2009

Avoiding the Blind Spots: Competitor Identification Using Web Text and Linkage Structure

Gautam Pant
The University Of Utah, gautam.pant@business.utah.edu

Olivia R.L. Sheng
The University Of Utah, olivia.sheng@business.utah.edu

AVOIDING THE BLIND SPOTS: COMPETITOR IDENTIFICATION USING WEB TEXT AND LINKAGE STRUCTURE

Completed Research Paper

Gautam Pant
Department of Operations and Information Systems
David Eccles School of Business
The University of Utah
1645 E Campus Center Dr., Salt Lake City UT 84112
gautam.pant@business.utah.edu

Olivia R. L. Sheng
Department of Operations and Information Systems
David Eccles School of Business
The University of Utah
1645 E Campus Center Dr., Salt Lake City UT 84112
olivia.sheng@business.utah.edu

Abstract

The importance of identifying competitors and of avoiding "competitive blind spots" in marketplace has been well emphasized in research and practice. However, identification of competitors is non-trivial and requires active monitoring of a focal company's competitive environment. The difficulty in such identification is amplified manifold when there are many more than one focal company of interest. As the web presence of companies, their clients/consumers, and their suppliers continues to grow, it is increasingly realistic to assume that the real-world competitive relationships are reflected in the text and linkage structure of the relevant pages on the web. However, finding the appropriate web-based cues that effectively signal competitor relationships remains a challenge. Using web data collected for more than 2500 companies of the Russell 3000 index, we explore the notion that web cues can allow us to discriminate, in a statistically significant manner, between competitors and non-competitors. Based on this analysis, we present an automated technique that uses the most significant web-based cues and applies predictive modeling to identify competitors. We find that several web-based metrics on an average have significantly different values for companies that are competitors as opposed to non-competitors. We also find that the predictive models built using web-based metrics that we suggest provide high precision, recall, F measure, and accuracy in identifying competitors.

Keywords: competitor identification, web metrics, data mining

Introduction

The importance of identifying competitors and of avoiding “competitive blind spots” (Zajac and Bazerman 1991) has been well documented in the literature (Walker et al. 2005). It can be challenging to continuously monitor a focal company’s environment to identify competitors. The challenge is greatly amplified when there are many more than one focal company of interest (e.g., analysis of a portfolio with hundreds of companies) (Ma et al. 2009). Several works have described the difficulties, both from cognitive and procedural standpoint, in identifying competitors. While some have highlighted the role of mental models and taxonomy (Porac and Thomas 1990) in identifying competitive threats, others have highlighted the “managerial myopia” (Bergen and Peteraf 2002) in recognizing competitors.

As the web presence of companies, their clients/consumers, and their suppliers continues to grow, it is increasingly realistic to assume that the real-world competitive relationships are reflected in the text and linkage structure of the relevant pages on the web. However, finding the appropriate web-based cues that effectively signal competitor relationships remains a challenge (Bao et al. 2008; Ma et al. 2009). Typically, competitor identification, which is a necessary first step for competitor analysis and strategy, is based on “supply-side” and “demand-side” considerations (Bergen and Peteraf 2002; Chen 1996; Desarbo 2006). For example, if two companies depend on similar resources and technology for production (say iron ore) then they may be considered competitors based on the supply-side substitutability. On the other hand if two companies produce products that address similar needs of the consumers (say personal computing) then they may be considered competitors based on demand-side substitutability. Web sites of companies can be expected to receive links from (i.e., in-links) and also link to (i.e., out-links) their supply-side or demand-side relations. Hence an overlap between in-links and out-links of two companies' web sites may be an indication of their substitutability and therefore a signal of potential competitor relationship. The content of the web site of a company provides a description of the company (e.g., its various products and services) by the company itself. An overlap between the self descriptions of companies could also be used as a signal of their substitutability. We study the extent to which competitor relationships are discernable from these and other web-based metrics.

Our study is based on more than 2500 companies and their competitors from the Russell 3000 index. We suggest three carefully crafted web metrics that require us to crawl hundreds of thousands of web pages across companies in our data set. We further obtain a list of more than a million URLs of pages that link to web sites of those companies. We suggest two additional control metrics that are derived from previous works in text and web mining for inter-company relationships. Using these web metrics we present a systematic study that explores the signal (if any) contained in these metrics that is relevant to the problem of automatic competitor identification. Our study makes the following main contributions:

1. We present three novel web metrics based on in-links, out-links and text of web sites corresponding to companies that may be used to measure the substitutability of companies. We find that these metrics, on an average, vary significantly between competitors and non-competitors hence validating their relevance to competitor identification problem.
2. This is the first study that uses a variety of web metrics based on linkage structure, web site content, online news and search engine results to feed predictive models for competitor identification. We find such predictive models to have high accuracy, precision, and recall.
3. We validate the predictive power of web metrics for competitor identification using a large data set that covers more than 2500 companies. The large data set allows us to make statistically robust arguments for the suggested metrics.

