Diversification Improvements Through News Article Co-occurrences

John Robert Yaros and Tomasz Imieliński
Department of Computer Science
Rutgers University
Piscataway, NJ 08854-8019
Email: {yaros,imielinski}@cs.rutgers.edu

Abstract—Intuition suggests that a set of companies mentioned in the same news article are more likely to be related than unrelated. For instance, an article discussing a retailer would more probably mention its competitors or supply chain partners than mention other companies with no economic connection. Correspondingly, we consider using news article co-occurrences as a means to determine company relatedness. We show that companies mentioned together frequently are more likely to have higher future stock-return correlation, and consider using this data source as a means to achieve portfolio diversification by avoiding having pairs of related companies in the portfolio. We find this approach reduces risk and can be used to improve standard approaches to diversification that use expert-defined industry taxonomies, seeking to avoid portfolio concentration in any given economic sector.

I. INTRODUCTION

Two main techniques in risk management are hedging and diversification. With hedging, risk in an investment is reduced by making additional investments that should generate a time series of returns that is the opposite of the original investment. That is, hedging can be achieved by finding assets with greatest anti-correlation, or, alternatively, by finding assets with greatest correlation and using those assets for short sales. Conversely, diversification reduces risk by investing in assets with correlation closest to zero. If a single asset in a portfolio loses value, the likelihood of losses in the other assets remains constant, whereas if they were correlated, the likelihood of simultaneous losses is increased.

In equity investing, diversification is often achieved by investing in stocks in different “sectors” or “industries.” This approach matches the old proverb “don’t put all your eggs in one basket,” since investments will be spread across different areas of the economy. It is well-established that such an approach will help to reduce risk. Furthermore, industry diversification has increased in importance as other means of diversification have weakened, such as country diversification, due the integration of the global economy.

At the same time, what defines industries is not straightforward and the creation of taxonomies to assign companies to industries is generally left to experts. These experts must consider multiple factors, including similarities in goods produced, inputs to production, market perceptions, sensitivity to consumer income, etc. They must compile all such information to be able to produce a reasonable taxonomy that accurately groups similar companies.

This article focuses on a similar aggregation of information, but by utilizing a separate set of experts: news writers. When a writer creates a news article that contains a company or set of companies, that writer generally communicates rich information about the companies, albeit only a sliver of the total information. In particular, they often convey much information about the relatedness of companies. Consider the following snippets from New York Times articles involving Wal-Mart (an American discount retailer with many stores in various countries):

1) The [advertising] agency changes were part of a strategy shift at Wal-Mart, the nation’s largest retailer, to compete more effectively against rivals like Kohl’s, J. C. Penney and Target.

2) Procter & Gamble, the consumer products company, reached an agreement yesterday to acquire the Gillette Company, the shaving-products and battery maker, ... The move is a bid by two venerable consumer-products giants to strengthen their bargaining position with the likes of Wal-Mart and Aldi in Europe, which can now squeeze even the largest suppliers for lower prices.

3) BUSINESS DIGEST ... Federal regulators wrapped up the first set of public hearings on Wal-Mart’s request to open a bank, but gave scant indication of how they might rule on the company’s application. ... Stocks tumbled as strength in the commodities market fed inflation fears and stifled investors’ enthusiasm over upbeat first-quarter earnings from Alcoa.

In the first snippet, several companies are identified as competitors to Wal-Mart. In a diversified portfolio, it would make sense to avoid large positions in several of these stocks because the companies face similar risks. For instance, a drop in consumer spending would likely affect all retailers.

In the second snippet, we see that Procter & Gamble and Wal-Mart hold different locations in the same supply chain. While the article clearly mentions a battle between the companies to extract more value in the supply chain, the profitability of each company is again linked to similar risks, such as a drop in consumer spending.

In the third snippet, we see Wal-Mart is mentioned together with Alcoa (a producer of aluminum), but there is no real
relation between the companies presented in the article, other than the fact they had notable events occurring on the same day and, therefore, appear together in a business digest.

