Feature Selection Methods For Understanding Business Competitor Relationships

Rahul Gupta
U. of Southern California
gupt507@usc.edu

Jay Pujara
U. of Southern California
jpujara@isi.edu

Craig A. Knoblock
U. of Southern California
knoblock@isi.edu

Shushyam M. Sharanappa
U. of Southern California
maligesh@usc.edu

Bharat Pulavarti
U. of Southern California
pulavart@usc.edu

Gerard Hoberg
U. of Southern California
hoberg@marshall.usc.edu

Gordon Phillips
Dartmouth College
gordon.m.phillips@dartmouth.edu

ABSTRACT

Understanding competition between businesses is essential for assessing the likely success of new ventures or products, for making decisions before investing capital in new businesses, and understanding the impacts of regulatory policy. One important resource for analyzing competitor relationships are business webpages, which can capture the mission, products, services, and key markets associated with a company. However, webpages also contain irrelevant, extraneous, or misleading text, hampering prediction. To address this challenge, predictive models use a process known as feature selection to identify only relevant terms. The diversity and specificity of business domains pose a challenge for automated approaches for feature selection. In this paper, we compare two approaches to feature selection: manually-curated lists of terms provided by experts and automated approaches to feature selection. We evaluate several approaches to feature selection and their impact on predicting competitor relationships, demonstrating that carefully designed automated feature selection approaches can surpass the performance of manually-curated word lists by 10%.

Keywords

Business Competitor Relationships, Feature Selection, Competition Graphs, Natural Language Processing, SEC Filings

1. MOTIVATION

Learning competitor relationships between businesses can enable new avenues of research in domains such as finance, management, and entrepreneurship. Research topics such as the evolution of competitive relationships over time, the impact of financial downturns on competitive structure, and the efficacy of government policy changes each require an understanding of the competitive landscape between businesses. However, precise characterizations of competitor relationships are difficult to define since companies may offer many products and services, participate in diverse markets, and evolve over time. In this paper, we contrast two data-driven approaches to that use textual analysis to predict competitor relationships using differing amounts of domain expertise.

Traditional approaches to understanding competitor relationships have focused on classifying companies into business sectors, with the expectations that companies in the same sector are competitors. Sector-based competitor relationships are easy to understand but have several drawbacks, particularly that sectors and company classifications must be updated as new industries and business models develop. In contrast, foundational microeconomic models of competition support the idea that competition between firms will manifest through product offerings that are similar but serve as inexact substitutes [7, 3].

In our work, we adopt this definition of competition, and define competitor relationships between pairs of companies based on the similarity of the products and services they offer. Capturing the similarity of products and services for a broad population of companies can be difficult, and automated, data-driven approaches are necessary to enable such analyses to be conducted at scale. Hoberg and Philips [6] developed such an approach, producing a resource for publicly-traded firms based on analyzing the text of SEC 10-K filings and using textual similarity as an indicator of competition. One limitation of this approach is coverage: only publicly traded firms are required to provide such filings, and the textual descriptions of products and services in these filings may be limited or incomplete.

One alternative to address this drawback is to use company websites [5] for deriving data to determine these competitor relationships, since both publicly-traded and privately-held companies provide information about products and services on the World Wide Web. However, the use of web data poses challenges as web pages contain substantial noise that hinders the process of identifying relevant business-related

DSMM'18, June 15, 2018, Houston, TX, USA

ISBN 978-1-4503-5883-5/18/06…$15.00

DOI: https://doi.org/10.1145/3220547.3220550

information. Apart from the main content, a web page often contains extraneous sections such as navigation panels, copyright and privacy notices, and advertisements for business purposes or sitemaps. Each of these sections can contain irrelevant information that can harm the performance of classification algorithms.

A key problem for using web data is identifying relevant content for analysis. In machine learning applications, this problem is referred to as "feature selection". Our main focus in this work is performing feature selection by identifying a list of relevant terms for competitor relationship prediction. We explore two different approaches to this problem: (1) using in-domain text (10-K filings) and extending this glossary using expert knowledge, and (2) using out-of-domain text (web pages) and automatically selecting relevant features using statistical metrics.