While the critical and challenging nature of competitor identification has been well emphasized, there is little literature on automatic competitor identification that involves no manual effort and utilizes publicly accessible data. Given that competitor identification remains a largely manual effort and the dynamic nature of global competitive environments poses additional challenges for accurate and timely competitor identification, there is a strong need for alleviating (at least partially) the complexity of the problem through automated tools. Our study hence provides a much needed systematic exploration of a variety of web metrics for automatic competitor identification.

Related Work

Several works in management literature have discussed the need for accurate identification of competitors and provided theoretical frameworks for that purpose. For example, Bergen and Peteraf (2002) suggest a broad framework for competitor identification through environmental scanning. It is a framework for mainly manual identification of competitors by managers. The competitor identification in their framework is performed on the "basis of similarities in terms of their [companies'] resource endowments and the market needs served" (Bergen and Peteraf 2002). These correspond to demand-side and supply-side considerations.

A few papers in the text and web mining literature have explored the idea of finding relationships between companies using news articles. The earliest work in this direction was by Bernstein et al. (2002) where they created virtual links between companies if they are mentioned in the same piece of news. A company that appears in large number of news stories with other companies will have a large number of links. Using link analysis, centrality of companies is measured and it is found that the 30 most central companies in the computer industry include several Fortune 1000 companies. Bernstein et al. (2003) use co-occurrence of stock tickers in news stories to identify connections between companies and utilize the resulting network to predict the industry membership of companies. More recently Ma et al. (2009) use the co-occurrence of stock tickers of companies in news stories to create connections between companies and use properties of the resulting network to predict competitor relationships. All of these works use just one source of information (business news) and are based on the assumption that co-occurrence of companies in news stories indicates a potential relationship between the companies.

Bao et al. (2008) described a competitor mining system that is based on the "observation" that competitors tend to co-occur in web pages (and hence search engine results) more often than non-competitor pairs. The authors provide no empirical support for this observation but utilize this observation/hypothesis in the design of their competitor mining system. Their proposed approach depends on just one data source (search engine results) and the evaluation is based on a very small set of companies (< 100). We feel that restricting web-based automatic identification of competitors (or other relationships) to a single data source (news or search engine) would limit the approach from gaining a broader view of companies' footprint on the web and hence limit its performance. As compared to previous works, we propose predictive models that utilize a variety of web resources that are publicly available. Also different from previous works, the web metrics suggested here are carefully explored with statistical tests to make robust observations about their behavior with respect to competitors and non-competitors before the metrics are incorporated within predictive models. We conduct our study on data collected for more that 2500 companies.

Test Bed

Our study is based on companies in the Russell 3000 index. For each of the companies listed in the index, we try to find the corresponding web site URL from Yahoo Finance. We also use the Hoovers[1] API to obtain a list of competitors for each of the companies. Since our analysis is restricted to Russell 3000 index companies, we consider only those competitors that are in the index as well. We are able to identify a data set of 16485 competitors (pairs) across 2694 companies. For each of the companies we identify an equal number of random (non-competitor) companies from Russell 3000 index to create a data set of 16485 non-competitors (pairs). The experiments and analysis will use either all or part of 32970 (16485 X 2) pairs of companies, half of which are competitors and the other half non-competitors.

Web Metrics

We suggest three web metrics that quantify the overlap in linkage structure and text of web sites associated with a pair of companies. In addition we use two other web metrics that are derived from previous works as control variables. Each of the five metrics can be easily computed for a pair companies. Next, we describe the computation of the suggested metrics.

[1] http://www.hoovers.com/

In-link Similarity

We would like to measure the pair-wise overlap or similarity in the web links that are directed towards web sites of companies of interest (that may or not be competitors). Hence for each of the companies in our test bed we obtain a list of top 500 in-link URLs using the Yahoo Boss API. The API provides programmatic access to the Yahoo search engine data. An in-link URL is the URL of a page that contains a hyperlink to the home-page on the web site (company) of interest. We omit in-link URLs that are from the same site as the web site of interest. We collect more than a million URLs of pages that link to web sites corresponding to the test bed companies.

Given a pair of companies (e.g., AMD and Intel in Figure 1), we now have a list of URLs from different web sites that have links to the two companies' web sites (e.g., pages on microsoft.com and apple.com contain links to home-pages of amd.com and intel.com). We note that many of these links may indicate supply-side or demand-side relationships. For example, a web site providing a forum for end-users of computer processors that links to AMD's web site is indicative of demand-side connections. Hence if the same web forum site also links to Intel's web site it may be indicative of the two companies' demand side substitutability and hence competition. However, it is important to gauge the strength and relevance of these connections. For example, Wikipedia has links to both AMD and Intel web sites. Is that indicative of their competitive relationship? Wikipedia links to a majority of companies in our test bed and hence its links are probably less discriminative and hence less relevant in terms of identifying a competitive relationship. In contrast, linuxinsider.com links to both AMD and Intel sites and it does not link to many other companies' web sites. However, this web site has only a single URL/link (among the in-link URLs collected) to AMD and Intel sites. While these lone links may be relevant to the competitive relationship, their strength (due to low number of links) is very week.