Previous research (e.g., [8]) has found that strong signals of company relatedness can be captured by simply examining the co-occurrences of companies, despite the presence of “noise,” such as in the case of the third snippet above. Yet, much of the existing research offers validation by simply drawing or listing company relations for the reader to approve, or by demonstrating some agreement with existing industry taxonomies. In contrast, this article seeks to validate the approach by applying news article co-occurrences to the fundamental task of diversification. We demonstrate that these articles can improve diversification beyond what is achieved by typical uses of industry taxonomies. For instance, in the first snippet above, Target is listed as a competitor to Wal-Mart. The correlation of average daily returns for Target and Wal-Mart is 1.78 times the average correlation of pairs of stocks for years 2004 to 2010. Yet, the Global Industry Classification System (GICS), a leading commercial industry taxonomy, placed Target and Wal-Mart in two separate sectors, Consumer Discretionary and Consumer Staples, respectively, for those years. A portfolio constructed to make sure that no stocks appear in the same sector could still contain these two highly related companies. By using news co-occurrences, we can avoid such situations.

Additionally, we show that as more news articles are included, signals become more robust. Given the abundance of not only news articles, but corporate filings, blog posts and other publicly available textual information, we believe more relationships can be captured, perhaps even for small or new companies that are not in industry taxonomies.

The remainder of the article is composed as follows. We begin by describing related work in section II. In section III we describe our datasets and outline the tools and heuristics used to identify companies in news articles and link them to stock price and sector/industry datasets. In section IV we describe our measure to quantify company relatedness through news and evaluate its predictiveness of pairwise company correlation. This measure is used in section V to select portfolios that are more diversified and, consequently, have less risk. We conclude in section VI.

II. RELATED WORK

The study of news and finance have intersected on a number of topics, including the speed of investor reactions to news stories [9] and the effects of media coverage (or lack thereof) on stock prices [10], [11]. Another area is sentiment analysis, which has been applied to measuring the impact of pessimism on stock price and volumes [12]. Sentiment analysis and use of other textual features have further been applied to create numerous signals for trading strategies [13], [14], [15]. [16]. Yet, those threads of research tend to focus on the use of news to predict the movements of single stocks, sectors or markets, rather than the relatedness and consequential co-movements of stocks.

The smaller thread of research more similar to this work is the use of news and textual data to extract inter-company relationships. In a seminal work [8], authors use ClearForest software (a pre-cursor to the Calais service used in this article) to extract entities from a corpus of business news articles, which are then cleaned for deviations in company naming (i.e., I.B.M. vs IBM). Authors then visualize the data with a network structure where edges are drawn between company vertices wherever the count of co-occurrences exceeds a set threshold. Authors highlight clusters in the network that appear to match common perceptions of industries, then develop a notion of “centrality” to measure a company’s importance to an industry. They further develop a cosine measure for the inter-relatedness of two pre-designated industries based on relatedness of their respective companies, as determined by news co-occurrences. As authors concede, the work is limited by its small set of articles covering only four months in 1999, where the technology bubble led to significantly high numbers of articles containing technology companies. Further, theresults rely on the reader to judge whether the industries and relatedness measurements are reasonable, rather than offering verification through external measures of company similarity, such as stock-return correlation.

Other researchers have also considered use of news or other textual information to determine various aspects of relatedness between companies. In [17], authors construct a network derived from news articles appearing on Yahoo! Finance over an eight month period. If the same article appears under the news pages for two different companies, a link is constructed between the two companies in the network. Multiple articles increase the weight of each link. Authors then use in-degree and out-degree measures as features for binary classifiers that seek to predict which company in a pair has higher revenue. In [18], authors similarly construct networks based on co-occurrences in New York Times articles, but instead study the evolution of networks over time and use network features along with regression models to predict future company profitability and value. In [19], authors suggest bank interdependencies can be inferred from textual co-occurrences, rather than the two traditional data sources, co-movements in market data (e.g., CDS spreads), which are not always efficient, and interbank asset and liability exposures, which are generally not publicly disclosed. They exemplify their approach using a Finnish data set and examine the temporal changes in the network, particularly following the Global Financial Crisis. In [20], a method for extracting competitors and competitive domains (e.g., laptops for Dell and HP) is presented that essentially uses a search engine to gather articles and then uses some sentence patterns to identify competitors. In [21], authors describe a system that extracts companies and corresponding relationships from news articles using predefined rules. They suggest an approach to detect events, such as acquisitions, by observing changes in the strength of relationships over time.

