A Comparative Study on Measuring Similarity between Companies Based on Financial Statements

Zhiwen Zhang [1, a]

[1] Rady School of Management, University of California, San Diego, USA.

[a] bachelorwho@163.com

Abstract. It is important for financial researchers to measure similarity between companies. This passage classifies references in this area based on research objects and research methods. To be more specific, there are four research objects based on data structure and data connectivity. Research methods can also be classified into four types based on data form and modeling type. It can be concluded that there are academic blanks on models based on graph theory: no matter independent data or unstructured data should be used more wildly. Moreover, interconnected structured data should be used more often when vectors are introduced into predictive modeling.

Keywords: company similarity; classification; creativity areas.

1. Introduction

Measuring similarity between companies is very important in financial researches. Many scholars have proposed measures to show the similarity between companies and created methods for classification.

Research objects can be divided into four types based on data structure and data connectivity: independent structured data, interconnected structured data, independent unstructured data and interconnected unstructured data.

As for research methods, there are four types based on data form and modeling type: vector-based inferential modeling, graph-based inferential modeling, vector-based predictive modeling and graph-based predictive modeling.

The rest of the paper is organized as follows. Section II gives the classification of research objects of company similarity. Section III introduces the classification of research methods. Section IV concludes the paper.

2. Classification of Research Objects

2.1 Criteria

In this section, two independent and different criteria would be used to divide research objects into different types: 1) Data Structure. There are two types here: structured data or unstructured data. Structured data refers to data with a clearly defined format and organization, with each data field carrying a clear meaning. As for unstructured data, it has no clear format or organization, but a high level of complexity and diversity. 2) Data Connectivity. There are two kinds of data connectivity here: independent data or interconnected data. Independent data means that each company's data is published separately, while interconnected data indicates that data from multiple companies are presented together.

2.2 The Classification

2.2.1 Type I: Structured Data & Independent

This type is independent structured data. In this case, data of each company is structured and independent, such as financial statements. References ([1][2][3][5][8][12]) belong to Type I. Reference [1] obtained the financial statement data from Silverfin, along with the financial ratios and industry activity codes. Reference [2] used the financial statement data of 1000 Belgian

companies, with their disclosure published separately before 2019. Reference [5] used a dataset consisting of 71295 firm-years' financial statement data. Reference [12] extracted financial statement in XBRL form, with each company containing an instance document and a taxonomy set. These references all made use of financial statements, which are both independent and structured. Reference [3] selected 50 largest oil and gas companies from the Herold database, in which the data is independent structured data. Reference [8] obtained data from 2015 EU Industrial R&D Investment Scoreboard, in which each company has its own independent structured patent portfolio.

2.2.2 Type II: Structured Data & Interconnected

This type is interconnected structured data. In this case, data of each company is structured and interconnected, such as U.S. input-output tables (IOTs). References [6] belongs to Type II. References [6] used input-output tables (IOTs) as datasets, which describe the selling and buying relationships between producers and consumers in an economy. The data from input-output tables is interconnected and structured.

2.2.3 Type III: Unstructured Data & Independent

This type independent unstructured data. In this case, data of each company is unstructured and independent, such as texts from 10-K. References ([4][7][9]) belong to Type III. Reference [4] extracted features from 69100 company websites, including URLs, titles, MetaDiscription, MetaKeywords, bodies and headings. The dataset can be considered as independent unstructured data. Reference [7] acquired 10-K forms filed by companies of the S&P Total Market Index from the EDGAR database. Reference [9] proposed a method to measure the differences between companies based on text analysis of 10-K product descriptions.

2.2.4 Type IV: Unstructured Data & Interconnected

This type is interconnected unstructured data. In this case, data of each company is unstructured and interconnected, such as texts from company websites. References ([10][11][13][14]) belong to Type IV. Reference [10] used 10-year collection of Reuters news articles, roughly 21 million articles published between 2003 to 2012, as data source. Reference [13] obtained a public dataset containing documents spanning various news topics. Reference [14] analysed the concurrences on news stories from 1999 to 2002 to get generic links between companies. Because there are usually more than one company involved in a news article, the datasets above are unstructured interconnected data. Reference [11] obtained traffic to the EDGAR website to study synergistic search effects between companies. The data is both unstructured and interconnected.

3. Classification of Research Methods

3.1 Criteria

In this section, two independent and different criteria would be used to divide research methods into different types: 1) Data Form. There are two types here: graph or vector. A graph is composed of given vertexes and edges representing interrelations between vertexes. Edges in a graph can be weighted to indicate the strength of the interrelation. Both numerical and non-numerical relationships between or within companies can be represented by graphs. In addition to the graph, financial information of companies can be represented by vectors. Each company has a list that includes data on all dimensions, which can be considered a vector.2) Modeling Type. There are two kinds of modeling here: inferential modeling or predictive modeling. Inferential modeling explores the causes of the data generation process, selects the model with the most reasonable assumptions and validates the model by fitting tests, leading to high model interpretability but uncertain validity. However, predictive modeling selects the model with the best performance and tries to predict the results of new samples. Although the validity is guaranteed, the model interpretability will be affected.