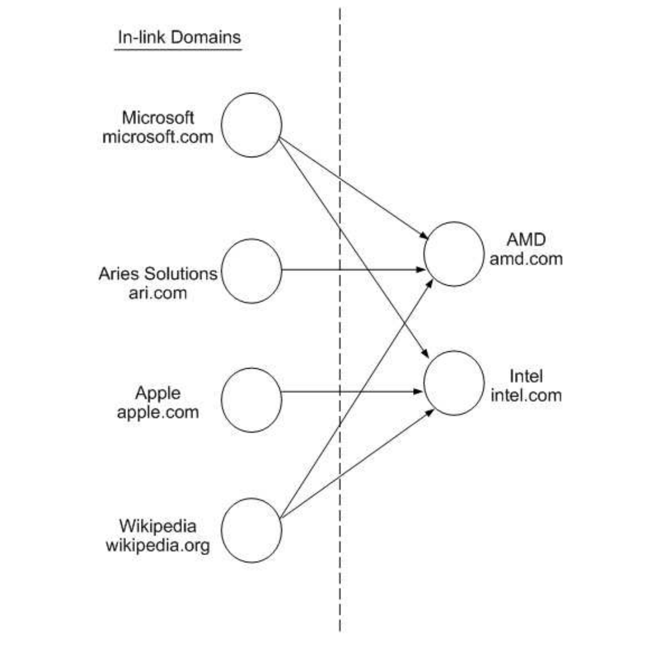

Figure 1. Example of computing in-link similarity based on overlapping in-link domains for AMD and Intel (does not show all in-link domains)

To quantify the concepts described above, we first extract the *in-link domains* which are the second-level domains (e.g., microsoft.com for http://www.microsoft.com/) for each of the in-link URLs. We then count the number of times an in-link domain appears among the in-link URLs for a given web site (company). We refer to this count as *domain frequency or DF*. For example, 10 URLs from wikipedia.org point to amd.com and hence the DF of wikipedia.org for AMD is 10. DF is intended to measure the strength of connection between an in-link domain and the company web site. We further count the number of companies for which a given in-link domain appears among the companies' in-link URLs. We refer to this count as *company frequency or CF*. For example, wikipedia.org appears as an in-link domain for 1658 companies in our data set and hence its CF is 1658. Due to the high CF of Wikipedia, the fact that two companies' home-pages are linked from Wikipedia pages is a weak indicator at best of their competitor relationship. Note that CF of an in-link domain is the same across companies while DF changes with the company (web site) of interest. Table 1 shows the top 10 in-link domains based on their CF.

Table 1. Top 10 domains by their company frequency computed using in-links

Domain	**Company Frequency (in-links)**
yahoo.com	2807
mffais.com	2295
google.com	2278
msn.com	2018
Prnewswire.com	1978
goldmood.com	1976
blogspot.com	1973
forbes.com	1970
finviz.com	1923
businessweek.com	1876

It is clear that some domains link to a large number of companies (web sites). Hence, in-links from high CF domains have lower relevance in terms of indicating competitor relationship. In any computation of in-link similarity of two companies we would need to weigh the in-link domains based on their CF values. Such a weighting function of CF would need to be an inverse of CF where higher CF values lead to lower weight for the in-link domain. We suggest the following formulation, which we call *inverse company frequency* or *ICF*, for the weight function which is analogous to the popular inverse document frequency or IDF measure in information retrieval:

$$ICF = \ln(\frac{N_c}{CF}) \qquad (1)$$

where N_c is the total number of companies over which CF of an in-link domain is computed. ICF increases with decreasing values of CF. ICF can be also written as $-\ln(\frac{CF}{N_c})$ where $\frac{CF}{N_c}$ can be seen as the probability of an in-link domain having a link to a company. As the probability of an in-link domain appearing among the in-links of companies increases it becomes less discriminating and hence its ICF decreases.

Each company is represented as a vector $\vec{w_c} = [w_{c1}, w_{c2}, ..., w_{cn}]$ of in-link domain weights where w_{cj} is the weight for an in-link domain j for company c and n is the total number of in-link domains across companies. The weight w_{cj} for domain j is computed as a product of DF and ICF:

$$w_{cj} = DF \times \ln(\frac{N_c}{CF}) \qquad (2)$$

Again, Equation 2 is analogous to the TF-IDF term weight formulation in information retrieval (Salton and McGill 1983). Once we have the vector representation for each company, we find the in-link similarity between two companies a and b as the cosine of the angle (or cosine similarity) between the two corresponding vectors as follows:

$$sim(a,b) = \frac{\overrightarrow{w_a} \cdot \overrightarrow{w_b}}{\left\|\overrightarrow{w_a}\right\| \cdot \left\|\overrightarrow{w_b}\right\|} \qquad (3)$$

The above cosine similarity measures the normalized overlap between in-link domains of two companies while accounting for differing weights of each domain.