To our knowledge, this work is the first to consider use of news in stock-portfolio risk management. In particular, the task of diversification is a well-suited application because it relies on correlation, which is a direct measure of the relatedness of two companies through their stock price co-movements. Further, whereas taxonomies are often regarded as “gold standards” in the literature, we take the view that improvements can be made beyond what they achieve in risk management and in determining company relatedness.
Fig. 1. Articles Per Year

III. DATA

Our collection of news articles is taken from two corpora at the Linguistic Data Consortium (LDC). The first is the New York Times Annotated Corpus, which contains over 1.8 million articles published by the New York Times (NYT) from January 1, 1987 to June 19, 2007. The second is English Gigaword Fourth Edition, which contains articles from the following five newswire services:

- AFP: Agence France-Presse
- APW: Associated Press Worldstream
- CNA: Central News Agency of Taiwan
- LTW: Los Angeles Times / Washington Post
- XIN: Xinhua News Agency

The data collection at LDC for most of the newswires was performed via dedicated lines that recorded article text and meta-data in real-time, although some articles were later received (or recovered) in bulk. Due to various collection issues, there are large gaps in the collection for particular newswires. There are also periods where fewer articles were collected due to changes in collection methods. Fig. 1 depicts the number of articles from each source, per year.

To extract company names from articles, we use Calais, which is developed and maintained by ClearForest, a group within Thomson Reuters. The free OpenCalais web service allows users to submit text and receive back annotations. Company name detection is a main feature, which we use here.

To quantify OpenCalais error rates, we randomly selected 100 NYT articles and manually marked companies in the text. We then computed precision and recall as shown in Table 1.

Table 1. OpenCalais Performance on 100 Sample Articles

<table>
<thead>
<tr>
<th>No. Companies in Text</th>
<th>True Positives</th>
<th>False Positives</th>
<th>False Negatives</th>
<th>F1 score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>288</td>
<td>213</td>
<td>34</td>
<td>75</td>
<td>0.796</td>
<td>0.862</td>
<td>0.740</td>
</tr>
</tbody>
</table>

Calais does reasonably well at the difficult task of identifying companies, including differentiating those companies from non-profit organizations (e.g., Red Cross) or governmental agencies (e.g., Air Force). Some examples of false positives are shown in Fig. 2 where the words alleged to be companies by OpenCalais are outlined in boxes. In example (1), a region was misidentified as a company. Examples (2) & (3) demonstrate a common problem, wherein only part of the company name is identified as a company, or other words are combined into the company name (possibly from other companies). We did not attempt to perform any correction, again with the expectation that such errors will amount to noise and enough articles will overcome any problems. The worst case occurs when a false positive has the name of an actual company. For example, if the fruit “apple” was somehow identified by OpenCalais as the technology company, Apple. We never observed such situations, although it is possible they did occur. For most false positives, the misidentified text would not be included in our analysis because it simply would not link to our universe of stocks.

Our stock universe is the constituents of the S&P 100, which is an index maintained by Standard and Poor’s (S&P). The index is a subset of 100 U.S. companies from the S&P 500 chosen for sector balance and such that the companies are larger, more stable and have listed options [22]. To identify the S&P 100 constituents each year, we use Compustat, a product of S&P.

We further use Compustat to get sector assignments for the S&P 100 stocks in the Global Industry Classification System (GICS), which is a leading commercial industry taxonomy produced by S&P and MSCI/Barra. GICS has been shown to offer better risk management characteristics than many historically-used government or academic taxonomies [23]. We focus on the 10 sectors, but GICS offers greater granularity by breaking the sectors into 24 industry groups, then 68 industries and finally 154 sub-industries. These counts reflect the state of GICS at the end of 2010, but they have changed over time to match changes in the economy (although there have always been 10 sectors). Additionally, a company’s GICS assignment may be updated for changes in the company’s business. For these reasons, it is important to use concurrent GICS assignments when performing historical analysis. Accordingly, Compustat offers these historical assignments and we use them in our analysis.

Fig. 2. Examples of False Positives

1) Ray Sullivan, Northeast regional director for American College Testing, ...
2) Specialty Products and Insulation Co., East Petersburg, ...
3) ... Credit Suisse First Boston/Goldman, Sachs & Co.
For stock returns, we use a dataset from the Center for Research in Security Prices (CRSP). The dataset includes daily returns that not only include returns due to daily price changes, but also all changes due to corporate actions, like cash dividends.