3.2 The Classification

3.2.1 Type I: Graph & Inferential Modeling

This type is graph-based inferential modeling. References [12] belongs to Type I. References [12] transformed the balance sheet into a labelled directed rooted tree graph to measure similarity between companies, and used the F-statistic to verify the differences between the results and the existing industry divisions. Therefore, graph-based inferential modeling was used in the reference.

3.2.2 Type II: Graph & Predictive Modeling

This type is graph-based predictive modeling. References ([1][2][13]) belong to Type II. Reference [1] presented a graph distance metric for financial statements using the earth mover's distance and tested the result with accuracy. Reference [2] changed the ledger accounts within a financial statement to a vertex-labelled tree, proposed a distance metric to measure the similarity between companies, and used a predictive method to verify the usefulness of the metric. Reference [13] used a deep learning method to encode the network graph of companies in a low-dimensional embedding space and validated the rationality of the method with accuracy-related indicators. These references all used graph-based predictive modeling.

3.2.3 Type III: Vector & Inferential Modeling

This type is vector-based inferential modeling. References ([3][5][6][9][10][11]) belong to Type III. Reference [3] input the data into Dividend Discount Model (DDM) in vector form and used Chow tests to identify firms with similar relationships between valuation multiples and relevant value drivers. Reference [5] proposed a measure of comparability in vector form, and tested the availability of the measure using hypothesis testing. Reference [6] proposed a vector-based method to measure interindustry relatedness based on vertical relatedness and complementarity and tested the effectiveness of the method by hypothesis testing. Reference [9] studied differences between companies with new vector-based measures based on text analysis of 10-K product descriptions and tested the result with R square and p-value. Reference [10] computed the cosine similarity of word vectors trained on 10- year financial news articles to find peer firms and tested the effectiveness of the method with R square. Reference [11] applied a “co-search” algorithm to Internet traffic at the SEC's EDGAR website with the input data in vector form, and tested the usefulness of the reference with p value and R square. The references above all used vector-based inferential modeling.

3.2.4 Type IV: Vector & Predictive Modeling

This type is vector-based predictive modeling. References ([4][7][8][14]) belong to Type IV. Reference [4] built a website classification system and tested its accuracy on a dataset of more than 20000 company websites. Reference [7] proposed an vector-based classification methodology based on business commonalities using topic features learned by the Latent Dirichlet Allocation from firms’ business descriptions and tested the usefulness with accuracy-related methods. Reference [8] proposed a vector-based data driven approach to classify companies to adapt to changing technological landscapes and used a predictive method to test the validity of the approach. Reference [14] proposed a relational vector-space model to abstract the linked structure and tested the usefulness of the model with area under curve and error reduction. The references above all used vector-based predictive modeling.

4. Conclusions

This paper studies the references on similarity measurements and classifies them based on two dimensions: research objects and research methods. There are some academic blanks in company similarity measurements. When introducing graph theory into researches, scholars should concentrate more on independent unstructured data and interconnected structured data.

References

[1] Noels S, De Ridder S, Viaene S, et al. An efficient graph-based peer selection method for financial statements[J]. Intelligent Systems in Accounting, Finance and Management, 2023.

[2] Noels S, Vandermarliere B, Bastiaensen K, et al. An Earth Mover's Distance Based Graph Distance Metric For Financial Statements[C]//2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics. IEEE, 2022: 1-8.

[3] Asche F, Misund B. Who's a major? A novel approach to peer group selection: Empirical evidence from oil and gas companies[J]. Cogent Economics & Finance, 2016, 4(1): 1264538.

[4] Berardi G, Esuli A, Fagni T, et al. Classifying websites by industry sector: A study in feature design[C]//Proceedings of the 30th Annual ACM Symposium on Applied Computing. 2015: 1053-1059.

[5] De Franco G, Kothari S P, Verdi R S. The benefits of financial statement comparability[J]. Journal of Accounting research, 2011, 49(4): 895-931.

[6] Fan J P H, Lang L H P. The measurement of relatedness: An application to corporate diversification[J]. The Journal of Business, 2000, 73(4): 629-660.

[7] Fang F, Dutta K, Datta A. Lda-based industry classification[J]. 2013.

[8] Gkotsis P, Pugliese E, Vezzani A. A technology-based classification of firms: Can we learn something looking beyond industry classifications?[J]. Entropy, 2018, 20(11): 887.

[9] Hoberg G, Phillips G. Text-based network industries and endogenous product differentiation[J]. Journal of Political Economy, 2016, 124(5): 1423-1465.

[10] Kee T. Peer firm identification using word embeddings[C]//2019 IEEE International Conference on Big Data. IEEE, 2019: 5536-5543.

[11] Lee C M C, Ma P, Wang C C Y. Search-based peer firms: Aggregating investor perceptions through internet co-searches[J]. Journal of Financial Economics, 2015, 116(2): 410-431.

[12] Yang S Y, Liu F C, Zhu X, et al. A graph mining approach to identify financial reporting patterns: an empirical examination of industry classifications[J]. Decision Sciences, 2019, 50(4): 847-876.

[13] Raman N, Bang G, Nematzadeh A. Multigraph attention network for analyzing company relations[C]//Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition. 2019: 426-433.

[14] Bernstein A, Clearwater S, Provost F. The relational vector-space model and industry classification[C]//Proceedings of the Learning Statistical Models from Relational Data Workshop at the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI). 2003.