2. RELATED WORK

Prior approaches for identifying competitor relationships range from rigid industry classifications to more flexible approaches using textual analysis. One traditional approach is manually curated industry classifications, such as SIC [12] and NAICS[11] classifications. These industrial classifications are based on fixed industry definitions or production processes, both relying on domain expertise to arrive at classifications. One drawback to industrial classifications is the inflexibility of the definitions of industrial sectors, which cannot capture the full complexity of modern businesses, which can participate in multiple sectors, or conversely be monopolists with no clearly defined sectors.

A different approach to identifying competitors uses differing financial metrics such as risk profile, revenues, margins, return on assets or equity, valuation, and debt leverage [4, 1] to classify companies. These approaches are founded on the expectation that competing firms will have similar performance and capital structure, as measured by financial metrics. One limitation of such approaches is that they still produce static classifications to which companies are assigned, providing a more rigid definition of competition. Furthermore, these approaches are only applicable when the companies publicly report financial results so that financial metrics can be computed.

A third approach to identifying competitors attempts to qualify the products and services each company offers, or using correlations between companies in consumer behavior. These approaches include the proprietary GICS system, self-reported products and markets [13], web-search coreferences [9], or textual analysis of company SEC filings [6] or preliminary approaches using web pages [5]. These approaches match our approach most closely, and each makes design decisions and derives appropriate metrics to capture products and services. However, in most cases the source data of these approaches is relatively noise-free (such as SEC filings or financial data). In order to scale these approaches to the broader company information available across diverse web data sources, a stronger model of relevancy is required, an argument we advance in the description of our method.

3. APPROACH

The fundamental challenge of using a noisy, diverse corpus of textual data, such as company web pages, is identifying the relevant terms for analysis. In domains where substantial labeled data (e.g., competitor relationships) is available, the relevance of terms can be automatically determined. However, in the absence of labeled competitor relationships it is necessary to consider alternative approaches to determining which terms on webpages are most indicative of competitor relationships.

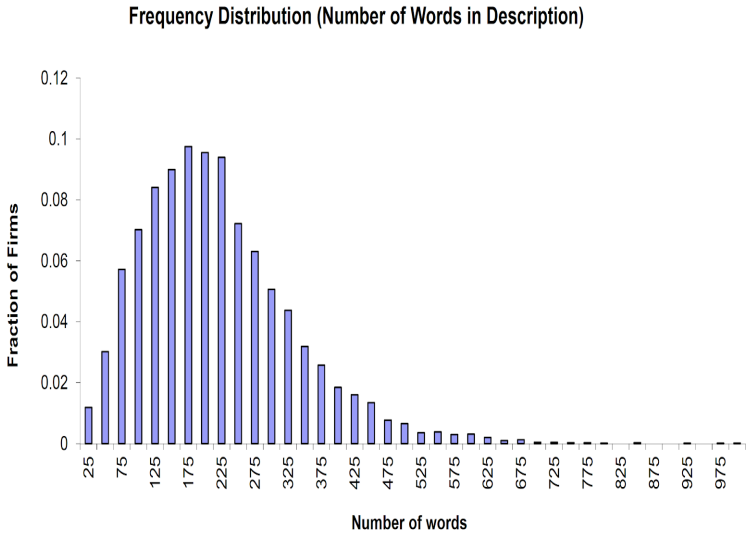

Figure 1: Frequency distribution of unique words in 10-K product descriptions.

We pose the task of feature selection for predicting competitor relationships as one of constructing a list of relevant terms for capturing information about products, services, markets, and processes. As a proxy for labeled data clearly indicating which terms will be relevant, we explore two different options for generating such a list. The first technique, introduced in subsection 3.1, relies extensively on domain knowledge to identify relevant features. The second technique, introduced in subsection 3.2, uses statistical measures on a large corpus of web pages to select the most meaningful terms.

