarchical clustering algorithm is considered to be the most reliable and popular. There are many instances where it has been used successfully. Müller (1999) uses hierarchical clustering for automatic taxonomy generation for a large document collection. Chuang & Chien (2002) use hierarchical agglomerative clustering algorithm to hierarchically group similar queries and generate the cluster hierarchies by a cluster partition technique. Wu and Gangolly (2000) describe classification of accounting concepts and use principal component analysis to reduce the dimensionality of the dataset, and then use AGNES to derive clusters of concepts. In this paper empirical data is combined with data from Statements of Financial Standards to create a knowledge base that is used to create a taxonomy applying the hierarchical clustering algorithm.

In this paper we use intelligent parsing together with object filtering and automatic indexing to extract informative data from 10K statements and then apply hierarchical agglomerative algorithm to generate the taxonomy automatically.

IV. Methodology

The sample:

A data corpus of 10K filings submitted by public companies is collected. 120 companies were randomly selected from the Fortune 500.10K statements of these companies filed in 2007 were collected from the SEC Edgar database. Altogether, one hundred and twenty 10K statements from 120 different companies were manually downloaded from the SEC Edgar website. 80 were used as the training dataset. 40 were used as the test dataset.

Generating the pension taxonomy

Figure 6 is an overview of the steps taken to partially automate the process of pension footnote taxonomy creation. The first stage, data collection and restructuring, consists of downloading data, changing its structure, and doing a word count to determine frequency of line items. In the second stage, the taxonomy structure is created. In the third stage line items are mapped to terms in the taxonomy. Finally, line items that could not be mapped to terms in the taxonomy are analyzed. If they are new and frequently occurring terms, they are added to the terms database; otherwise, they are checked for synonymy. Three iterations of the four steps are carried out before we arrive at the results.

Fig. 7 gives a detailed explanation of the four steps discussed above. The steps shown within the dotted line are repetitive. Three iterations of these steps were completed before we arrived at the final result.

1. First we download the 10K statements manually from the SEC Edgar website.
2. Then we identify and parse the pension plan footnote from the 10Ks.
3. Then we identify each table in the pension footnote and make each table a document of its own. This is a data restructuring process. This step is necessary and facilitates the clustering process.
4. The individual line items are then extracted from each of these documents.

5. The line items and their synonyms retrieved from the 10K statements and then stored in the access database.

6. These line items are then included in the knowledge base.

7. Simultaneously the standards SFAS 87 and SFAS 158 are referred to prepare a list of terms that should be in the taxonomy. These terms are then added to the central database. This step begins the creation of the knowledge base.
8. The knowledge is used to create a document term matrix.

9. Hierarchical clustering algorithm is then applied to this document-term matrix.

10. An output of this process is the taxonomy structure. Fig. 11 shows how the data is clustered into different groups.

11. This structure is used to create tags.

12. The line items in the pension footnotes are then mapped to the tags.

13. The success rate of tagging is measured based on the number of line items that could not be mapped.

14. Finally the line items that could not be mapped are analyzed.

15. After the analysis any new relevant terms are added to the list..

Steps 6 to 15 are iterative steps. Three iterations of these steps are carried out to improve the comprehensiveness of the taxonomy. Some of the steps mentioned above are explained in details 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 87, a list of items was prepared which should ideally be included in the pension footnote taxonomy.