CRSP and Compustat were linked using a variety of identifiers, such as CUSIP and exchange tickers, with attention to ensuring the identifiers are matching at concurrent time periods. Any unmatched constituents from the S&P 100 at any time period were manually linked.

The more significant task was linking the news article companies because the same company may be referenced in multiple ways. For instance, “DuPont,” “El DuPont,” “E.I. Du Pont De Nemours” and “E.I. du Pont de Nemours and Company” are all aliases for the same company. Calais does offer company “resolutions,” where these multiple aliases are resolved to a single company name and Reuters Instrument Code (RIC). However, these resolutions do not account for the time period of the article. For example, “Mobil” will resolve to ExxonMobil, which is not helpful if the article is prior to the 1998 merger of Exxon and Mobil. Therefore, we use only the original company names identified by Calais, not their resolutions.

To link the identified companies to the other datasets, we use a manually constructed mapping of company aliases to CRSP permno (CRSP’s unique identifier). Each entry of the mapping includes beginning and ending dates of validity to avoid mapping to incorrect companies for the given article’s date, such as in the ExxonMobil example above. We further use a series of standardization rules on the raw company names, such as removal of periods, consolidation of abbreviation characters (i.e. “A. T. & T.” → “AT&T”), removal of suffixes (i.e., remove “Co,” “Company,” “Inc,” etc.) and multiple other rules to reduce the possible number of alias derivatives for a given company.

Finally, an important note is that we do not consider subsidiaries in determining co-occurrences. A main reason is that joint ownership of a subsidiary (by two or more companies) is frequent and may easily obscure the strength of a co-occurrence in news. For example, suppose 32% of Hulu is owned by NBCUniversal, which in turn is wholly owned by Comcast. Further suppose there is a co-occurrence in news between Hulu and Google. What should be the strength of the relationship between Google and Comcast? What about the other owners of Hulu? We leave this for future work.

\[
\cos(C_1, C_2) = \frac{|W_{c_1} \cap W_{c_2}|}{\sqrt{|W_{c_1}| \cdot |W_{c_2}|}}
\]

where \(W_{c_1}\) and \(W_{c_2}\) are the sets of articles for companies \(c_1\) and \(c_2\), respectively. For two vectors, the cosine measures the angle between them, but not the magnitude. Likewise, the cosine for two sets reflects the similarity between them rather than the sizes of the sets \([23]\). This favorable property makes cosine appropriate for our dataset where some companies may have a much greater volume of news than others.

Another favorable aspect of cosine similarity is that it has the “null addition property,” wherein adding or removing
unrelated data does not affect the measure. More precisely, increasing or decreasing the number of articles that contain neither c1 nor c2 will not alter C(c1, c2). Many other measures, such as Collective strength, odds ratio, All-confidence, etc., do not have the null addition property [25].

In Table II we exemplify the similarity ranking for a single company by showing the top 15 similar companies to Target in 2005, as measured by cosine similarity C. For reference, Target was in 189 articles in 2005 and was in the Consumer Discretionary (“Discret.”) GICS sector. As seen in the table, the top companies ranked by the news are intuitively related to Target. Wal-Mart directly competes with Target as a discount retailer. Limited Brands, Home Depot and Radioshack are also retailers, but with different focuses and/or store sizes than Target. Black & Decker and P&G produce goods that are frequently sold in Target stores. Wal-Mart is selected as most similar, even though GICS has no indication of similarity as it places the company in a different sector, Consumer Staples (“Staples”). We also see that the top stocks tend to be those most correlated with Target in the subsequent year, 2006. As the strength of the cosine weakens, so do the intuitive relatedness of the companies and their correlation.

### A. Correlation Prediction

Following Modern Portfolio Theory (MPT) [26], risk can be measured by the variance of returns. Diversification can best be achieved by including assets with correlation close to zero. We first measure how well news can predict correlation in order to validate our intuition that stocks mentioned together in news are more correlated. Since correlations vary greatly among stocks (and sectors), we normalize correlation for each stock by dividing by its average daily correlation with all other stocks.9

\[
\psi_{i,j} = \frac{\rho_{i,j}}{\bar{\rho}_i}
\]

where \(\rho_{i,j}\) is the correlation between stocks \(i\) and \(j\) and the average correlation \(\bar{\rho}_i\) for stock \(i\) is

\[
\bar{\rho}_i = \frac{\sum_{j \in U, j \neq i} \rho_{i,j}}{|U| - 1}
\]

where \(U\) is the universe of stocks.