3.1 Manually Curated Term Lists

The first method we consider for assembling a list of relevant terms relies extensively on domain knowledge. First, domain experts identify a restricted set of documents that are known to contain relevant terms describing the business activities of companies. Next, terms are extracted from those documents and automatically filtered to remove overly-frequent words. Finally, an expert manually reviews terms to add missing relevant terms and remove irrelevant terms. We describe this process in more detail below.

The number and diversity of web pages make manual analysis of content difficult. To allow a more tractable foundation for curating word lists, our approach begins with a different corpus, SEC 10-K filings. In prior work [6], domain experts identified the business description section of 10-K filings as a source of relevant terms about products and services. They applied additional filters, using only nouns and proper nouns, and removing geographical names and frequent terms appearing in more than 25% of descriptions. As in previous work, we filter business descriptions by re-

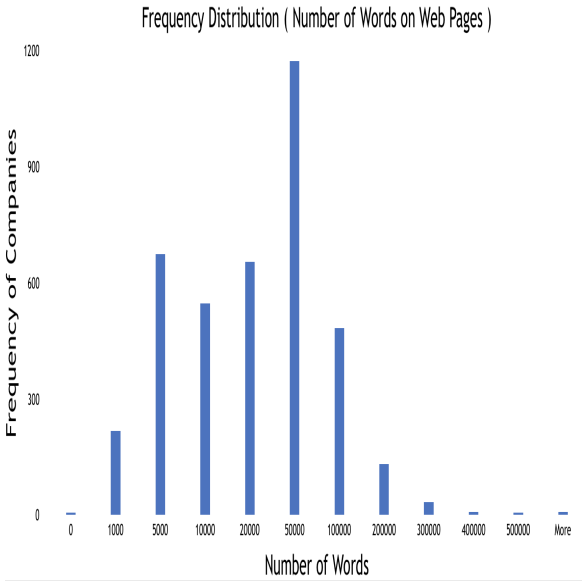

Figure 2: Frequency distribution of unique words on web pages of companies

moving the most frequent terms and including only noun phrases. By restricting our analyses to 10-K filings, we are able to create an initial highly relevant word list for feature selection.

Using 10-K filings to determine relevant words has potential drawbacks. The primary data source for our system is company web pages, which we anticipate to be substantially different from the text of 10-K filings. For example, 10-K descriptions are far more terse, using fewer words. In an analysis of the 10-K corpus and our web corpus, we found that the average company would use hundreds of words in the business description section, whereas a company's web pages use tens of thousands of words. Figure 1 depicts a histogram of unique word counts in 10-K reports while Figure 2 shows a similar histogram for company web pages. Furthermore, 10-K business descriptions show less variation in the words used, with 200K unique words across all companies, relative to 1.7M unique words used in company web pages. We also hypothesize that business descriptions are often expressed in more formal language, and may use more general terms than the product descriptions found on web pages. As a result, we attempt to extend the word list generated from the out-of-domain regulatory filings with terms from in-domain web pages.

Our approach to revising the word list generated from regulatory filing was structured in terms of two additional lists. The first list was a "whitelist" that identified missing terms that were judged to be highly relevant for predicting competitor relationships. The second list was a "blacklist" of terms that were likely to be irrelevant and were excluded from analysis.

The whitelist was constructed by selecting frequent or discriminative terms generated from the business descriptions found on web pages and having a domain expert review these terms and classify them as "relevant" or "irrelevant". Relevant terms were included in the whitelist and irrelevant terms were discarded. This process allowed us to identify terms such as "*ethernetcarrier*", "*sleeper*", "*tumor*" which were missing from the initial word list. The whitelist was constructed by analyzing the discriminative words from the webpages that were not in 10-Ks. Each of these words were marked "valid" or "invalid" based on definition mentioned above. The words deemed to be "valid" were added to the whitelist.

The blacklist was constructed by examining the initial word list. Words judged to be unrelated to business activities in the context of a web page were excluded from future analysis. Using this process, words such as "*admiralty*", "*gardner*", "*steinberg*" were added to the blacklist.

