Large-Scale Analysis of Document Similarity Measures of Financial Statements

Abstract

In this paper I design and implement a system for the retrieval, parsing, and clustering based on document similarity of a large web-based corpus of documents. Using corporate filings from the United States Securities and Exchange Commission EDGAR database, I construct a measure of document similarity and implement a fast, scalable, distributed system for the calculation of the similarity measure. I validate the empirical usefulness of the textual similarity measure in a finance context.

1. Introduction

There have been many proposed measures of document similarity but the most commonly used measure is that based on "cosine similarity." While the mathematical construction of cosine similarity for text documents is straightforward, the calculation of the measure on real-world documents is non-trivial. This is especially true when the size of the corpus to be analyzed makes calculation on a single local machine impractical. Therefore, I describe an implementation using a cloud-based distributed platform. Namely, the Elastic Compute Cloud (EC2) from Amazon Web Services (AWS).

2. Cosine Similarity Measure

The construction of textual similarity, SIM, between two sets of accounting policy disclosures involves three components. First, denote the set of all words occurring in the "Summary of Significant Accounting Policies" descriptions in a year is Ω_t. We measure a scalar, $J_t \equiv ||\Omega_t||$, that is, the number of unique words used in the descriptions of both firms. For each firm i in year t, an ordered binary vector W_{it} is constructed such that each element $W_{it}^{(j)}$ of W_{it} is one if and only if word j occurs in the given firm's text in the given year. Let $N_{it} \equiv W_{it}/||W_{it}||$, the normalized vector of word occurrences for each firm-year. For each year, construct a similar vector U_t of size J_t in which each element $U_t^{(j)}$ is equal to the sum of the number of occurrences of each word j among both sets of text. From this vector, compute a measure of the aggregate change in text from year $t-1$ to year t as the change in the number of times the word was used. Denote this aggregate word use change variable as $D_{t-1,t}$.

Formally,

$$D_{t-1,t} \equiv \left|\left| \sum_j (U_{j,t} - U_{j,t-1}) \right|\right|. \tag{1}$$

The textual similarity of a firm's accounting policy disclosures in a year is constructed as

$$SIM_{it} \equiv \frac{W_{it}}{||W_{it}||} \cdot \frac{D_{it}}{||D_{it}||}, \tag{2}$$

which is the cosine of the angle between the firms' vectors of word occurrences and the aggregate word change vector.

In subsequent analysis I make use of a measure of a firm's *self* similarity,

$$SELF\text{-}SIM_{it} \equiv 1 - \frac{W_{it}}{||W_{it}||} \cdot \frac{W_{it}}{||W_{it}||}, \tag{3}$$

which attempt to measure the degree to which the firm is changing its description of its own accounting policies from one year to the next. By construction, the similarity measure is on the closed interval from zero to one.

3. Implementation

3.1. Data Retrieval from EDGAR

Number of files File size distribution (Mb, tokens) Arrangement of files into years, industry pairs Distribution of pairs for which cosine similarity will be calculated

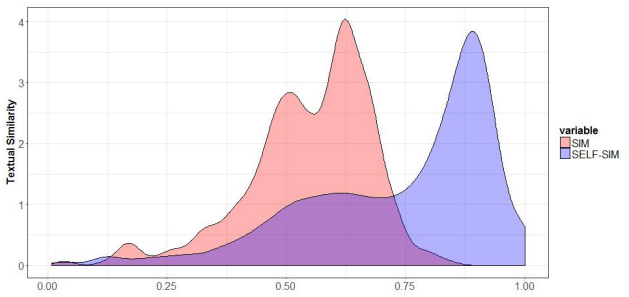

Figure 1. (Top panel) Time series trends of accounting comparability measure SIM. (Bottom panel) Pooled cross-sectional distribution of SIM.

3.2. Cluster Setup

3.3. Preprocessing

To more accurately analyze the 10-K sections I perform several preprocessing steps on the corpus. Specifically, I remove alphanumeric sequences and punctuation, convert all words to lowercase, and perform stemming and lemmatization on each document. Stemming and lemmatization are methods from the fields of computational linguistics and natural language processing that are designed to normalize textual corpora for automated analyses. Stemming requires to conversion of a word token to its root, removing pluralization and conjugation. Lemmatization is designed to convert word tokens that are different but often used to refer to the same physical entity to a common token.