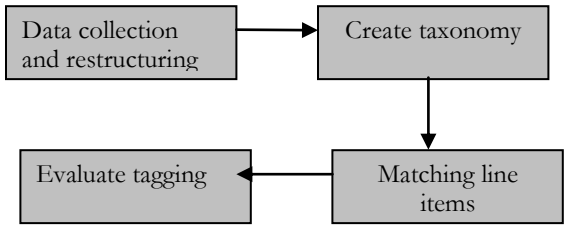

Fig.6 An overview of the process to generate taxonomy

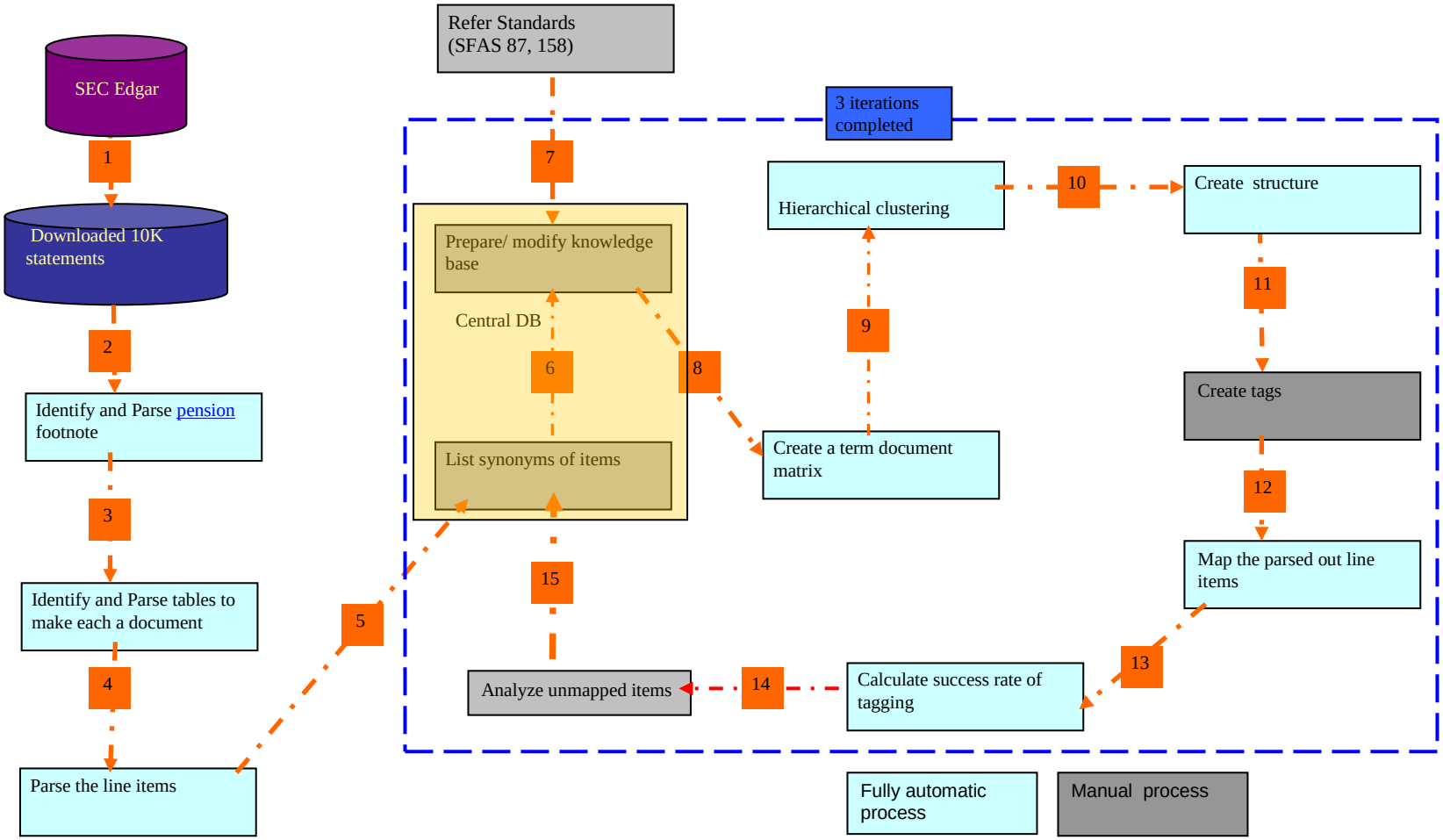

Fig.7 Details of the procedure for taxonomy creation

Automatic Indexing: After the 10K statements were filtered using the list, the remaining text was processed by the automatic indexing routine. The following steps are part of the automatic indexing procedure:

(a)Word identification: The prototype system first identified the remaining words (after object filtering) in each document. The system recognized words by monitoring space and everything was converted to upper case to avoid confusion.

(b)Stop-wording: To eliminate unwanted words, a combination of two stop words list is used. The function of a stop word list is to eliminate frequently occurring words that do not have any semantic bearing. The first stop word list contains 571 words and was built by Gerard Salton and Chris Buckley for the experimental SMART information retrieval system at Cornell University The second stop word list was obtained from the Onix Text Retrieval Toolkit.

Word stemming has not been used since in many cases it completely changes the context of the word, which is crucial for our purposes.

Developing the Parsing tool

Developing the parsing tool was a challenging process due to variations in the reporting styles of subject companies.

The first task was to identify the heading of the pension plan footnote. Analysis of the 10K documents showed that companies do not follow any standard for naming the headings of the pension plan footnotes, nor does any structure exist in the paragraphs.

After identifying the pension footnote paragraph, the detailed content including the tables and line items were parsed out. Semantic parsing using synonyms becomes necessary due to variations in reporting structures and variations in terminologies used by companies.

Restructuring the data format

Clustering permits grouping of documents of the same type. In the case of pension footnotes, once the pension disclosure is extracted from the 10K statements of the different companies they are all of the same type, posing a problem. Documents which are of the same type would automatically be clustered in one single group.

In order to tackle this problem, restructuring of the data was necessary. First the pension disclosure is extracted from the 10K statements. After that each of the tables in the pension disclosure are extracted and stored separately, so a single pension disclosure now becomes a set of approximately ten documents. This set of newly formed documents can now be more effectively clustered.

Creation of the document-term matrix:

A separate program has been developed to create the document-term matrix. The program does a word count to measure how many times a particular word or phrase occurs in a document. Following which the term frequency in the collection of documents is measured. A high term frequency indicates that a term is highly related to a particular document. Terms which appeared at least 20 times were included in the list of synonyms and subsequently added to the knowledge base. Terms identified through object filtering are usually more accurate than terms generated by automatic indexing because automatic indexing is a relatively "noisy" Process.

The document-term matrix which resulted from the training dataset has 80 rows (for the number of companies) and 1275 columns to represent the words and phrases occurring in the documents.