### V. Diversification

To quantify the risk of a given portfolio, we use a simple method of counting the number of ordered pairs of stocks in the portfolio that are expected to be highly similar through news. More specifically, we count the number of occurrences of an ordered pair \((c_1, c_2)\) where \(C(c_1, c_2)\) is in the top \(K\) highest values for the given stock \(c_1\). Observe that two stocks \(\{c_a, c_b\}\) may be counted twice - once for \(C(c_a, c_b)\) and once for \(C(c_b, c_a)\). From Fig. 5 we can see news is best at predicting roughly the top 5-10 most similar stocks related to a given stock, so we set \(K=10\) in the following experiments.

For each year in our dataset, we randomly generate 1,000,000 portfolios of five stocks each from the S&P 100. Since each stock in the portfolio may itself have varying levels of risk, we weight each stock in the portfolio by the inverse of its historical standard deviation of 20-trading-day returns (approx. one calendar month) over the previous three years. That is, the weight \(w_i\) of stock \(i\) in the portfolio is proportional to \(1/\sigma_{h,i}\), where \(\sigma_{h,i}\) is the historical standard deviation of

### Fig. 5. Performance of Correlation Prediction by News

Since we ultimately wish to use news for prediction, we use walk-forward testing [27], [28] where we split our data into years and use news articles from each year to predict correlation in the subsequent year. Results are depicted in Fig. 5. Each point along the curves indicates the correlation ratio \(\psi\) (y-axis) between each stock and its corresponding k-th similar stock at the given rank of news similarity (x-axis). The figure displays the results averaged over all stocks and over all years (1995 to 2008), but each individual year follows the same general pattern. For stocks that are most highly similar by news (left-most in the figure), the correlation is highest. The correlation rapidly diminishes with the first 5-10 stocks, then has a slower downward trend towards lower correlation. This indicates that news is strongest at determining the most correlated companies. It can still differentiate at lower levels, but at weaker strength. Also evident in the inset of the figure is that as more articles are included, the better news does at determining the most correlated stocks. The smallest source, CNA, has weakest predictive power, while the larger APW does better. Using all articles does the best.
We compute the following "risk improvement" measure

\[ R = \frac{\sigma_A}{\sigma_H} \]

where \( \sigma_A \) is the actual standard deviation of 20-trading-day returns of the weighted portfolio over the subsequent 12 periods (essentially monthly returns over the next year). The "homogeneous" standard deviation \( \sigma_H \) is a theoretical value that measures the variance if the stocks had been perfectly correlated (i.e., \( \rho_{ij} = 1 \) for all pairs of stocks \( i, j \)).

\[
\sigma_H^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_i \sigma_j \rho_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_i \sigma_j
\]

Of course, most stocks are not perfectly correlated, so \( R \) will range between zero and one with lower values indicating less risk. A value of 0.5 means that the risk (as measured by variance) is half the risk that would be encountered if the stocks were perfectly correlated.

Results are shown in Table III where we display the risk improvement measure, \( R \), averaged over all sample portfolios for the given year. We also show averages for several subsets of the portfolios, selected by the news cosine, GICS sectors, or the combination method that will be described shortly.

The risk improvements measures for all portfolios are substantially less than 1.0, which is unsurprising given that stocks are rarely perfectly correlated and it is well-known that including more stocks in a portfolio quickly reduces risk [29], [30]. Even using only five stocks offers a dramatic risk improvement as seen in the second column ("All Portfolios") of the table.

For news articles, we see that having fewer instances of pairs of stocks in the top K similarity by news leads to lower risk. This is evident in the table, but also in Fig. 6 where the values are plotted relative to random (i.e., the "All Portfolios" column). We plot values in this ratio since the random risk reduction changes greatly from year to year and using the ratio provides a more stable view. As seen in the figure, portfolios with fewer top K similar stocks by news tend to have lower risk. For most years, portfolios with zero top K similar stocks have least risk. Years 1999, 2000 and 2007 are exceptions. Still, zero has lower risk than one to three pairs of top K stocks at the 5% level of significance under a paired t-test.
In Table III and in Fig. 2, results are shown for GICS. Our expectation is that the more sectors the portfolio spans, the more diversified the portfolio should be. Indeed, results indicate this is nearly always true. In Fig. 7, we can see that portfolios with stocks in five sectors should have least risk. Correspondingly, at the 0.1% level of significance under a paired t-test, a portfolio spanning five sectors has less risk than a portfolio spanning only four sectors. In general, we see that this GICS approach to diversification offers a more consistent risk reduction than the news approach.