Below is an example from the 2011 annual report of IBM. The filing originally contained the line:

> Additionally, changes to noncontrolling interests in the Consolidated Statement of Changes in Equity were $(29) million, $8 million and $(1) million for the years ended December 31, 2011, 2010 and 2009, respectively The accounts of variable interest entities (VIEs) are included in the Consolidated Financial Statements, if required.

After the pre-processing steps described above, the line became:

> addit chang noncontrol interest consolid statement chang equiti million million million year end decemb respect account variabl interest entiti vie includ consolid financi statement requir

Table 1. Univariate Determinants of Accounting Comparability. This table presents univariate regression results of SIM on firm characteristics. SIM_{t-1} indicates the one-year lag of similarity. All other variables are as defined in Appendix A.

	Coeff	Std Err	t	R^2
AGE	0.00	0.00	2.05	0.02
AT	0.00	0.00	6.08	0.16
BASPREAD	-0.33	0.05	-6.37	0.17
MVE	0.00	0.00	6.38	0.17
PE	-0.00	0.0	-2.40	0.02
SP500	-0.01	0.00	-5.51	0.13
TANG	-0.01	0.01	-1.85	0.01
TURNOVER	0.00	0.0	10.48	0.46
SIM_{t-1}	0.65	0.01	111.73	38.33

To reduce further the amount of irrelevant information contained in each piece of text extracted from the 10-K, I weight each word token that occurs via a method known as term frequency-inverse document frequency (tf-idf). This weighting scheme reduces the influence of overly common terms in a text corpus on the analysis of that corpus. These steps should reduce the amount of noise in the construction of accounting similarity.

3.4. Calculation of Cosine Similarity

4. Datasets and Validation

Publicly traded firms in the United States are required to file detailed annual reports, which are subsequently made available to the public. These annual reports, known as "10-K's" from the Securities and Exchange Commission form number of the report, must contain descriptions of the operations of the firm, financial results, and the accounting treatment used to prepare the financial statements. By accounting treatment, it is meant the set of allowed accounting rules chosen by the firm to recognize revenues and expenses and to measure assets and liabilities. Firms discuss these accounting treatments in the section of the 10-K known as the "Summary of Significant Accounting Policies," which itself is a subsection of the "Notes to the Financial Statements" (henceforth "the notes"). The accounting policies section is typically the first subsection of the notes, which themselves typically are the first section after the financial statements (balance sheet, income statements, statement of cash flows, and statement of shareholders' equity).

In this paper I find novel evidence that the accounting policies section contains significant predictive content

in an financial context. Namely, I find that the similarity of these texts is predictive of the behavior of large institutional investors. Institutional investors are large investors who often trade on behalf of a large number of individuals. Such institutions make up the vast majority of dollars invested in public equity markets.

5. Results

6. Related Work

7. Conclusions

Acknowledgments

This work is based, in part, on my doctoral dissertation at the University of Michigan Ross School of Business. I would like to thank my dissertation committee chairs Raffi Indjejikian and Venky Nagar and members Allison Earl and Robert F. Dittmar. I am also grateful for the advice and comments of Ryan Ball, Greg Miller, Cathy Shakespeare, Christopher Williams, and Reuven Lehavy, as well as workshop participants at the University of Michigan. I appreciate the financial support of the Ross Doctoral Fellowship, the Paton Accounting Fellowship, and the University of Michigan Rackham Merit Fellowship.

References

Admati, AR. 1985. "A Noisy Rational-Expectations Equilibrium for Multi-Asset Securities Markets." *Econometrica*, 53(3): 629–657.

Almeida, Heitor, and Murillo Campello. 2007. "Financial Constraints, Asset Tangibility, And Corporate Investment." *Review of Financial Studies*, 20(5): 1429–1460.

Ayers, BC, J Jiang, and PE Yeung. 2006. "Discretionary accruals and earnings management: An analysis of pseudo earnings targets." *Accounting Review*, 81(3): 617–652.