Creating a hierarchy of terms

Cluster analysis (Tan, Steinbach, Kumar 2005) is the starting point in the taxonomy creation process. The document-term matrix created in the previous step is an input to the hierarchical agglomerative clustering algorithm. The hierarchical clustering algorithm has been implemented using CLUTO. [4] It is a freely available software package used for clustering datasets and analyzing the characteristics of these datasets.

Creating the taxonomy

Hierarchical groups created automatically using the agglomerative algorithm create the basic structure of the taxonomy. However, this is by no means a comprehensive taxonomy, especially since such a hierarchical structure does not have any provisions to include Hypercubes and domains as explained in the requirements for XBRL Dimension 1.0. As a result, using the hierarchical structure generated by clustering as a guideline, a basic structure for the taxonomy has been created. When we compare the two taxonomies we take this into consideration and compare only their basic hierarchical structures.

Term matching in the test data set

The tagging process begins by first identifying all the elements, their names and the corresponding labels. The test dataset of forty 10K statements were then parsed. Finally the line items in the test dataset were matched with the tags created using the training dataset. A separate program has been developed as a part of this project to complete this matching process automatically.

Analyzing the unmapped items

After the first iteration of term matching was completed automatically, remaining unmatched items were stored in an Access database. The terms could be unmapped due to absence of a matching tag or absence of a matching synonym or inefficiency of the parsing program. Studying these unmapped items allows us to identify new synonyms for line items and make the tagging process more comprehensive. These new terms or phrases are then incorporated into the database and made part of the knowledge base, and a new iteration begins. Three iterations are completed before we arrive at the final result. The number of iterations is decided more on a general perception of the comprehensiveness of the knowledge base rather than any other recommendations.

[4] Family of Data Clustering Software Tools; http://glaros.dtc.umn.edu/gkhome/views/cluto

The prototype system:

The prototype system consists of several modules: (i) a semantic parsing tool for automatic extraction of data from the pension disclosures of 10K filings; (ii) modules for automatic indexing and object filtering; (iii) a document term matrix creation module which automatically creates a list of terms or phrases that occur in the 10K filings of companies as well as a count of the number of times that they occur; and (iv) a term matching program which is used to automatically tag parsed data from pension disclosures of 10K filings. Unmapped terms are stored in a central database for future analysis. All of these programs are written using Visual Basic

Generic use of the Tool:

We have attempted to make the design of the prototype tool generic so that data extraction can be done for exploratory research on other areas of 10K statements, especially footnotes. To facilitate this data is stored in a MS-Access database. The database consists of 12 tables.
Here we consider the example of the ClassFile table as shown in Fig.8. This table stores information regarding each section of a pension footnote. However adding a unique identifier like "Category" would allow its use even in cases other than pension footnotes.

	Field Name	Data Type
	IDENTIFICATION	Text
🔑	Category	Text
	ClassCode	AutoNumber
🔑	ClassHeading	Text
	Synonym1	Text
	Synonym2	Text
	Synonym3	Text
	Synonym4	Text
	Synonym5	Text

Fig. 8 ClassFile table stores information about each group of items

V. Results

Performance evaluation:

We ran a performance evaluation to determine taxonomy comprehensiveness. This essentially is a measure of how well the pension disclosure data can be mapped to the tags from the taxonomy. This is done by calculating the success rate of tagging different data items. The evaluation of the parsing logic is performed on both the test dataset and the training dataset.

Figure 9 shows the matrix visualization of the data. The original data matrix is displayed and different shades are used to graphically represent the values present in the matrix. White or light grey represent values near zero and darker shades represent larger values. Shades of grey signify that these phrases occur less frequently. Black horizontal dividers separate the clusters. In this diagram if we take the example of the first cluster which is highlighted we can see that the phrases "actuarial losses," "benefit obligation end of year," "benefit obligation beginning of year," "benefit payments," "curtailment losses," "interest cost," "service cost," "settlements," "special termination benefits," "plan amendments," and "prepaid pension assets" are shown in black. As a result, these phrases were clustered under "Change in Benefit Obligation". Similar results can be found for other

clusters. This visual was useful while developing the taxonomy structure.

	Pension header			Section identifier			Line item parsing			Overall		
	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 3 Performance evaluation of the Parsing module for Training and Test dataset

Manual checks were performed to determine the success rates of the parsing logic at different levels by determining whether actually those terms exist and whether they were correctly parsed out.