Out-link Similarity

The out-link similarity between two companies is measured in a manner similar to the in-link similarity. However, the web links used for the computation are not the links coming into web sites of interest but those going out from the sites (or out-links). For the purpose of gathering out-links, we crawl up to first 200 pages from the web sites of each of the companies in our test bed. We use a multi-threaded web crawler that starts from the home-page of each company and follows the links outwards in a breadth-first manner to download pages. The crawler follows only those links that are from the same domain as the company web site and hence avoids downloading pages that are external to the company.

After downloading web pages from each company's web site, an HTML parser is used to identify all of the out-links appearing in those pages. For a given company, a list of out-links is created that corresponds to all web links that lead to pages outside of the company's web site (domain). Figure 2 shows an example of two investment firms and some of their *out-link domains*. Similar to in-link domains, out-link domains are second-level domains corresponding to out-links of company web sites.

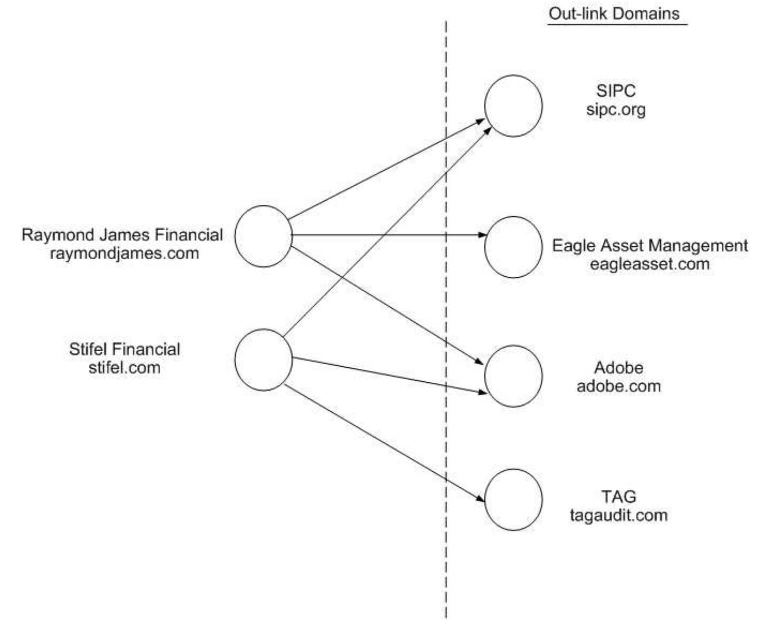

Figure 2. Example of computing out-link similarity based on overlapping out-link domains for two brokerage firms (does not show all out-link domains)

We may expect some of the overlapping out-link domains between two companies to be indicative of their competitive relationship. In Figure 2, both of the firms of interest (left hand side) have a link to SIPC's web site.

This may indicate that both of the firms provide products for investors since SIPC provides investor protection (demand-side substitutability). The issue of strength and relevance of the connection between companies of interest and their out-link domains is the same as that discussed for in-link similarity computation. For example, both of the firms in Figure 2 have a link to adobe.com, but this overlap may not be indicative of their competitive relationship. A large number of firms in our test bed have an out-link to some page on adobe.com. Hence we can define company frequency or CF in this context as the number of companies for which a given out-link domain appears among the companies' out-link URLs. The CF for adobe.com is 1116. Table 2 shows the top 10 out-link domains by company frequency computed using the out-links. Again, the need for appropriately discounting the high CF domains is clear. We use the ICF function defined in Equation 1 (but now based on CF values computed using out-links) for weighing different domains.

As in the case of in-link similarity, we represent each company using a vector of weights, where each weight corresponds to an out-link domain and the weight is computed using Equation 2 (note that DF and CF are computed using the out-links). Finally, the out-link similarity between two companies is measured as the cosine similarity (see Equation 3) between the out-link weight vectors of the two companies.

Table 2. Top 10 domains by their company frequency computed using out-links

Domain	Company Frequency (out-links)
corporate-ir.net	1394
adobe.com	1116
sec.gov	619
yahoo.com	565
microsoft.com	501
google.com	435
shareholder.com	402
10kwizard.com	370
wsw.com	303
apple.com	300

Text Similarity

Web sites of companies describe various aspects of their business. The text in the first few pages on a company's web site can be considered as a type of self-description by the company. As explained in the previous section, we have crawled up to the first 200 pages from each company's web site. We now concatenate the text from these crawled pages and use it as a self-description of the company. Two companies that have similar demand-side and supply-side considerations will tend to describe themselves similarly. Hence we treat the similarity in the self-description of companies as an indicator of their competitive relationship. To measure the similarity between self-descriptions of two companies we use a standard TF-IDF representation (Salton and McGill 1983) from information retrieval. The words in each of the self descriptions are identified. We remove stop words or common words such as "and", "or", "the" etc. and also ignore numeric or alphanumeric words (e.g., "67.2", "a12"). We compute the weights of each of the remaining words as follows:

