On the Automatic Classification of Accounting concepts: Preliminary Results of the Statistical Analysis of Term-Document Frequencies

Jagdish Gangolly, Department of Accounting & Law, School of Business, State University of New York at Albany, Albany, NY 12222. Email: j.gangolly@albany.edu.

Yi-Fang, Wu, School of Information Science & Policy, State University of New York at Albany, Albany, NY 12222. Email: yw4698@cnsunix.albany.edu.

Abstract

In this paper we study the feasibility of automatic classification of financial accounting concepts by statistical analysis of term frequencies in the financial accounting standards. The analysis uses principal-components analysis to reduce the dimensionality of the dataset, and then uses agglomerative nesting algorithm (agnes) to derive clusters of concepts.

1 Introduction

The proliferation of financial accounting literature and the explosive growth of the World Wide Web as an application deployment platform have necessitated the development of mechanisms to provide efficient integration of text and structured databases from the viewpoint of their organization and retrieval. Anecdotal evidence suggests that around 90 per cent of databases for large American Corporations consist of text (the remaining 10 per cent consisting of structured databases such as the traditional accounting records). The preponderance of text in the corporate repositories of information, its pervasive role in electronic commerce and yet the scant attention accorded to its study in accounting is truly astonishing.

While the theoretical foundations of structured databases are now well-understood and their study has permeated the education of an accountant, the study of text databases that is now becoming the focal point of research in a variety of fields such as corpus based computational linguistics (of the natural language processing as well as statistical variety), data mining and knowledge discovery in databases, data visualization, neural networks and the like, is truly deserving of our attention. The contributions of these fields have enabled the development of infrastructures for the use of electronic texts in new and innovative ways for organizing, traversing, browsing, skimming, summarizing, categorizing, and searching in text databases. The accounting profession can gain tremendously in terms of productivity by the development of analogous infrastructure blending text and non-text accounting databases to support workflow as well as document management.

One ingredient in the development of such infrastructure is the classification of concepts. While traditionally the development of lists of concepts to be classified & their classification has been accomplished manually, the increasing size of text databases and the high cost of domain expertise needed to develop the classifications has necessitated the development of methods for automatic indexing and classification of concepts, or for the augmenting (or supporting) human classifiers by providing them with information generated by the automatic analysis of text. We can broadly classify these methods into those based on the statistical analysis of term-document frequencies pioneered by Salton's [12] vector space model of information retrieval, and those founded on the analysis of concepts based on facet analysis (Ranganathan [10]), and those based on the application of the theory of complete lattices (Ganter & Wille [5]). (For a review of the various approaches to automatic classification see Kent [8]).

While facet analysis seems to be experiencing a resurgence of interest, and the lattice theoretic approach is appealing on account of its sound mathematical foundations, they are often infeasible since they need massive human efforts devoted to organizing the concepts for classification, especially in the case of domains with large and/or complex vocabularies. The statistical methods based on term-document frequencies, even though they lack formal theoretical justification, on the other hand, are attractive in such situations since their analyses can always be subjected to the scrutiny of the domain experts before a classification is accepted. Moreover, the statistical analyses of term-document frequencies can themselves shed light on the Organization of concepts for analysis using either facet analysis or lattice-theoretic methods.

In this paper, we provide some results obtained in a very preliminary investigation involving the statistical analysis of the term-document frequencies for the pronouncements of rulemaking bodies for financial accounting in the United States (Accounting Principles Board and the Financial Accounting Standards Board). Since this is only a preliminary investigation, our objectives are quite modest. While sophisticated methods such as Kohonen networks (Self Organizing Maps) and Latent Semantic Indexing are envisaged for the near future, the objective here is a rudimentary exploratory data analysis of the financial accounting standards to gain some insights into the feasibility of using corpus-based statistical analysis of data in the financial accounting standards domain. Efforts are under way, however, in the text analysis group at the State University of New York at Albany to develop corpora and tagging schemes for accounting text to support empirical research in automatic classification, thesauri construction, and temporal reconstruction of accounting standards (Gangolly [4]) in the accounting domain. In the next section we provide a brief list. of the various statistical approaches to automatic classification and the techniques used in data preparation and analysis (detailed review of those methods must await a final version of this paper). Finally, in section 3 we describe the procedures used to prepare the data for indexing, the procedures used for pruning the list of index terms, the nature of data finally used in the analysis, the preliminary analysis of the results, and our perceived future directions for research in the area.