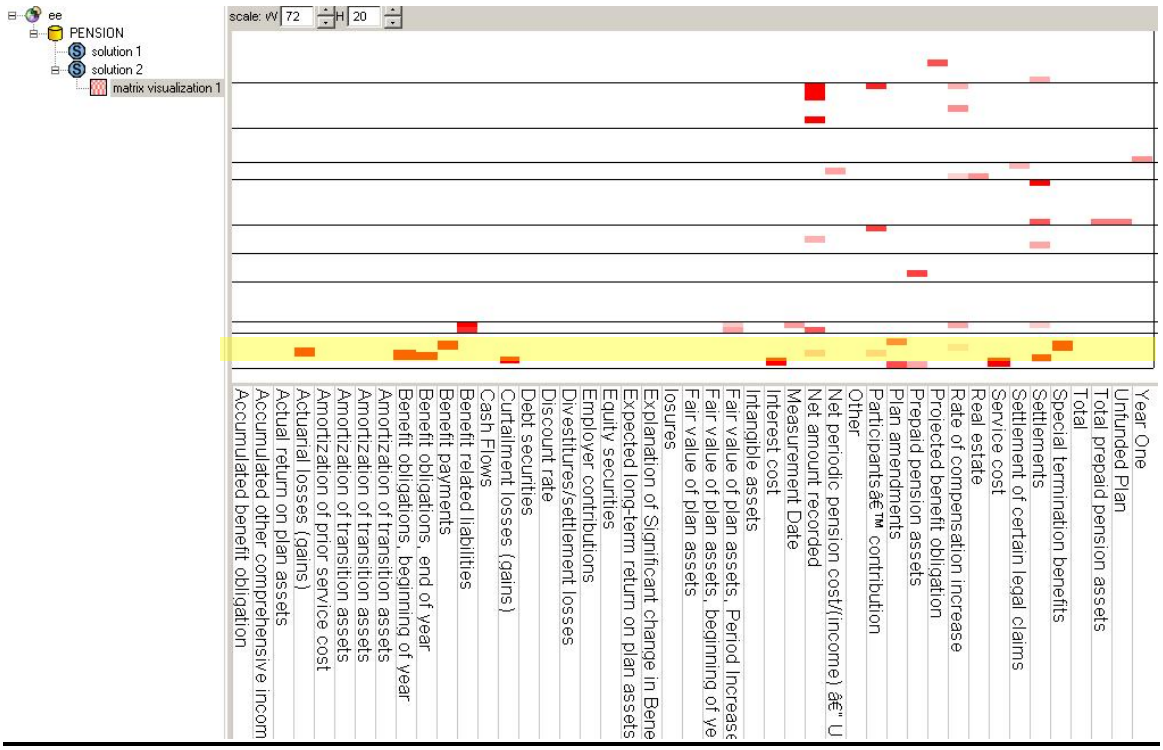

Fig 9 Graph showing concentration of phrases in different clusters

Comparison of taxonomy structure:

It needs to be mentioned here that the comparison between the two taxonomies is a general comparison of their hierarchical structure rather than calculating distances. The following is a basic hierarchical structure of the taxonomy. The general structure of the two taxonomies is similar and different fonts, as shown below, have been used to make a comparison between the two.

XBRL taxonomy
Alternative method taxonomy
Common to both types of taxonomies

1. Cost/(income) of pension plans / Components of the net periodic benefit cost (credit) for continuing operations (COST)
- Service cost
- Interest cost
- Expected return on plan assets
- Recognized Net Gains(Losses) due to Settlements and Curtailments
    - ***Recognized Net Gains(Losses) due to Settlements***
    - ***Recognized Net Gains(Losses) due to Curtailments***
    - ***Total***
- Amortization of Gains(Losses)
- Amortization of transition assets
- Amortization of prior service cost
- Unrecognized Net loss
- Settlement of certain legal claims
- Recognized actuarial losses
- Divestitures/settlement losses
- Net periodic pension cost/(income) — U.S. Plan and material non-U.S. Plans
- Cost of other defined benefit plans
- Net cost of multi-employer plans

Total net periodic pension cost/(income) for all defined benefit plans

Fig. 10 Cost of pension Plans

"Recognized Net Gains (Losses) due to Settlements and Curtailments" is reported as a single term as found in the companies' data. In the XBRL taxonomy, this is broken down further into separate entries for "Settlements" and "Curtailments". There are several items in the historical data that do not occur in the XBRL taxonomy. A possible explanation could be that "unrecognized net loss" and "recognized actuarial losses" might be a part of the "amortization of gains (losses)" and so the former two do not appear as separate line items in the XBRL taxonomy. On the other hand, "cost of other defined benefit plans", "net cost of multi-employer plans" and "net periodic pension cost/(income)" could have been added to the "total net periodic pension cost/(income) for all defined benefit plans".

2. Change in benefit obligations (ChBO)

- Benefit obligations, beginning of year
- Service cost
- Interest cost
- Participants' contribution
- Plan amendments
- Actuarial losses (gains)
- ***Gross Prescription Drug Subsidy Receipts Received***
- Business Combinations and Acquisitions, Benefit Obligation
- ***Divestitures***
- Foreign Currency Exchange Rate Changes
- Benefit payments
- Special termination benefits
- Settlements
- Curtailment losses (gains)
- Settlement/curtailment/acquisitions/dispositions
- Benefit obligations, end of year

Fig. 11 Change in benefit obligations

"Gross Prescription Drug Subsidy Receipts Received" and ***"Divestitures"*** are not commonly used by companies. In several cases, settlements and curtailments are combined whether or not they appear separately in the XBRL taxonomy.