Still, news can offer some benefit by improving GICS results. The final column (“News / GICS Combination”) in Table III displays results for portfolios that span five GICS sectors and have zero top K similar pairs of stocks by news. This method has least risk in many years and only does worse than both the GICS and news methods in 2005.

Since the use of variance has been criticized as a measure of risk [31], we consider an alternative: expected shortfall (ES), which we compute as the average of the worst 5% of portfolio returns. This measure is also sometimes called “expected tail loss.” Unlike the standard deviation of returns, expected shortfall is more interpretable since it tells the investor “there is a 1 in 20 chance your returns will be $X, on average (although they can be worse).” The measure is usually applied to select portfolios based on risk, but it can also be used to evaluate portfolio selection methods, as we do here.

For our ES measure, since each randomly generated portfolio may contain stocks that are more volatile than others, we (de)lever each portfolio such that they all have equal volatility in terms of the historical variances of their stocks, specifically the standard deviation of the thirty-six periods of 20-trading-day returns (i.e., roughly three years of monthly returns). The investment for a portfolio with average volatility is $10,000 and the amount invested will be lower or higher for other portfolios based on whether they are more or less volatile, respectively. Results are shown in Table IV where the expected shortfall is shown per year for the given diversification methods. We find that the combination method performs only marginally better than GICS alone.

However, by increasing K and making the news article portfolios more selective, the margin between the combination method and GICS increases. This is shown in Fig. 9 for expected shortfall and in Fig. 8 for the risk improvement measure. R. In the figures, the best portfolios for news (zero news top K pairs), GICS (5 sectors) and the news / GICS combination (satisfying both) are shown, averaged over all years. For both the ES and R risk measures, we see risk improvements in the news and combined portfolios as K increases. (GICS remains constant because it is unaffected by K.) At the same time, the number of possible portfolios quickly decreases, as shown in Fig. 10. Hence, trade-off for selecting K is improving risk at the cost of fewer alternatives.

VI. SUMMARY AND FUTURE DIRECTIONS

We evaluate the intuition that companies that frequently co-occur in news articles are related. We consider a large set of news articles from several major services over a 1994 to 2007 period and use OpenCalais to identify company names in the articles. To quantify company relatedness through co-occurrences, we use cosine similarity because it is well-suited for the large disparities in the number of articles each company appears in. We find that when ranking companies by their cosine similarity to a given company, the top ranked companies tend to have highest future stock-return correlation with that given company. We suggest a simple portfolio diversification method, wherein potential portfolios are avoided if they contain any pairs of stocks that have high news-article cosine similarity. Results show that such an approach tends to reduce risk as measured by variance in returns. We find that a standard approach of avoiding stocks in the same sector (as defined by GICS) is generally more reliable in reducing risk. However, combining the approaches, such that similar stocks as determined both by news and by GICS sectors are avoided, does even better at reducing risk.
Several opportunities exist to be more exact in extracting company relatedness in the articles beyond the approach used in this work. In [18], authors suggest considering the proximity of the company names in the news article may be beneficial. For example, more weight should be given to occurrences in the same sentence. Further, using natural language processing to try to actually identify the relationship (competitor, supply chain partner, etc.) between two companies may help to determine the future interrelationship of their stock returns. Conversely, identifying and reducing the weight of “noise” articles, such as market summaries, that do not present strong relationships may help in improving the similarity signals.

Finally, whereas this article focuses on large companies because industry taxonomies are available to use for baselines, greater potential may exist when examining smaller, new or private companies. In such situations it may be difficult to determine company similarity because traditional methods, such as industry taxonomies or even historical prices, are unavailable. Still, there may be a large desire to be able to identify “comparables” for use in company valuation or hedging. News articles and other textual documents, such as blog posts, may be able to fill this void, particularly because they are becoming more prevalent.

ACKNOWLEDGEMENTS

Authors thank Kenneth Liau and Gerard O’Neill for their contributions during preliminary experiments.

REFERENCES