Badrinath, SG, JR Kale, and HE Ryan. 1996. "Characteristics of common stock holdings of insurance companies." *Journal of Risk And Insurance*, 63(1): 49–76.

Bajo, Emanuele, Massimiliano Barbi, Marco Bigelli, and David Hillier. 2013. "The role of institutional investors in public-to-private transactions." *Journal of Banking & Finance*, 37(11): 4327–4336.

Beatty, AL, B Ke, and KR Petroni. 2002. "Earnings management to avoid earnings declines across publicly and privately held banks." *Accounting Review*, 77(3): 547–570.

Beneish, M. D. 2001. "Earnings Management: A Perspective." *Managerial Finance*, 27: 3–17.

Bhojraj, S, and CMC Lee. 2002. "Who is my peer? A valuation-based approach to the selection of comparable firms." *Journal of Accounting Research*, 40(2): 407–439.

Bradshaw, Mark T., Gregory S. Miller, and George Serafeim. 2009. "Accounting Method Heterogeneity And Analysts' Forecasts." *Working Paper*.

Bushee, BJ, and CF Noe. 2000. "Corporate disclosure practices, institutional investors, and stock return volatility." *Journal of Accounting Research*, 38(S): 171–202.

Bushee, Brian J., and Theodore H. Goodman. 2007. "Which Institutional Investors Trade Based on Private Information About Earnings and Returns?" *Journal of Accounting Research*, 45(2): 289–321.

Campbell, John L., and P. Eric Yeung. 2016. "Earnings Comparability As A Signal of Earnings Quality And Future Stock Returns: Evidence From Peer Firms' Earnings Restatements." *Working Paper*.

Cascino, Stefano, and Joachim Gassen. 2015. "What Drives The Comparability Effect of Mandatory Ifrs Adoption?" *Review of Accounting Studies*, 20(1): 242–282.

Chung, Kee H., and Hao Zhang. 2011. "Corporate Governance and Institutional Ownership." *Journal of Financial And Quantitative Analysis*, 46(1): 247–273.

Chung, KH, and SW Pruitt. 1994. "A Simple Approximation of Tobins-Q." *Financial Management*, 23(3): 70–74.

Dechow, Patricia M., and Douglas J. Skinner. 2000. "Earnings Management: Reconciling The Views of Accounting Academics, Practitioners, And Regulators." *Accounting Horizons*, 14(2): 235–250.

DeFond, Mark, Xuesong Hu, Mingyi Hung, and Siqi Li. 2011. "The impact of mandatory IFRS adoption on foreign mutual fund ownership: The role of comparability." *Journal of Accounting & Economics*, 51(3): 240–258.

De Franco, Gus, S. P. Kothari, and Rodrigo S. Verdi. 2011. "The Benefits of Financial Statement Comparability." *Journal of Accounting Research*, 49(4): 895–931.

Easley, David, and Maureen O'hara. 2004. "Information and the Cost of Capital." *The Journal of Finance*, 59(4): 1553–1583.

Eaton, Tim V., John R. Nofsinger, and Abhishek Varma. 2014. "Institutional Investor Ownership and Corporate Pension Transparency." *Financial Management*, 43(3): 603–630.

Falkenstein, EG. 1996. "Preferences for stock characteristics as revealed by mutual fund portfolio holdings." *Journal of Finance*, 51(1): 111–135.

Fang, Vivian W., Mark Maffett, and Bohui Zhang. 2015. "Foreign Institutional Ownership and the Global Convergence of Financial Reporting Practices." *Journal of Accounting Research*, 53(3): 593–631.

FASB. 2010. "Qualitative Characteristics of Accounting Information."

Ferreira, Miguel A., and Pedro Matos. 2008. "The Colors of Investors' Money: The Role of Institutional Investors Around The World." *Journal of Financial Economics*, 88(3): 499 – 533. Darden - {JFE} Conference Volume: Capital Raising in Emerging Economies.

Gompers, PA, and A Metrick. 2001. "Institutional investors and equity prices." *Quarterly Journal of Economics*, 116(1): 229–259.

Granger, Clive, and P. Newbold. 1974. "Spurious Regressions In Econometrics." *Journal of Econometrics*, 2(2): 111–120.

Grossman, SJ, and JE Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review*, 70(3): 393–408.