3. Change in fair value of plan assets:
- Fair value of plan assets, beginning of year
- Actual return on plan assets
    - ***Actual Return on Plan assets still held***
    - ***Actual Return on Plan assets sold during period***
    - ***Actual Return on Plan assets, Total***
- Employer contributions
- Participants' contribution
- Fair value of plan assets, Period Increase (Decrease)
- Foreign Currency Exchange Rate Changes
- ***Transfers between Measurement levels***
- Benefit payments
- ***Plan, Purchases, Sales and Settlements***
    - ***Business Combinations and Acquisitions, Plan assets***
    - ***Divestitures, Plan assets***
    - ***Settlements, Plan assets***
    - ***Purchases, Sales and Settlements, Total***
    - Settlement/curtailment/acquisitions/dispositions
- Fair value of plan assets, Period Increase (Decrease)
- Fair value of plan assets, end of year
- Funded (unfunded) benefit obligation
- Unrecognized net loss
- Unrecognized prior service cost
- Net amount recorded

Fig. 12 Change in fair value of plan assets

Fig. 12 also shows that companies generally aggregate the "actual return on plan assets" and "plan, purchases, sales and settlements" whereas they are disaggregated in the XBRL taxonomy.

4. Amounts recognized in the Consolidated Statement of

Financial Position captions include or amounts recorded in our consolidated balance sheets: (ABS)

- Prepaid pension assets
  - ***Current assets***
  - ***Assets for Plan benefits, Noncurrent***
  - ***Assets for Plan benefits, Total***
- Benefit related liabilities
  - Current Liabilities
  - Liabilities, Noncurrent
  - Liabilities, Total
- Intangible assets
- Accumulated other comprehensive income
- Accrued Benefit Liability
- Total prepaid pension assets
- Deferred Income Taxes
- Amount recognized in Balance Sheet, Total

Fig 13 Amounts recognized in Balance Sheet

Fig. 13 again shows a greater level of disaggregation for "prepaid pension assets" in the XBRL taxonomy whereas that is not reflected in the disclosure of companies. "Intangible assets", Accumulated other comprehensive income, Accrued Benefit Liability, Total prepaid pension assets, Deferred Income Taxes do not appear in the XBRL taxonomy.

5. Information for pension plans with an accumulated benefit obligation

- Projected benefit obligation
- Accumulated benefit obligation
- Fair value of plan assets

6. Weighted-average asset allocation of the pension and postretirement plans

- Equity securities
- Debt securities
- Real estate
- Long duration bonds
- U.S Stocks
- International stocks
- Emerging markets stocks and bonds
- Alternative investments
- Other
- Total

7. Information for pension plans with an accumulated benefit obligation in excess of plan assets:(BOPA)

- Projected benefit obligation
- Accumulated benefit obligation
- Accumulated postretirement benefit obligation(APBO)
- Fair value of plan assets
- ABO less fair value of plan assets

8. Assumptions used in calculations

- Narrative Description
- Weighted Average Assumptions Used in Calculating Benefit Obligation
  - Discount Rate
    - ***Discount Rate Support, Methodology and Source Data***
    - ***Discount Rate Support, Bond Indices***
  - Rate of compensation increase
- Weighted Average Assumptions Used in Calculating Net Periodic Benefit Cost
  - Discount rate
  - Expected long-term return on plan assets
  - Rate of compensation increase
  - ***Change Due to Interim Measurement***

9. Estimated Future Benefit Payments

- Description
- Year One
- Year Two
- Year Three
- Year Four
- Year Five

Fig. 14 Information for pension etc.

In Fig. 14, the line items for 5 and 6 do not appear in the XBRL taxonomy, even though they are found in the historical data. It should be mentioned that many companies have an "accumulated postretirement benefits obligation" instead of an "accumulated benefit obligation."

10. Information on Plan assets

- ***Investment Policies and Strategic Narrative Description***
  - ***Target Allocation Percentage***
  - ***Investment Goals***
  - ***Risk Management Practices***
  - ***Significant Concentrations of Risk***
  - ***Permitted Investments***
  - ***Prohibited Investments***
  - ***Derivatives Use***
  - ***Diversification***
  - ***Relationship between Plan Assets and Benefit Obligations***
- Weighted-Average Asset allocation (Actual Plan asset allocation)
  - Equity securities
  - Debt securities
  - Real estate
  - Other Plan assets
  - Total
- Assets, Target Allocations
  - Target allocation Percentage of Assets, Debt securities
    - ***Debt Securities***
    - ***Debt Securities, Range Minimum***

        - ***Debt Securities, Range Maximum***
    - Target allocation Percentage of Assets, Equity securities
        - ***Equity Securities***
        - ***Equity Securities, Range Minimum***
        - ***Equity Securities, Range Maximum***
    - Target allocation Percentage of Assets, Real Estate
        - ***Real Estate***
        - ***Real Estate, Range Minimum***
        - ***Real Estate, Range Maximum***
    - Target allocation Percentage of Assets, Other
        - ***Other***
        - ***Other, Range Minimum***
        - ***Other, Range Maximum***
- Narrative Description of Basis Used to Determine Overall Expected Long-term Rate-of Return on Assets Assumption
- ***Additional Disclosures about Plan assets Type of Employer and Related Party Securities Included in Plan assets***
- ***Amount of Employer and Related Party Securities Included in Plan assets***
- ***Number of shares of Equity Securities Issued by Employer and Related Parties Included in Plan assets***

Fig. 15 Information on plan assets, etc.