$$t_{kc} = (0.5 + \frac{f_{kc}}{\max_{k' \in T_c} f_{k'c}}) \times \ln(\frac{|E|}{d_k}) \qquad (4)$$

where t_{kc} is the weight of the word k in self-description of company c, f_{kc} is the frequency of the word k in the self-description of the company c, T_c is the set of words appearing in the self-description of company c, d_k is the document frequency of the word k, and $|E|$ is the set of pages over which document frequencies are computed. Document frequency of a word is the number of pages in which the word appears. We compute document frequencies over all of the pages crawled across all of the companies in our test bed. A word that tends to appear in a large number of pages can be considered to be less specific and hence probably less meaningful while measuring the similarity between self-descriptions of two companies. Equation 4 is a standard TF-IDF formulation (Salton and McGill 1983). Each of the self-descriptions is represented as a vector of word weights and their similarity is measured through cosine similarity as depicted in Equation 3.

News Count

If two companies are mentioned in the same piece of news, it may indicate some type of connection between the two. This has been a fundamental assumption behind some previous works that have utilized such co-occurrence to suggest web metrics that may indicate inter-company relationships (Bernstein et al. 2002) such as competitive relationships (Ma et al. 2009). While not developing any sophisticated metrics, we use a metric that utilizes the fundamental assumption that co-occurrence in news stories indicates a potential inter-company relationship. In particular, using the Yahoo Boss API, we obtain the number of new stories in which two companies' names co-occur. We call this measure *name news count*. For this purpose, we first canonicalize the company names by removing words such as inc and corp, and then concatenate the two names to create a keywords that is searched against news stories using the Yahoo Boss API. The API returns the number of news pages that match the keyword (this data is similar to the count of results that typically appears at the top of search engine result pages). Thus we obtain the number of news stories with the two company names. As an additional measure we also count the number of news stories that contain the tickers of the two companies (instead of names). We call this measure *ticker news count*.

Search Engine Count

Bao et al. 2008 utilize co-occurrence of company names among search engine results as a fundamental step in their process of mining competitor relationships. Hence, in a manner similar to news count, we query Yahoo Boss API with company names and tickers of the two companies of interest and obtain *name se count* and *ticker se count* as metrics that we will utilize in our analysis. Again, the API returns the data on the number of web pages that match the keyword containing the company names (or tickers).

We note that while metrics based on news count and search engine count have been utilized in the past, to the best of our knowledge, we are the first to suggest the metrics of in-link similarity, out-link similarity, and text similarity for identifying competitor (or any other) relationship between companies. Hence we will use the news count and SE count as control variables to verify if the other three metrics provide predictive power beyond these control variables. This is also the first work to combine such a wide variety of web metrics for the problem of competitor identification.

Empirical Analysis

While suggesting the above mentioned web-based variables we hypothesize (although implicitly) that these variables behave differently for companies that are competitors versus those that are not. We now explicitly state several hypotheses that we would like to test to explore the presence or absence of discriminating signal (competitor versus non-competitor) contained in these variables:

H1: The average in-link similarity between competitors is greater than between non-competitors

H2: The average out-link similarity between competitors is greater than between non-competitors

H3: The average text similarity between competitors is greater than between non-competitors

H4a: The average name news count for competitors is greater than for non-competitors

H4b: The average ticker news count for competitors is greater than for non-competitors

H5a: The average name se count for competitors is greater than for non-competitors

H5b: The average ticker se count for competitors is greater than for non-competitors

We test these hypothesis with the web metrics computed for the 16485 competitors (pairs of companies) and 16485 non-competitors (pairs of companies). Table 3 shows the results of the various t-tests. The table also shows the average values of the various web metrics for competitors and non-competitors along with the respective standard errors. All of the hypotheses other than H4b and H5b are supported at the 0.01 significance level. H5b is supported at the 0.05 level while H4b can be rejected. In other words, the three metrics suggested by us on an average behave differently for competitors versus the non-competitors. This is also true for the news and search engine count metrics (using company names) that are derived from fundamental assumptions of previous literature dealing with text or web mining for discovering inter-company relationships.

Table 3. Results of hypothesis testing for various web metrics along with the average values of the metrics for competitors and non-competitors

Hypothesis	Class	Mean	Std. Error	Sig. (2-tailed)
H1: in-link similarity	Competitor	0.0493	0.001	0.000
	Non-competitor	0.008	0.000	
H2: out-link similarity	Competitor	0.020	0.001	0.000
	Non-competitor	0.013	0.001	
H3: text similarity	Competitor	0.0613	0.001	0.000
	Non-competitor	0.026	0.000	
H4a: name news count	Competitor	26.65	1.813	0.000
	Non-competitor	3.96	.564	
H4b: ticker news count	Competitor	254.16	114.655	0.121
	Non-competitor	72.69	22.653	
H5a: name se count	Competitor	77761.55	4755.712	0.000
	Non-competitor	17945.25	1565.988	
H5b: ticker se count	Competitor	218893.84	23464.719	0.041
	Non-competitor	146673.22	26537.293	

Given that H4a and H5a are more significantly supported than H4b and H5b, we will use just the company names-based news and search engines count for further analysis and leave out the ticker-based counts.