The manual curation process described above yielded a high-precision set of terms relevant for defining business activities and predicting competitor relationships between businesses. The primary drawback for such an approach is the amount of time a domain expert must spend classifying terms to curate the word list. Furthermore, since business areas and processes continually change, a manual curation approach must be repeated regularly to ensure that current terms are included in the word list. These considerations suggest that an automated approach based on statistical measures can provide a promising alternative to manual curation, an idea we explore in the next section.

3.2 Statistical Measures of Term Relevancy

Natural language processing, information retrieval, and data mining applications are often confronted with a similar problem of determining relevant terms in documents and capturing the most important aspect of queries. A common approach to solving this problem is using statistical measures of word importance. The TF-IDF (term-frequency, inverse document frequency) statistic is among the most popular measures for determining term relevance. We provide a quick overview of TF-IDF and then detail some particular observations of applying TF-IDF to company webapges.

3.2.1 Term frequency

Term frequency captures the number of times each term, t, appears in each document, d. We define the term frequency based on an indicator function, $fr(x,t)$, for token x and term t, where the indicator is 1 when a token matches the term. Formally:

$$tf(t,d) = \sum_{x \in d} fr(x,t)$$

where the fr(x,t) is a function defined as:

$$fr(x,t) = \begin{cases} 1 & x = t \\ 0 & x \neq t \end{cases}$$

3.2.2 Inverse Document Frequency

One weakness of the term frequency measure is highly ranking frequent terms even when those terms lack discriminative power. To measure the specialization of terms, the inverse document frequency captures how many documents in the corpus use a particular term. By penalizing terms that occur in a large fraction of the documents, the IDF measure can help identify more unique terms. IDF is defined as the ratio of the number of documents in the corpus, $\|D\|$, to the number of documents containing a term t.

$$idf(t) = \log \frac{|D|}{1 + \sum_d f(t,d)}$$

where the f(t,d) gives the number of documents where the term t appears:

$$f(t,d) = \begin{cases} 1 & t \in d \\ 0 & otherwise \end{cases}$$

Multiplying the term frequency by the inverse document frequency provides a weight for each term in each document. The TF-IDF score is defined as: $tf - idf(t,d) = tf(t,d) \cdot idf(t)$. For a corpus-wide feature selection, we take the sum of all TF-IDF scores across the corpus, $tf - idf(t) = \sum_{d \in D} tf(t,d) \cdot idf(t)$. After computing these scores, we choose the top 15% of terms (by TF-IDF score) for inclusion in the glossary.

3.3 Computing Competitor Relationships

Using the feature selection techniques in the previous sections, we are able to identify relevant terms from each company's web pages. We use these terms to predict competitor relationships by measuring the overlap of terms for each pair of companies. We denote the terms in company i's webpage as T_i, and use the Jaccard index to assess the similarity of two companies:

$$sim(i,j) = \frac{T_i \cap T_j}{T_i \cup T_j}$$

. Computing the exact Jaccard index can be computationally expensive, so as an approximation, we use the MinHash data sketch to store the terms for each company and the Locality Sensitive Hashing [8, 10] technique to efficiently retrieve the most similar companies for each query company. We rank the pairwise similarities of all company pairs and choose the top 2% of these pairs as business competitors. In the next section, we describe our quantitative evaluation of these competitor relationships.

4. EVALUATION

Our experimental evaluation of the two different approaches to feature selection used approximately 1.9M web pages from nearly 4,000 publicly traded firms, which we describe in Section 4.1. In our experiments, we introduced additional post-processing on the TF-IDF score based on our observations of data inconsistencies, which we describe in Section 4.2. We describe our evaluation metric in Section 4.3 and discuss results and their sensitivity to parameter settings in Section 4.4.

