A Supervised Approach to Predict Company Acquisition With Factual and Topic Features Using Profiles and News Articles on TechCrunch

Guang Xiang[†] **Zeyu Zheng**[†] **Miaomiao Wen**[†] **Jason Hong**[†] **Carolyn Rose**[†] **Chao Liu**[§]
[†]School of Computer Science
Carnegie Mellon University
{guangx, zeyuz, mwen, jasonh, cprose}@cs.cmu.edu
[§]Internet Services Research Center
Microsoft Research
chaoliuwj@gmail.com

Abstract

Merger and Acquisition (M&A) is a critical corporate strategy for companies to preserve their competitive advantages, and M&A prediction has been an interesting and challenging research topic in the past a few decades. However, past work has only adopted numerical features such as accounting, financial and market variables in building models, and yet the valuable textual information from the great variety of social media sites like news portals and microblogs, which discusses tech trends a lot and is potentially helpful for M&A prediction, has not been touched at all. To fully explore this information, we used the profiles and news articles for companies and people on CrunchBase, the largest public database for the tech world, which anybody can edit. Specifically, we explored topic features via topic modeling techniques, as well as a set of other novel features of our design within a machine learning framework. We conducted experiments of the largest scale in the literature, and our approach achieved a high true positive rate (TP) between 60% to 79.8% with a false positive rate (FP) mostly between 0% and 8.3% over categories with less missing attributes in the CrunchBase profiles.

Introduction

Merger and acquisition refers to the process of buying, selling, dividing or combining other companies to boost the growth of an enterprise. In particular, M&A prediction deals with choosing proper target companies for the bidder company, and is an important and challenging task. First, M&A prediction is the critical step that makes assessments on each company about its chance of being acquired, which will facilitate both parties to develop the best strategies. Moreover, M&A prediction is valuable for venture capital (VC) firms to choose investment targets, which are typically those with the potential to grow rapidly. Since the distinction between a merger and an acquisition has become increasingly blurred, we will use M&A and acquisition as synonyms in this paper.

Although quite a few techniques for M&A prediction have been proposed in the literature, there are a few common weaknesses among them. First, the scale of previous work is limited by the volume of their data sets, the largest of which only had $2,394$ M&A cases with 61 acquired instances of acquisitions (Wei, Jiang, and Yang 2009). Abundant empirical evidence over the past decade has suggested that the size of training data eventually becomes more critical than the sophistication of the algorithms themselves, especially for the scale as enormous as the world wide web and social media (Banko and Brill 2001; Norvig 2008). Second, prior work has employed numerical operationalizations of financial, managerial, and technological variables in predictive models, while ignoring the potentially valuable textual data that is available as a rich resource from social media sites. Prior work has demonstrated that social media is effective in detecting events and expressing public opinions (Sakaki, Okazaki, and Matsuo ; O'Connor et al. 2010). In this paper, we utilized topic modeling techniques over news articles from TechCrunch to augment a manifold of other numerical features of our design for M&A prediction, achieving a high TP of up to 79.8% with the FP mostly between 0% and 8.3%. Another benefit of exploiting topic features is that values for numerical features may not be accessible, while news articles typically abound. Moreover, public data sets, like CrunchBase, are typically sparse with missing entries despite their scale, which adds another benefit to the use of text articles and topic modeling techniques.

Our contributions to the literature are two fold.

1. To the best of our knowledge, our work is the first in exploring topic modeling over news articles to enhance traditional features for M&A prediction.
2. Our work is also the first in utilizing one of the premier sources for tech news and startups nowadays, i.e., CrunchBase, which maintains a high volume of profiles and media articles covering a wide spectrum of aspects about companies, people, financial organizations and products, and relies on the web community to edit most of its pages.

Related Work

Previous Research on Acquisition Prediction

Prior studies on M&A prediction generally fall in three categories. The first exploits financial and managerial variables (Hyytinen and Ali-Yrkko 2005; Gugler and Konrad 2002; Meador, Church, and Rayburn 1996; Pasiouras and Gaganis

2007; Wei, Jiang, and Yang 2009) in building models. Typical financial features include firm size, market to book value ratio and cash flow, etc., and for managerial variables, management inefficiency, resource richness, industry variations, etc. were often used.