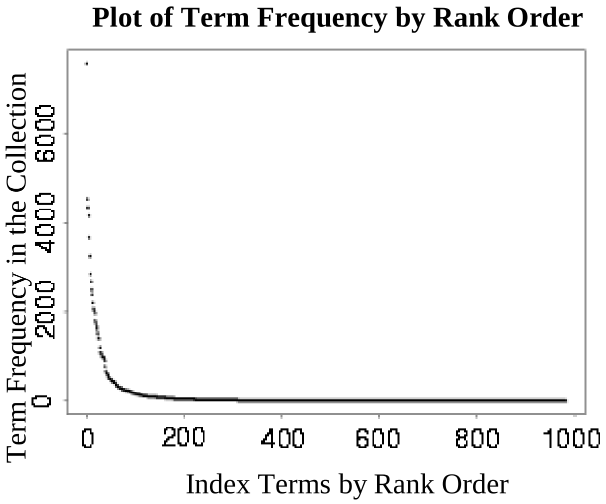

Figure 1: Zipf Curve of the FARS Database

2 Classification and Term-Document Frequencies

The idea that the frequencies of words in documents convey information regarding their importance has a long history in Linguistics and Information Science. In fact, the well-known Zipfs law ([15]; [6]), which states that the plot of ordered term frequencies in any corpus resembles a rectangular hyperbola, where as the information value of such ordered terms has a familiar bell-shaped curve, has been used to remove non-content fluff words (also referred to as stop words) in arriving at index terms used in thesauri construction and information retrieval (See [13]; [12]). Most statistical studies in conceptual classification can be divided into two phases. In the first phase, often called the text pre-processing phase, a set of index terms is derived for classification (by removal of fluff words, stemming, and inclusion of phrases denoting concepts), and the term-document frequencies are computed. In the second statistical analysis phase, the term-document frequencies are analyzed in order to gain an understanding of the associations between the terms.

The text pre-processing phase involves the five step process of removal of fluff words by using a list of such words (usually consisting of articles, conjunctions, and prepositions) called stop list, stemming (reducing syntactic variations of words to their common grammatical roots by removal of prefixes and suffixes, considering plurals, gerund forms, past tense suffixes, etc.), and detection of phrases, pruning the list by eliminating very high and very low frequency words, and finally computing the frequency of the terms in the resultant list for each document in the corpus. The result of such text pre-processing phase is a matrix of term-document frequencies.

In the statistical analysis phase, the rows in the term-document frequency matrix are interpreted as Datapoints, and analysis is performed using one of the standard techniques of multi-variate analysis. The techniques used have included factor analysis, multi-dimensional scaling ([2]), and some of the more recent methods including hopfield neural networks, latent semantic indexing ([3]).

3 Data, Analysis, and Future research

The basic data for this study was derived from the FARS database of the Financial Accounting Standards Board. The documents considered included Statements of Concepts 1 - 6, Statements of Financial Accounting Standards 1 - 132, Financial Interpretations 1 - 42, and FTBs 79-1 to 94-1. This set of documents was indexed using Indexicon, an indexing utility that works in conjunction with Microsoft Word. Indexicon scans and preprocesses the text to prepare a back-of-the-book index. The index prepared by Indexicon includes words as well as phrases. We removed the page numbers in such index and terms with all capital letters (since they appear only in statement/section/paragraph headings. Next, we moved sub-headings in the index terms before or after the headings to create new entries. For example, "accounting - principle" was used to create the new entry "accounting principle". Since the index entries prepared by Indexicon are case sensitive, there were many duplicates (for example, accounting, Accounting, and duplicates on account of inclusion of both singular and plural numbers). We also removed all proper nouns, and terms that had little import to accounting (for example, earthquakes, floods, etc.), to finally arrive at a list of 983 terms.

Term-document frequency matrix (983 terms x 232 documents) was computed and the Zipf curve plotted (Figure 1).