4.1 Dataset

For our experiments, we created a list of the web addresses of 4,000 publicly traded firms using the Compustat database of financial firms. From this list of web domains, we collected 1.9 million individual web pages stored by the Internet Archive[1] in the year 2015. For each company domain, we chose up to 500 web pages for our analysis, prioritizing web pages by the depth of their URLs. For example, `https://www.google.com/about` would be prioritized above `https://www.google.com/about/stories` during the selection process. Each web page was processed to extract readable text using the BeautifulSoup library.[2] In all of our experiments, we use this corpus of processed webpages when computing competitor relationships.

4.2 Implementation Details

4.2.1 Manually Curated Term List

To determine a relevant term list for our curation-based approach, we extracted terms from the business description sections of companies 2015 10-K filings. We removed common stop words. We determined the part of speech for each term, and included only terms that were classified as nouns, proper nouns, or adjectives in the final list. A research assistant in a business school program assisted in selecting terms for inclusion in the whitelist and blacklist.

4.2.2 Statistically Selected Terms

We computed scores for terms occurring in company web pages, yielding scores for 1.7M unique terms. After the computation of TF-IDF scores, we noticed several abnormalities that we corrected through post-processing. These corrections include the removal of proprietary terms based on a minimum document frequency and removal of incorrectly extracted terms by enforcing a maximum word length.

One abnormality we noticed was that many top-scoring terms appeared frequently in very few company webpages, including "countsbaker", "geon", "ultratuf", "wilflex", and "oncap." To avoid selecting words that are strongly associated with a single company, signaling a proprietary term, we introduced a filter to remove terms that only occurring in a few companies. We discuss the impact of this parameter in more detail in Subsection 4.4. Anecdotally, after filtering proprietary terms occurring in only one or two companies, the highest-scoring terms based on the TF-IDF score were "blog", "accessories", "clinical", "shop", and "cloud."

Another observation based on our analysis of processed web pages was that extremely long terms that were artifacts of our data processing pipeline would often score highly on TF-IDF. Examples of terms that appeared to be the result of processing abnormalities included words such as "WeightedAverageNumberOfDilutedSharesOutstanding" which was present in companies including Footlocker (retail sales), Alsic (manufacturing), Alliqua (biotechnology), and Finjan (enterprise software). Since these terms appeared to lack discriminative power, we introduced another filter to remove terms that had a length of more than 20 characters, assuming that most terms of this form were due to abnormalities in the web processing libraries.

4.3 Evaluation Metric

Following the approach of Bhojraj et al. [2], we evaluate the effectiveness of our competitor relationships based on the ability to predict financial outcomes using these competitor relationships. Specifically, for each company c_i, we define a set of competitors or rivals, R_i based on the competitors defined in Section 3.3. We also define a financial metric of interest, F, such as asset-adjusted company profits (the ratio of profits to assets). We fit a regression model to estimate the financial metric, $\hat{F}$ based on the average metric value,

[1] https://archive.org

[2] https://www.crummy.com/software/BeautifulSoup/

$min(d)$	R^2
0	0.258
3	0.262
5	0.259
10	0.252

Table 1: Removing proprietary terms initially improves performance, but higher thresholds are harmful.

$maxlen(t)$	R^2
NA	0.262
17	0.284
20	0.286
25	0.285

Table 2: Removing long words improves performance, but the precise threshold has fairly limited impact.

top %	R^2
10	0.289
15	0.286
20	0.220

Table 3: As more terms are included by feature selection, performance can degrade.

$\overline{F(R_i)}$, of the rivals across all companies (learning λ and c):

$$\hat{F}(c_i) = \lambda\overline{F(R_i)} + c$$

We then evaluate a set of competitor relationships on the basis of the coefficient of determination:

$$R^2 = 1 - \frac{\sum_i (F(c_i) - \hat{F}(c_i))^2}{\sum_i (F(c_i) - \overline{F})^2}$$

A high R^2 suggests that the regression model successfully explains observed financial results using the results of rival companies.

4.4 Results

In this section we explore how the parameters introduced earlier, the minimum number of documents a term appears in, the maximum length of a term, and the top terms included in an automated feature selection approach, affect our overall results. We then choose a fixed set of parameters and compare the results of a manually curated list of terms to automated feature selection. Based on our preliminary findings, our automated feature selection approach outperforms a manually curated approach and remains robust across different parameter settings.