Fig. 15 gives several examples where the official taxonomy is disaggregated, but company data is aggregated.

11. Funded Status of the Plan
    - Fair Value of Plan assets
    - Benefit Obligation
    - Funded Status of Plan, Total

12. Unfunded Plan

13. ***Other Comprehensive Income, Adjustment before tax***

- ***Amounts recognized in other comprehensive income, net gain(loss) before tax***
    - ***Other Comprehensive Income, net amortized gain(loss) before tax***
    - ***Other Comprehensive Income, reclassification of defined benefit plan's net gain(loss) recognized in net periodic benefit cost, before tax***
    - ***Amounts recognized in other comprehensive income, net gain(loss), total***
- ***Amounts recognized in other comprehensive income, net prior service cost(credit) before tax***
    - ***Other Comprehensive Income, net prior service cost(credit) before tax***
    - ***Other Comprehensive Income, amortization of net prior service cost(credit) before tax***
    - Amounts recognized in other comprehensive income, net prior service cost(credit) before tax, Total
- ***Other Comprehensive Income, Reclassification of Net transition asset (obligation), recognized in Net Periodic Benefit Cost ,before tax***
- ***Other Comprehensive Income, Minimum pension liability, Net adjustment before tax***
- ***Other Comprehensive Income, Finalization of pension and non-pension Postretirement plan valuation, before tax.***
- ***Other Comprehensive Income, Adjustment before tax, Total***

14. Accumulated Benefit Obligation

15. Accumulated other comprehensive income, before tax

- ***Net Gains (losses), before tax***
- ***net prior service cost(credit) before tax***
- ***Net transition assets(Obligations), before tax***
- ***Minimum pension liability, before tax***
- ***Total***

16. Amounts Amortized from Accumulated Other Comprehensive Income (Loss) in next Fiscal year
    - Amortization of net gains(losses)
    - Amortization of net Prior service cost(credit)
    - Amortization of net Transition Asset(Obligation)
    - Total

17. Pension plans with a benefit obligation in excess of plan assets
    - Aggregate Benefit Obligation
    - Aggregate Fair value of Plan assets

20. Estimated Future employer contributions in Next Fiscal Year

21. Alternative Methods to Amortize Prior Service Amounts

22. Alternative Methods to Amortize net gains and losses

23. Method to Determine Vested Benefit Obligation

24. Description of any Substantive Commitment Used as Basis for Accounting for Benefit Obligation

25. Special Termination Benefits during Period

- ***Description of Event Resulting in Special or Contractual Termination benefits recognized during period***
- ***Cost of providing Special termination benefits***

25a.Plan Amendment

- ***Description***
- ***Effect on Accumulated Benefit Obligation***
- ***Effect on Net Periodic Benefit Cost***

25b. Explanation of Significant change in Benefit Obligations or Plan assets not apparent from other disclosures

26. Settlement and Curtailments

- ***Description***
- ***Effect on Accumulated Benefit Obligation***

27. Measurement Date

28. Pension plans with a Accumulated benefit obligation in excess of plan assets

- Aggregate Projected Benefit Obligation
- Aggregate Accumulated benefit obligation
- Aggregate Fair value of Plan assets

Fig. 16 Funded status of the plan etc.

VI. An application of the methodology and tool:

Comparison of pension footnote reporting structure between 2000 and 2010

The parsing tool and the data matching module developed for the taxonomy generation process were designed to be used elsewhere in the 10K statements. As an experimental study, forty five companies were chosen randomly. 10K statements of these companies from year 2000 to 2010 were retrieved manually. Line items were then extracted and term matching done on those line items to understand whether companies repeatedly disclosed in the same format or their pension reporting format changed over the years. Examples typical of our results follow.

Fig 17. A new term added in a year

Components of net periodic benefit cost

Defined benefit plans:
Service cost
Interest cost
Expected return on assets
Amortization of:
Transition Asset
Prior service cost
Settlement
Unrecognized net loss
Net periodic benefit cost for defined benefit plans

Defined contribution plans
Service cost
Interest cost
Expected return on assets
Amortization of:
Prior service cost
Settlement missing in 2007
Unrecognized net loss

Net periodic benefit cost for defined benefit plans

Fig. 19 A term missing in a year

Change in Benefit Obligation
Benefit obligation at January 1
Service cost
Interest cost
Plan participant contributions
Medicare PartD subsidy in 2010
Plan amendments
Actuarial (gain) loss
Acquisitions included in 2010
Divestitures included in 2010
Benefits paid
Curtailment
Recognition of termination benefits
Foreign currency exchange rate change
Benefit obligation at December 31
Accumulated benefit obligation portion of above at December 31

Fig. 20 Group of items added in a year

As we can see from the examples in Fig. 17 through Fig. 24 several instances have been found where a company has either added one (Fig. 17) or many(Fig. 20,21,22,23,24) new terms in their pension disclosures. In some cases terms have found to be missing(Fig. 19) in a particular year. Complete new section (Fig.18) has been found to be added in certain cases. A more detailed study of why these changes occur would be done in a future research project.