Due to limitations of data visualization for such a large collection of index terms, we selected 93 index terms with frequencies in the range 100 and 994 (i.e., we omitted very high and very low frequency terms) for further analysis. The corresponding term-document frequency matrix (93 terms x 232 documents) formed our final dataset for analysis. Since this is a rather large dataset, we reduced dimensionality by doing a principal components analysis of this matrix and computed the scores on the first 23 principal components that explain 90 per cent of the variance. Figure 2 gives the scree plot of the first 23 principal components.

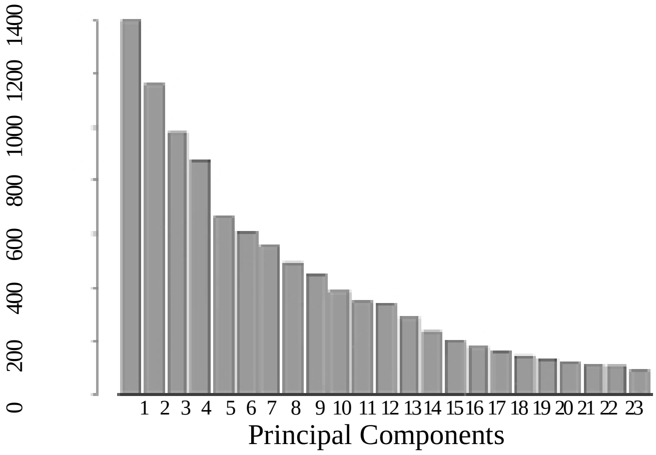

Figure 2: Scree Diagram of the First 23 Principal Components (94 Index Terms)

Finally, we performed cluster analysis on the matrix of principal components scores (93 terms x 23 principal component scores) using the agglomerative nesting routine described in Kaufman & Rousseeuw ([9]) implemented in the system S-Plus ([14]). The resulting dendrogram is given in Figure 3.

As can be seen from the dendrogram in Figure 3, the results are rather mixed. While the agglomerative coefficient is high (0.80238 of a maximum 1.0) and we do see the clusters for tax, leases, marketable securities, and financial statements, the hierarchies are not as clear-cut as one would wish. The fact that we restricted our attention to just the pronouncements of FASB, just one of the venues for discourse in financial accounting, and the severe pruning of the index term list for further analysis also would have restricted the validity of the results. Nevertheless, the fact that rudimentary clusters are differentiable even in this preliminary statistical analysis suggests that further research in the area of automatic classification in financial accounting can be a fruitful endeavor.

References

[1] Ricardo Baeza-Yates and Berthier Rebeiro-Neto, *Modern Information Retrieval*, AddisonWesley, 1999.

[2] B.T. Bartell, G.W. Cottrell and R.K. Belew, "Latent Semantic Indexing is an Optimal Special Case of Multi-dimensional Scaling", *Proceedings of the 15th International Conference on Research and Development in Information Retrieval*, 1992, pp.161-167.

[31 S. Deerwester, S.T. Dumais, G.W. Furnas and T.K. Landauer, "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science,* 1990, pp.391-407.

[4] Jagdish Gangolly, "Some Thoughts on the Engineering of Financial Accounting Standards", *Artificial Intelligence in Accounting & Auditing III,* Miklos Vasarhelyi (ed), Markus Weiner Publishing (1994), pp. 177-190.

[5] Bernhard Ganter and Rudolf Wille, Formal *Concept Analysis: Mathematical Foundations*, Springer, 1999.

[6] G. Gonnet and Picardo Baeza-Yates, *Handbook of Algorithms and Data Structures, 2nd. edition*, Addison-Wesley, 1991.

[7] Robert E. Kent and C. Mic Bowman, "Digital Libraries, Conceptual Knowledge Systems, and the Nebula Interface", *Transarc Working Paper,* 1994.

[8] Robert E. Kent, "Automatic Classification", *Intel Corporation surveys the field of "automatic classification".* March 1995. http://wave.eecs.wsu.edu/WAVE/IntelSurveyps

[9] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley and Sons, 1990.

[10] S.R. Ranganathan, *Colon Classification,* Asia Pubhshing House, 1963.

[11] Gerard Salton, *Automatic Information Organization and Retrieval*, McGraw-Hill Book Company, 1968.

[12] Gerard Salton, *Automatic text processing: the transformation, analysis, and retrieval of information by computer*, McGraw-Hill Book Company, 1989.

[13] C.J. Van Rijsbergen, *Information Retrieval*, Butterworths, 1979.