In addition to the financial and managerial view of this problem, data mining and machine learning strategies were also explored (Meador, Church, and Rayburn 1996; Ragothaman and Ramakrishna 2002; Slowinski, Zopounidis, and Dimitras 1997; Wei, Jiang, and Yang 2009). Based on the Naive Bayes classification model, Wei et al. proposed a set of features (Wei, Jiang, and Yang 2009) that model a company's technological quantities. They proposed to utilize the ensemble learning algorithm on resampled data to solve the problem of data skewness, resulting in a TP of 46.43% on $2,394$ companies out of which 61 actually got acquired. In another work, an expert system, ACQTARGET, was designed as a useful evaluation tool to classify firms into acquisition and non-acquisition target categories (Ragothaman and Ramakrishna 2002).

Lastly, researchers have also studied business failures and bankruptcies. Among them, the first dated back to (Altman 1968; Beaver 1966) 1960s, which used empirical methods and proposed several financial ratios as features, giving rise to multivariate statistical analysis (Karels and Prakash 1987) and discriminant analysis (Deakin 1972) for this task. Since early 1990s, machine learning and data mining techniques dominated the domain of bankruptcy prediction, yielding a few representative works such as (Olson, Delen, and Meng 2012; Cho, Hong, and Ha 2010; shik Shin, Lee, and jung Kim 2005; Wilson and Sharda 1994). In (Wilson and Sharda 1994), Wilson and Sharda compared the predictive power of neural networks and discriminant analysis based on five financial ratios, and concluded that neural networks clearly outperformed the traditional discriminant analysis. Moreover, Shin et al. showed in (shik Shin, Lee, and jung Kim 2005) that the Support Vector Machine (SVM) also achieved a competitive performance especially with small training sets.

The originality of this paper is that, we proposed to utilize topic features from the text of tech news articles to augment other numerical features for M&A prediction, and conducted experiments using data from CrunchBase and TechCrunch, the largest public database with profiles and news about companies.

TechCrunch and CrunchBase

In our paper, we used the public company and people profiles as well as tech news articles from TechCrunch and CrunchBase. TechCrunch (Arrington 2005a), founded in 2005, is a popular technology publication, dedicated to profiling startups, reviewing new Internet products and breaking tech news. TechCrunch and its network of web sites have more than 12 million unique visitors and over 37 million page views per month, with over 2 million friends and followers on Twitter, Facebook, and other social media. It supports social features by which users of Facebook, Twitter, LinkedIn and Google$+$ can see what their friends commented and liked. Its columnists and contributors bring insights about the tech community, pushing articles and thought pieces out daily. CrunchBase (Arrington 2005b) is TechCrunch's open database with information about startups, investors, trends, milestones, etc. It relies on the web community to edit most pages (though profiles for large companies like Facebook and Google are not editable by the general public). As of January 10, 2012, CrunchBase had profiles for $81,219$ companies, $107,274$ persons, $7,328$ financial organizations, $3,955$ service providers, $25,895$ funding rounds and $6,173$ acquisitions. CrunchBase provides API that allows public access of its data (CrunchBase 2011) in the JSON format. The local copy we crawled from CrunchBase in December 2011 is slightly smaller than that in volume, and yet still exceeds the corpus used in all previous works, the largest of which had $2,394$ cases (Wei, Jiang, and Yang 2009). Moreover, 94.1% of the companies in our corpus have at least one revision on their profiles, with a total of $359,986$ edits by the web users.

Algorithmic Method

Features are the backbone of any machine learning task, and in this paper, we designed two types of features to classify company acquisitions, including 22 factual features based on the CrunchBase profiles and a varied number of topic features using TechCrunch articles. We will use "successful" to denote companies that got acquired, and "unsuccessful" to represent other companies including failed ones and those that went public on the initial public offering (IPO).

Factual Features

Our factual features can be classified into three categories: basic features, financial features, and managerial features.