Healy, Paul M., Amy P. Hutton, and Krishna G. Palepu. 1999. "Stock Performance And Intermediation Changes Surrounding Sustained Increases In Disclosure*." *Contemporary Accounting Research*, 16(3): 485–520.

Healy, Paul M., and Krishna G. Palepu. 2007. *Business Analysis and Valuation: Using Financial Statements.* . 4th ed., Thomson South-Western.

Hellwig, MF. 1980. "On the Aggregation of Information in Competitive Markets." *Journal of Economic Theory*, 22(3): 477–498.

Hoberg, Gerard, Gordon Phillips, and Nagpurnanand Prabhala. 2014. "Product Market Threats, Payouts, And Financial Flexibility." *The Journal of Finance*, 69(1): 293–324.

Hribar, P, and NT Jenkins. 2004. "The effect of accounting restatements on earnings revisions and the estimated cost of capital." *Review of Accounting Studies*, 9(2-3): 337–356. Conference on Accounting Disclosure and the Cost of Capital, UCLA, Los Angeles, CA, SEP, 2003.

Huang, Jiekun. 2015. "Dynamic Liquidity Preferences of Mutual Funds." *Working Paper*.

Kim, Jeong-Bon, Leye Li, Louise Yi Lu, and Yangxin Yu. 2016. "Financial Statement Comparability And Expected Crash Risk." *Journal of Accounting And Economics*, 61(2–3): 294 – 312.

Kim, O., and RE Verrecchia. 1994. "Market Liquidity and Volume Around Earnings Announcements." *Journal of Accounting & Economics*, 17(1-2): 41–67.

Koller, Tim, Marc Goedhart, and David Wessels. 2005. *Valuation: Measuring and Managing the Value of Companies.* . 6th ed., McKinsey & Company, Inc.

Leary, Mark T. 2009. "Bank Loan Supply, Lender Choice, And Corporate Capital Structure." *The Journal of Finance*, 64(3): 1143–1185.

Lee, Charles M. C., Paul Ma, and Charles C. Y. Wang. 2015. "Search-based peer firms: Aggregating investor perceptions through internet co-searches." *Journal of Financial Economics*, 116(2): 410–431.

Lehavy, Reuven, and Richard G. Sloan. 2008. "Investor recognition and stock returns." *Review of Accounting Studies*, 13(2-3): 327–361. Conference on Uses of Accounting Data for Firm Valuation and Performance Measurement, Arizona State Univ, WP Carey Sch Business, Tempe, AZ, OCT, 2007.

Maffett, Mark. 2012. "Financial Reporting Opacity And Informed Trading By International Institutional Investors." *Journal of Accounting And Economics*, 54(2–3): 201 – 220.

Merton, RC. 1987. "A Simple Model of Capital-Market Equilibrium with Incomplete Information." *Journal of Finance*, 42(3): 483–510.

OECD. 2014. "Annual Survey of Large Pension Funds And Public Pension Reserve Funds."

Ostrower, John. 2016. "Boeing's Unique Accounting Method Helps Improve Profit Picture."

Shane, Philip D., David B. Smith, and Suning Zhang. 2014. "Financial Statement Comparability And Valuation of Seasoned Equity Offerings." *Working Paper*.

Srivastava, Anup. 2014. "Selling-Price Estimates In Revenue Recognition And The Usefulness of Financial Statements." *Review of Accounting Studies*, 19(2): 661–697.

Stuart, Alix. 2008. "Why VSOE Spells Trouble." *CFO Magazine*.

Wang, J. 1993. "A Model of Intertemporal Asset Prices Under Asymmetric Information." *Review of Economic Studies*, 60(2): 249–282.

Yaniv Grinstein, Roni Michaely. 2005. "Institutional Holdings And Payout Policy." *The Journal of Finance*, 60(3): 1389–1426.

Yip, Rita W. Y., and Danqing Young. 2012. "Does Mandatory IFRS Adoption Improve Information Comparability?" *Accounting Review*, 87(5): 1767–1789.

Young, Steven, and Yachang Zeng. 2015. "Accounting Comparability and the Accuracy of Peer-Based Valuation Models." *Accounting Review*, 90(6): 2571–2601.