Table 4 shows the pairwise correlation between the five web metrics identified for further analysis. We find that almost all of the correlations are weak (<0.2) other than a moderate correlation (0.53) between news count and se count. Hence the web metrics suggested by us potentially capture different aspects of competitor relationship.

Table 4. Pairwise correlation between the 5 web metrics

	in-link similarity	out-link similarity	text similarity	news count	se count
in-link similarity	1				
out-link similarity	0.151	1			
text similarity	0.173	0.111	1		
news count	0.017	-0.002	0.037	1	
se count	0.011	-0.011	0.050	0.530	1

Predictive Models

Our empirical analysis suggests that the five web metrics identified earlier may act as good predictors for competitor identification. Using the web metrics computed for competitors and non-competitors pairs (companies), we build predictive models where inputs to a model are the 5 web metric values and the output is a class label C (competitor) or NC (non-competitor). We train the models using 66% of randomly selected data (pairs) and test it on the remaining data. To maintain the disjoint nature of the training and testing data sets we remove the pairs of competitors from the testing data whose reverse[2] instances appear in the training data. We also maintain the 1:1 ratio of competitors to non-competitors in both the training and the testing data (see the next section for experiments with more skewed data sets). We use two different predictive modeling techniques: C4.5 decision tree and logistic regression. These modeling techniques are popular in the data mining and econometric modeling literature. We repeat the training and testing process 50 times using each of the modeling techniques. In other words, we divide the overall data into training-testing data sets 50 times and each time we train the modeling techniques using the training data and observe their performance on the testing data. Hence we obtain 50 observations on the performance of each modeling technique. Our results are based on average testing data performance from the 50 repeated experiments for each of the modeling techniques.

The performance of a modeling technique is measured using the following measures:

1. Average Precision: The precision of a given technique is the fraction of company pairs identified as competitors by the technique that are actual competitors. The precision is computed using the testing data for each test-train run and the average precision is computed over the 50 runs.
2. Average Recall: The recall of a given technique is the fraction of actual competitor pairs that are correctly identified as competitors by the technique. The recall is computed using the testing data for each test-train run and the average recall is computed over the 50 runs.
3. Average F Measure: For a given technique, precision and recall measures often present a tradeoff (i.e., attempts to increase precision can lead to lower recall and vice-versa). Therefore F measure, a harmonic mean of precision and recall, is popularly used to integrate precision and recall into one measure as follows:

[2] "Exxon is a competitor of Chevron" is the reverse instance of "Chevron is a competitor of Exxon". This not always true since competitor relationship can be asymmetric.

$$F = \frac{2 \times P \times R}{P + R} \qquad (5)$$

where P is precision and R is recall for the given technique. The F measure is computed using the testing data for each test-train run and the average F measure is computed over the 50 runs.

4. Average Accuracy: The accuracy of a given technique is the fraction of company pairs that are correctly classified as competitors or non-competitors. The accuracy is computed using the testing data for each test-train run and the average accuracy is computed over the 50 runs.

To understand the additional predictive power provided by the three web metrics suggested by us over the two control variables that have been previously suggested in the literature, we build three predictive models using the control variables (one variable at a time as well as the two variables together). The models use C4.5 decision tree technique. The performance of these *control models* is also measured and reported. Table 5 reports the results corresponding to the different models. Since the data contains 50% competitor pairs and 50% non-competitors, the prior for the data set is 0.5 (i.e., trivial classification models can achieve an accuracy of 0.5).

Table 5. Performance of various predictive models

	Average Precision ± std. error	Average Recall ± std. error	Average F measure ± std. error	Average Accuracy ± std. error
Decision Tree	0.787 ± 0.009	0.709 ± 0.017	0.746 ± 0.007	0.759 ± 0.005
Logistic Regression	0.803 ± 0.001	0.573 ± 0.001	0.668 ± 0.001	0.716 ± 0.001
Control Model – News Count	0.763 ± 0.001	0.364 ± 0.001	0.493 ± 0.001	0.626 ± 0.001
Control Model – Search Engine Count	0.661 ± 0.002	0.721 ± 0.007	0.688 ± 0.002	0.674 ± 0.001
Control Model – Search Engine Count + News Count	0.664 ± 0.003	0.713 ± 0.009	0.685 ± 0.003	0.675 ± 0.001

We find that all of the models perform reasonably well as compared to the prior distribution of the classes in the data set. Decision tree model that uses all of the five web metrics has the best performance in terms of average accuracy and average F measure while close to the best performance on average precision and average recall. Logistic regression provides a slightly better (about 1% higher) average precision than decision tree but it comes at the cost of much lower (about 19% lower) average recall as compared to decision tree. Similarly, two of the three control models achieve a slightly (<2%) better average recall than the decision tree with all five web metrics, but it comes at the cost of much lower (about 15% lower) average recall. In other words decision tree that uses the five web metrics provides a better trade-off between precision and recall (hence a higher F-measure) and also achieves higher accuracy than any of the other models. While the control models that use previously suggested metrics achieve reasonable performance, their average F measure and average accuracy is statistically significantly ($p <$ 0.001) lower than the decision tree model with the five web metrics. In other words, the metrics that we suggest provide predictive power in addition to what is available through these control variables. In particular, with the five web metrics combined we achieve about 12% to 21% higher accuracy and 8% to 51% higher F-measure than models with control variables.