The Company's pension plan weighted-average asset allocations at December 31, by asset category, are as follows:

Long duration bonds
U.S. stocks
International stocks
Emerging markets stocks and bonds
Alternative (private) investments
Total

Fig. 18 A new section added in a particular year

Fig. 21 New Terms added in a year

Change in Fair Value of Plan Assets
Fair value of plan assets at January 1
Acquisitions included in 2010
Divestitures included in 2010
Actual return on plan assets
Company contributions
Plan participant contributions
Medicare PartD subsidy included in 2010
Benefits paid
Foreign currency exchange rate change
Fair value of plan assets at December 31:

Fig. 22 New Terms added in 2010 as per SFAS 158 requirements

Funded Status AS in 2010
Amounts Recognized in the Consolidated Balance Sheet at December 31
Noncurrent assets
Current liabilities
Noncurrent liabilities
Total recognized

Fig. 23 New Terms added in a year

The fair values of our pension plan assets at December 31, 2009, by asset class are as follows: As in 2010
Cash and cash equivalents
Diversified equity securities
Government debt securities
Diversified corporate debt securities
Mortgage-backed securities
Common/collective trusts
Mutual funds
Derivatives
Private equity funds
Insurance contracts
Preferred stock
Real estate

Fig. 24 New Terms added in in 2010

Severence Accrual
As a result of the 2008 business environment's impact on our operating and capital plans, a reduction in our overall employee work force occurred in 2009.
Beginning balance
Accruals
Benefit payments
Accrual reversals
Ending Balance

VII. Conclusion

The objectives of this paper are threefold: (1) to develop a semiautomatic method of taxonomy creation; (2) to compare the structure of the taxonomy created against the XBRL US GAAP taxonomy and (3) to demonstrate how the tool developed as a part of this process can be used for more exploratory research.

A prototype tool has been developed to extract and restructure information from pension footnotes of 10k statements. Hierarchical clustering algorithm is applied to the data to develop a taxonomy structure. The parsing module has been evaluated and functions well, with an overall success rate of 97% for the training data set and 95% for the test data set.

Comparison of the taxonomy structure(from historical data) with that of the XBRL taxonomy reveals some differences between the two. In general it is found that companies tend to aggregate some of the data whereas a more disaggregated structure is followed in the XBRL taxonomy. Apart from this some new terms are found in the historical data taxonomy which does not exist in the XBRL taxonomy. In some cases the positioning of the terms might be different. These new terms are mostly seen under "Cost of pension plan", "Change in fair value of plan assets", "Amount recognized in balance sheet", "Weighted average asset allocation of the pension and postretirement plans".

The parsing tool developed could be used for exploratory research. Its use has been demonstrated by comparing some randomly selected Fortune 500 companies across a span of ten years. It is seen that in some cases there are additions of new terms or even addition of a completely new section by a company. Future research can be carried out to explore some of these pension footnote reporting trends.

10K filings of Fortune 500 companies have been used, and the results obtained may represent the trends observed in these companies rather than representing a more varied and larger cross section of companies. Another drawback could be using data directly from the filings which may lead to a taxonomy that is a reflection of the way companies report rather than being a true representation of the reporting standards. Future research maybe be carried out to address some these issues by using a larger dataset of Fortune 1000 companies.

VIII. References:

Bates, M. J. 1986. Subject access in online catalogs: a design model. *Journal of the American Society for Information Science*, 37(6):357-376, November

Bies, Susan Schmidt, 2003. Remarks by Governor Susan Schmidt Bies, At the Conference on Market Discipline, Federal Reserve Bank of Chicago, Chicago, Illinois October 31, 2003 http://www.federalreserve.gov/BoardDocs/Speeches/2003/20031031/default.htm

Bovee, M., A. Kogan, R. P. Srivastava, M. A. Vasarhelyi, K. M. Nelson, 2005. Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible

Business Reporting Language (XBRL). *Journal of Information Systems*, 19 (1 Spring) 19-41.

Bovee, M., M. Ettredge, R. P. Srivastava, and M. A. Vasarhelyi. 2002. "Does the Year 2000 XBRL Taxonomy Accommodate Current Business Financial Reporting Practice? *Journal of Information Systems* 16 (2) 165-182.

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

Chen, H., & Lynch, K. J. (1992). Automatic Construction of Networks of Concepts Characterizing Document Databases. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(5), 885-902.

Chen, H., Martinez, J., Kirchhoff, A., Ng, T. D., & Schatz, B. R. (1998). Alleviating Search Uncertainty through Concept Associations: Automatic Indexing, Co-Occurrence Analysis, and Parallel Computing. *Journal of the American Society for Information Science*, *49*(3), 206-216.

Chen, H., Ng, T. D., Martinex, J., & Schatz, B. R. (1997). A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. *Journal of the American Society for Information Science*, *48*(1), 17-31.