Basic Features This category measures the basic statistics of a company, including *1:#employees, 2:company age (months), 3:number of milestones in the CrunchBase profile, 4:number of revisions on the company CrunchBase profile, 5:number of TechCrunch articles about the company, 6:number of competitors, 7:number of competitors that got acquired, 8:headquarter location, 9:number of offices, 10:number of products, 11:number of providers*.

In computing company age, feature 2 considers the date a company was founded and the date of its acquisition or the current day if acquisition data is absent. Similarly, features 3, 4 and 5 collect corresponding statistics prior to the acquisition of the target company. Feature 3 is obtained from the "milestones" attribute of the JSON company profile on CrunchBase, which has multiple entries with date and description. For instance, one milestone for Facebook is "Facebook adds comments to Mini-Feed", added on Jun 25, 2008. For feature 4, we speculated more revisions might mean better management of the public profile. Features 6, 9, 10 and 11 come directly from the CrunchBase company profile. Feature 11 captures entities providing services, data, hardware, etc. to the target company. Moreover, we assigned a total of five values to feature 8, including "Silicon Valley", "New York City", "Seattle", "Boston" considering their prominent positions in the startup world, and "other area" for all the rest of the areas.

Financial Features Strong financial backing is generally considered critical to the success of a company. As such, we designed a total of eight features to capture the influence of finance-related factors, using the company and person profiles on CrunchBase.

This category includes *12:number of funding rounds, 13:number of investments by the company, 14:number of acquisitions by the company, 15:number of venture capital and private equity firms investing in the company, 16:number of people with financial background investing in the company, 17:number of key persons in the company with financial background, 18:number of investors per funding round, 19:amount of investment per funding round*.

Our hypothesis here is that successful companies are more likely to have more funding and financial experience. Feature 15 involves VCs and PEs, and toward that end, we obtained 92 VCs and 266 PEs from Wikipedia (Wikipedia 2011b; 2011a), and then refer to the funding information in the CrunchBase profiles to extract values for this feature. Moreover, we consider it a good sign for a company to have more people with financial background involved with it. Specifically, feature 16 inspects persons with financial experience (from the people category of the CrunchBase corpus) that ever invested in the target company. Feature 17 examines the "relationships" field of the company profile for persons, both former and present, with experience in financial organizations.

Managerial Features The conventional wisdom is that the experience and influence of founders have an invaluable impact on a company, and in this category, we evaluate companies along that dimension. Specifically, we have: *20:number of companies founded by founders of the target company, 21:number of successful companies by founders, 22:founder experience (months)*.

Feature 22 measures the experience (months) of the target company's founders in founding other companies prior to the acquisition of that company. "founders" here denotes people with keywords "founder", "director", and "board" in the "title" field of their CrunchBase profile.

Topic Features

In this social world, news articles, especially those by authoritative web portals, typically have headlines discussing interesting technologies, products, and trends over time. Previous research endeavors (O'Connor et al. 2010; Ramage, Dumais, and Liebling 2010) have shown the efficacy of social media in expressing public opinions and the power of topic modeling techniques to categorize web content. In this paper, we chose to extract topic features from articles in the most popular tech news web site, TechCrunch, to augment the traditional factual features for M&A prediction.

The central idea here is to treat the news articles for each company as a finite mixture over an underlying set of topics, each of which in turn can be characterized by a distribution over words, and build models via such topic distributions using machine learning techniques. Intuitively, topics strongly associated with acquisition will have higher probabilities for words that constantly occur in acquisition-related themes. By condensing the high dimensional n-gram space into the succinct yet highly representative composite topic distributions, our approach is more robust against slight disturbance in the traditional bag-of-words approach for text analysis. Specifically, we adopt the latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003), a renowned generative probabilistic model for topic discovery, to build the composite topical features. The detailed procedure of topic feature extraction is given in Algorithm 1. Since the news corpus we currently have for topic discovery is of a small volume, as analyzed in the experiment setup section, learning too many topics would split useful signals across multiple weak features instead of a few strong ones, and therefore, we set the number of topics to 5.