Analysis: Skewed Data Sets

The results that we have presented until now are based on a balanced data set where it is equally likely to find a competitor and a non-competitor (i.e., 50% prior probability of each class). In many realistic scenarios the data is expected to be much more skewed towards non-competitors. In other words, we may expect the non-competitors to far exceed competitors in many portfolios (set of companies) of interest. To understand the sensitivity and relative performance of models on skewed data we create data sets with different ratios of competitors and non-competitors. For that purpose, we again split the data into training and testing as before. However this time we randomly filter out data instances to obtain a desired skewed competitor to non-competitor ratio (1:2, 1:5, and 1:10) within the training and the testing data sets. A ratio of 1:5 indicates that for every competitor (pair) in our training and testing data sets we have 5 non-competitors (pairs). We build our models using the training data set and evaluate their performance on the testing data set. All models evaluated use decision tree algorithm but vary in the web metrics used. The first model uses all of the 5 web metrics (*all*), the second model uses the news count alone (*news*), the third model uses the search engine count alone (*se*), and the fourth model uses both the news count and the search engine count (*news + se*). Hence the last three models are control models that utilize previously suggested web metrics.

Figure 3 shows the performance of the four models for different competitor to non-competitor ratio data sets. We find that using the web metrics suggested in this paper consistently provides an advantage over using only the control variables for all of the skewed data sets (ratios of 1:2, 1:5, and 1:10). Moreover, this advantage is seen across the four different performance measures. These results further strengthen the argument to include the suggested web metrics in models that attempt to identify competitor relationships.

