Automating the Process of Taxonomy Creation and Comparison of Taxonomy Structures

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.