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,.

Chen, H., Yim, T., & Fye, D. (1995). Automatic Thesaurus Generation for an Electronic Community System. *Journal of the American Society for Information Science*, *46*(3), 175-193.

Crouch, C. J. (1990). An Approach to the Automatic Construction of Global Thesauri. *Information Processing and Management*, *26*(5), 629-640.

Crouch, C. J., & Yang, B. (1992). *Experiments in Automatic Statistical Thesaurus Construction.* Paper presented at the Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen Denmark.

Dumais, S. T. (1991). Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments, & Computers*, *23*(2), 229-236.

Dumais, S. T. (1993). LSI Meets TREC: A Status Report. *[Online], Available: http://lsi.research.telcordia.com/lsi/LSIpapers.html*, (5/30/02).

Dumais, S. T. (1994). Latent Semantic Indexing (LSI) and TREC-2. *[Online], Available: http://lsi.research.telcordia.com/lsi/LSIpapers.html*, (5/30/02).

Dumais, S. T. (1995). Latent Semantic Indexing (LSI): TREC-3 Report. *[Online], Available: http://lsi.research.telcordia.com/lsi/LSIpapers.html*, (5/30/02).

Dumais, S. T., Furnas, G. W., & Landauer, T. K. (1988). Using Latent Semantic Analysis to Improve Access to Textual Information. *, [Online]*(Available: http://lsi.research.telcordia.com/lsi/LSIpapers.html), (5/30/02).

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 http://www.nysscpa.org/cpajournal/2005/1005/essentials/p28.htm http://www.frbsf.org/publications/economics/letter/2003/el2003-19.html

Fisher, Dorothy M. and Steven A. Fisher 2001. "Integrating Financial Reporting Systems With XBRL" Seventh Americas Conference on Information Systems

Fisher, Ingrid E. 2004. "On the structure of financial accounting standards to support digital representation, storage and retrieval" *Journal of Emerging Technologies in Accounting* 1, 23 – 40.

Frazier, Ingram and Tennyson, 1984. "A methodology for the analysis of narrative accounting disclosures" *Journal of Accounting Research* 22(1) Spring

Edgar Online http://www.edgar-online.com/

Gangolly J. 1995. Some thoughts on the engineering of Financial Accounting standards. In *Artificial Intelligence in Accounting and Auditing*

Gerdes, 1. Jr. 2003. EDGAR-Analyzer: Automating the analysis of corporate data contained in the SEC's EDGAR database. *Decision Support Systems 35* (1): 7-29.

Gopalakrishnan, V. 1994. "The effect of recognition vs. disclosure on investor valuation: The case of pension accounting"; *Review of Quantitative Finance and Accounting*, 4(4) / December

Jacobs P.S, Rau Lisa F. 1990. **"**SCISOR: extracting information from on-line news", Communications of the ACM 33(11 November): 88 - 97

Kieso, Donald; Jerry Weygandt and Terry Warfield "Intermediate Accounting"

U.S. Department of Labor www.dol.gov

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

Katriel and Rafsky "Weight-Rate: A neoclassical approach to text categorization" *ACM SIGIR '97 Conference*

Nelson, K M., A. Kogan, R P. Srivastava, M. A. Vasathelyi. and H. Lu. 2000. V i auditing agents: Tbe EDGAR agent challenge. *Decision Support Systems* 28 *(3):* 241-253.

Shui-Lung Chuang, Lee-Feng Chien; Taxonomy Generation for Text Segments: A Practical Web-Based Approach, *ACM Transactions on Information Systems, Vol. 23, No. 4, October 2005, Pages 363–396.*

Salton, G. 1989.*Automatic Text Processing*. Addison-Wesley Publishing Company, Inc., Reading, MA

ShuiLung, Chuang; Chien LeeFeng; 2004. A Practical Web based Approach to Generating Topic Hierarchy for Text Segments; *CIKM'04*, November 8–13, 2004, Washington, DC, USA.

Stavrianou, Anna; Periklis Andritsos; Nicolas Nicoloyannis; 2007. Overview and Semantic Issues of Text Mining; *SIGMOD Record*, 36(3) September

Tan, Pang-Ning; Michael Steinbach; Vipin Kumar "Introduction to data mining ";Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA Year of Publication: 2005

Tergesen, Ann "The Fine Print: How to Read Those Key *Footnotes" Business Week* 2/4/2002 Issue 3768, p94-96.

Wayman, Rick, An Investor's Checklist To Financial Footnotes http://www.investopedia.com/articles/analyst/03/100103.asp

Wikipedia http://en.wikipedia.org/wiki/Retirement_plans_in_the_United_States Wikipedia; Semantic web services

Wu and Gangolly 2000 On the Automatic Classification of Accounting concepts: Preliminary Results of the Statistical Analysis of Term-Document Frequencies.

XBRL.ORG http://xbrl.org/FRTApproved/ http://xbrl.us/preparersguide

IX. Appendix:

Table 1 List of sections and number of synonyms for each section

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

Table2: Details of items under each section

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