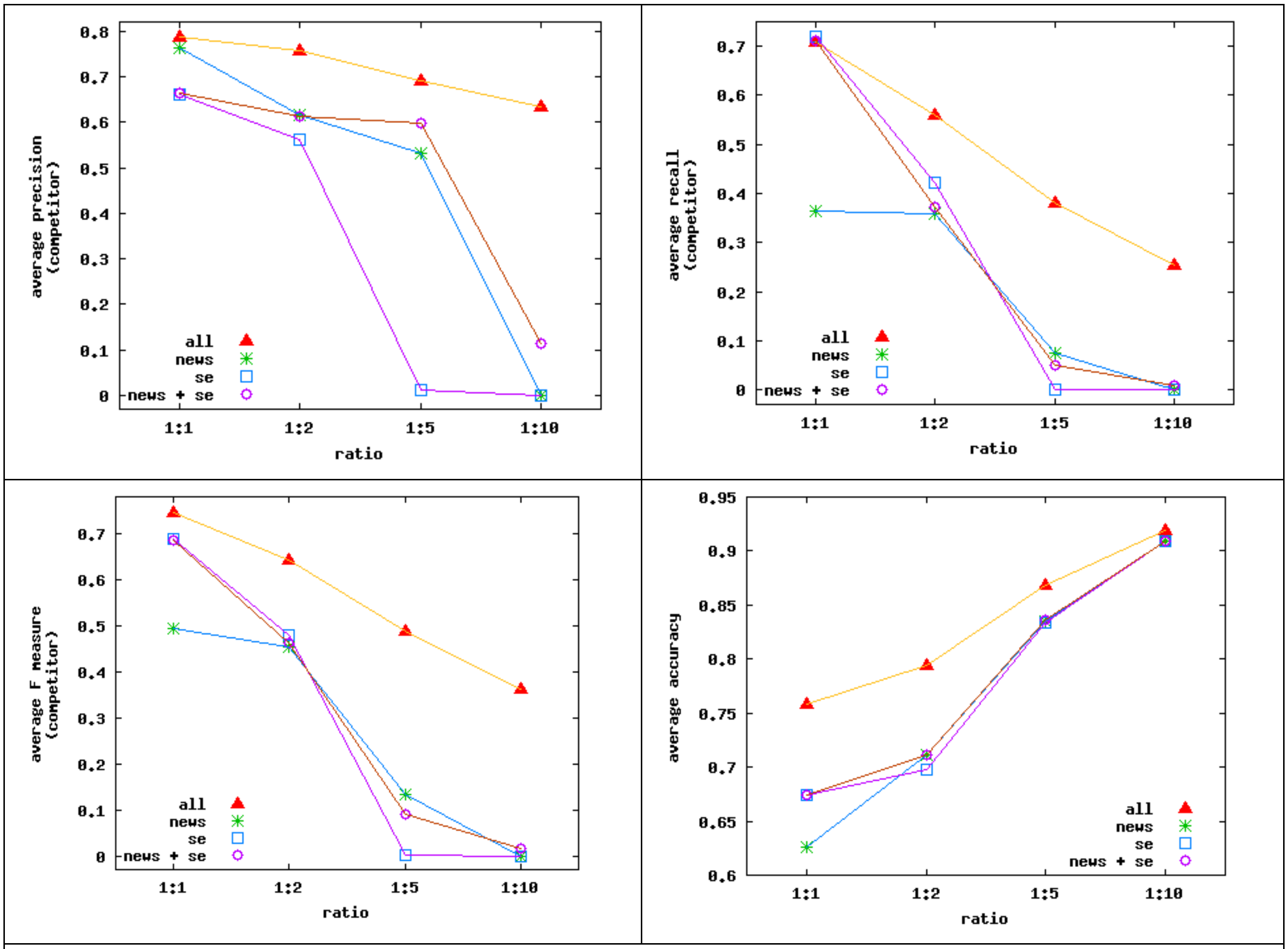

Figure 3. Performance of decision tree models using all of the 5 web-based metrics (all), only the news count (news), only the search engine count (se), and the search engine count with the news count (news + se).

Conclusion

We present a systematic study of various web metrics that may contain relevant cues for competitor identification problem. While competitor identification has been highlighted as a critical and challenging step in competitive analysis and strategy, there is limited literature on automatic identification of competitors. We argue that with increasing footprint of companies and their clients/suppliers on the web, it is now realistic to assume that some web metrics may provide effective signals for automatically identifying competitors. However such metrics need to be carefully formulated and explored for their relevance to the competitor identification problem. We suggest three new web metrics for the purpose of competitor identification and add another two web metrics that are derived from fundamental assumptions in previous works in text and web mining. We find that all of these metrics, on an average, have statistically different values for competitors as compared to non-competitors. This finding prompts the use of these variables as inputs in predictive models that attempt to classify company pairs as competitors or non-competitors. We find the resulting predictive models provide high accuracy, F measure, precision, and recall. The models also indicate that the new web metrics suggested by us provide statistically significant and strong benefit as compared to just using the individual control variables that are derived from previous literature. The observed benefit is especially strong for skewed data sets where non-competitors outnumber competitors. We note that such skewed data sets are expected in many real-world scenarios where competitor identification is needed.

As a future direction, we plan to extend our current cross-sectional study into a longitudinal study where we evaluate the predictive models in terms of being able to predict future competitors. In addition, using human experts, we would like to understand if our predictive models identify indirect or potential competitors that are likely to be missed by most manually created/updated databases. We also plan to evaluate our predictive models on their ability to create ranked list of candidate competitors instead of simple binary classification. Currently, our metrics do not try to identify specific types of pages on company's web site such as those describing products or partners. Such pages could provide additional value for competitor identification. Our metrics are also agnostic to differences between the supply side and demand side competitors and solely concentrate on overall substitutability or overlap between companies. In the future we would extend our study to identify additional subtleties of competitor relationships.

Acknowledgements

We thank the anonymous reviewers and the associate editor for their valuable suggestions.

References

Bao, S., R. Li, Y. Yu, and Y. Cao. "Competitor mining with the web," *IEEE Transactions on Knowledge and Data Engineering* (20:10), 2008, pp 1297–1310.

Bergen, M. and M. A. Peteraf (2002). "Competitor identification and competitor analysis: A broad-based managerial approach," *Managerial And Decision Economics* (23), 2002, pp 157—169.

Bernstein, A., S. Clearwater, S. Hill, and F. Provost. "Discovering knowledge from relational data extracted from business news," *In Proc. of the KDD 2002 Workshop on Multi-Relational Data Mining*, 2002.

Bernstein, A., S. Clearwater, and F. Provost. "The relational vector-space model and industry classification," *In Proc. of Workshop on Learning Statistical Models from Relational Data*, 2003.

Chen, M.-J. "Competitor analysis and interfirm rivalry: Toward a theoretical integration," *Academy of Management Review* (21:1), 1996, pp 100–134.

Desarbo, W. S., R. Grewal, and J. Wind. "Who competes with whom? a demand-based perspective for identifying and representing asymmetric competition," *Strategic Management Journal* (27), 2006, pp 101—129.

Ma, Z., G. Pant, and O. R. L. Sheng. "Mining competitor relationships from online news: A network-based approach," *Working Paper*, 2009.

Porac, J. F. and H. Thomas. "Taxonomic mental models in competitor definition," *Academy of Management Review* (15:2), 1990, pp 224–240.

Salton, G. and M. J. McGill. *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

Walker, B. A., D. Kapelianis, and M. D. Hutt. "Competitive cognition," *MIT Sloan Management Review*, 2005.

Zajac, E. J. and M. H. Bazerman. "Blind spots in industry and competitor analysis: Implications of interfirm (mis)perceptions for strategic decisions," *Academy of Management Review* (16:1), 1991, pp 37–56.