[14] W.N. Venables and B.D. Ripley, *Modern Applied Statistics with S-Plus,* Springer, 1997.

[15] G. Zipf, Human *Behavior and the Principle of Least Effort,* Addison-Wesley, 1949.

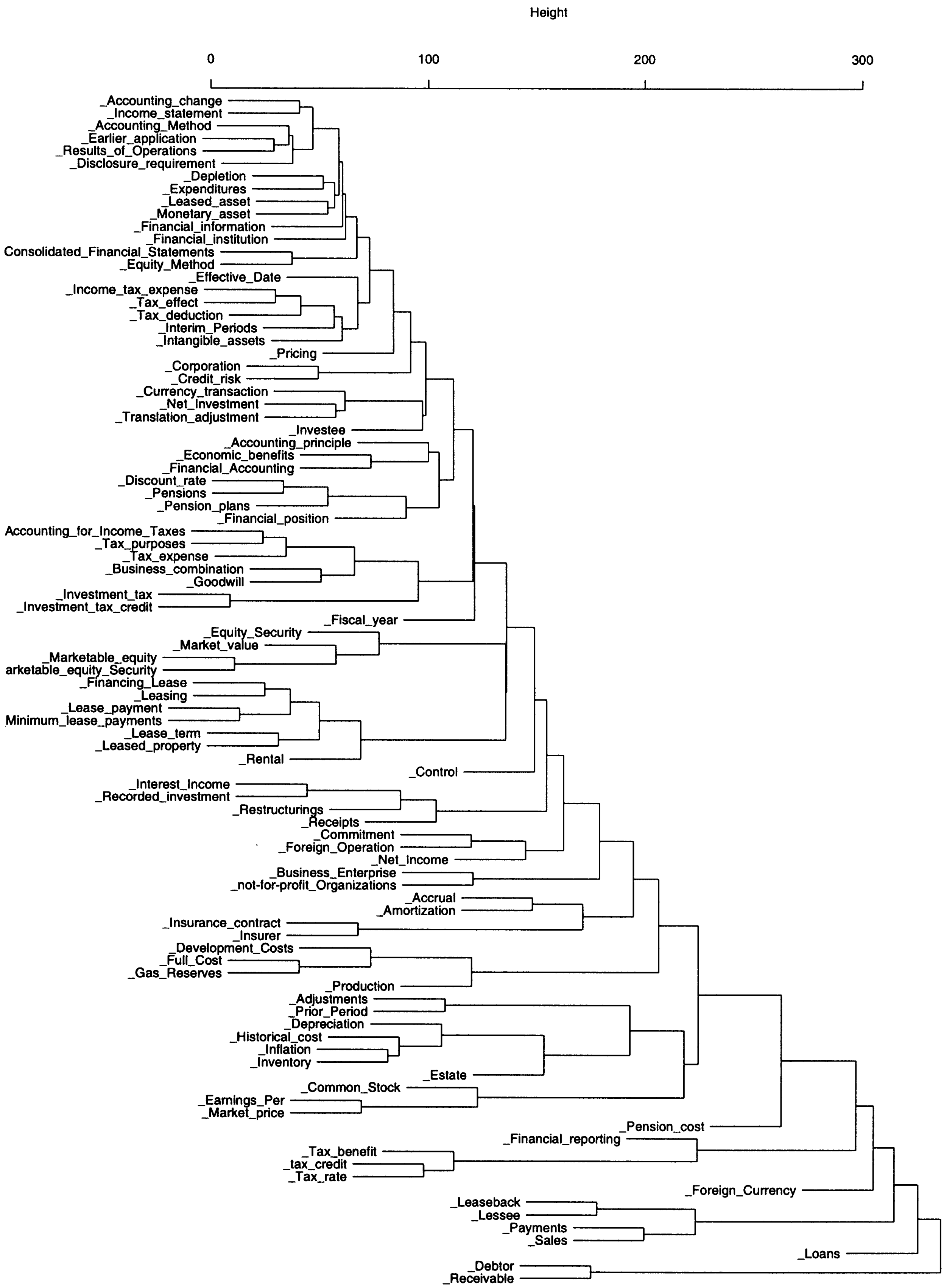

Figure 3: Results of Cluster Analysis (Agglomerative Nesting)