In Table 1 we compare the results based on choosing a minimum document threshold for terms. When terms occurring in only one or two documents are eliminated, the R^2 metric improves slightly, but increasing the minimum document threshold to 5 or 10 reduces performance. Based on this experiment, we set the minimum document threshold to 3 for subsequent experiments. We next investigate removing long words that appear to have been introduced by errors in document processing. Eliminating words with lengths greater than 17, 20, or 25 all seem to have a similar impact on performance, demonstrating a marked increase when these artifacts from document processing can be removed. Finally, we experiment with the number of top-scoring terms we include in an automatically generated term list. We find that both the top 10% and 15% of terms show similar performance, but as more terms are included performance begins to suffer.

Finally, we compare the results from a manually curated term list and one that is automatically generated term list in Table 4. We compute TF-IDF scores after applying a minimum document count threshold of 3, a maximum term length threshold of 20 and using the top 15% of terms. We found that the improvement using an automated feature selection approach increased the R^2 metric about 10%, suggesting that a data-driven, automated approach has advantages over more limited feature selection relying heavily on human curation.

Feature Selection	R^2
Manual	0.261
Automatic	0.286

Table 4: Automatic feature selection approaches improve performance about 10% over manual curation

5. CONCLUSION

As our experiments show, web pages can provide a powerful resource for determining competitor relationships. In this paper we contrasted two approaches for feature selection, which is necessary to overcome the inherent heterogeneity and noisiness of web pages. The first approach defined an initial set of relevant terms from in-domain text (business descriptions found in SEC 10-K filings) and extended this with manual curation. The second approach used a common statistical measure, TF-IDF, frequently used in natural language processing and information retrieval, and filtered the terms using three basic techniques. Comparing the competitor relationships derived after applying these feature selection methods, we found that the performance of the automated, statistical approach exceeded that of the manually curated term list by 10%. In future work, we plan to examine more powerful feature selection and text mining techniques for identifying competitor relationships, as well as scaling these techniques to the tens of millions of company web pages in our collected corpus.

Acknowledgements

This material is based on research sponsored in part by the National Science Foundation under Grant Nos. 1561057 and 1561068.

References

[1] S. Bhojraj and C. M. C. Lee. Who is my peer? a valuationâĂŘbased approach to the selection of comparable firms. *Journal of Accounting Research*, 40(2):407–439, 2002.

[2] S. Bhojraj, C. M. C. Lee, and D. K. Oler. What's my line? a comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, 41(5):745–774, 2003.

[3] E. Chamberlin. *A Theory of Monopolistic Competition*. Harvard University Press, 1933.

[4] E. F. Fama and K. R. French. Industry costs of equity. *Journal of financial economics*, 43(2):153–193, 1997.

[5] E. Heiden, G. Hoberg, C. A. Knoblock, P. Modi, G. Phillips, G. Raul, and P. Szekely. Web text-based network industry classifications: Preliminary results. In *SIGMOD Workshop on Data Science for Macro-Modeling with Financial anD Economic Datasets (DSMM)*, 2017.

[6] G. Hoberg and G. Philips. Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5), 2016.

[7] H. Hotelling. Stability in competition. *Economic Journal*, 39(153):41–57, 1929.

[8] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *ACM Symposium on Theory of Computing*, 1998.

[9] C. M. Lee, P. Ma, and C. C. Wang. Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, 116 (2):410 – 431, 2015. ISSN 0304-405X.

[10] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*. Cambridge university press, 2014.

[11] Office of Management and Budget. *North American Industry Classification System*. United States Census, 2017.

[12] E. Pearce. *History of the Standard Industrial Classification*. Bureau of the Budget, Office of Statistical Standards, 1957.

[13] J. D. Rauh and A. Sufi. Explaining corporate capital structure: Product markets, leases, and asset similarity. *Review of Finance*, 16(1):115–155, 2011.