Algorithm 1 ExtractTopicFeatures

Require: TechCrunch articles TC, all companies C, number of topics n
Ensure: topic distributions TD for all companies
1: $raw\_text \leftarrow \phi$
2: **for all** $c \in C$ **do**
3: $raw\_text \leftarrow raw\_text \cup$ all articles about c in TC
4: **end for**
5: $text \leftarrow$ tokenize raw_text, remove stopwords, retain words consisting entirely of letters, -, and '
6: $TD \leftarrow$ learn topics on $text$ via LDA
7: $TD \leftarrow TD \cup$ uniform topic distributions for companies with no articles in TC
8: **return** TD

Machine Learning Algorithms

In our approach, we chose Bayesian Network (BN) as the primary learning algorithm, which is a probabilistic graphical model that makes inferences via a directed acyclic graph (DAG). BN is suitable for our task in that it could discover and represent the probabilistic relationships between features via local conditional dependencies, which is more robust than those simple linear classifiers. Specifically, we build models over a training data set composed of profiles from a portion of the companies in our corpus, using the learning algorithm to handle missing values and attribute discretization for the numerical features, and then apply the models to the companies in a holdout testing set.

Experiment Setup

Evaluation Metrics

We adopted *true positive rate* and *false positive rate* as the main evaluation metrics, which are the standard for many binary classification tasks. We also used the area under the ROC curve (AUC) (Cortes and Mohri 2003), a summary statistic portraying the trade-off between TP and FP. For M&A prediction, the most important aspect of an algorithm is its TP. We believe companies are very careful in acquiring others, using techniques like ours in this paper as guideline only. Therefore, no liability issues are involved as in other domains such as spam detection, and thus a relatively higher FP does not matter as much.

Category-wise Evaluation

Each company typically offers products in a certain domains, such as web, mobile, and has its own specialty like marketing strategy and life cycle, which lends legitimacy to our methodology of investigating companies by categories. The category information is obtained from the "category_code" field of the JSON company profile on CrunchBase. Moreover, some categories are similar in nature, and so we also evaluate our technique after combining them. Specifically, we created a "computer" category, which includes "ecommerce", "enterprise", "games video", "mobile", "network hosting", "search", "security", "software", "web", and a "hardware-related" category, which is a combination of "hardware" and "semiconductor". A full list of categories is shown in Table 3. For each individual category and aggregate category, we conducted experiments with 10-fold cross-validation (cv), which is a standard evaluation strategy in machine learning to reduce the variance of the resulting estimates. Furthermore, this evaluation strategy can be easily generalized to unseen companies or new ones, because the area of a company is one of the attributes that are easily obtainable, and we can always apply the models built a priori for the corresponding category to classify those instances.

Profiles on CrunchBase and Ground Truth Labels

A startup typically needs to develop for a few years prior to its acquisition, if it were to be acquired at all, which means that the ground truth for new companies are usually unavailable. To further investigate this issue, we crawled the profiles on CrunchBase in the beginning of December in 2011 and plotted the average time toward acquisition in Fig. 1, which clearly indicates that traditional sectors like hardware and security took much longer to an acquisition than the new areas like web and mobile. The average time across all categories is 59.14 months, or about 5 years.

Interestingly, the average time toward acquisition is 50.04 months for companies founded between 2002 and 2008, indicating the trend of a faster exit for new startups. To alleviate the impact of the aforementioned temporarily unavailable labels on our category-wise evaluation, we used companies founded between 1970 and 2007, as well as those with missing founding date in their CrunchBase profiles, leading to a total of 105,795 person profiles and 59,631 company profiles in the corpus for evaluation.

For the ground truth about M&A, we checked the "acquisition" field of each company's CrunchBase profile, and extracted a total of 5,915 class labels.

TechCrunch News Articles

Based on those company profiles we crawled from CrunchBase, we scraped TechCrunch in December 2011 and collected 38,617 tech news articles for 5,075 companies out of 59,631, with no articles for the remaining 54,556 companies. Among those downloaded articles, 36,642 were posted prior to the acquisition of the corresponding companies, which composed the training text for topic extraction by LDA. In particular, the 5,075 companies have an average of 7.22 TechCrunch articles per company, with a standard

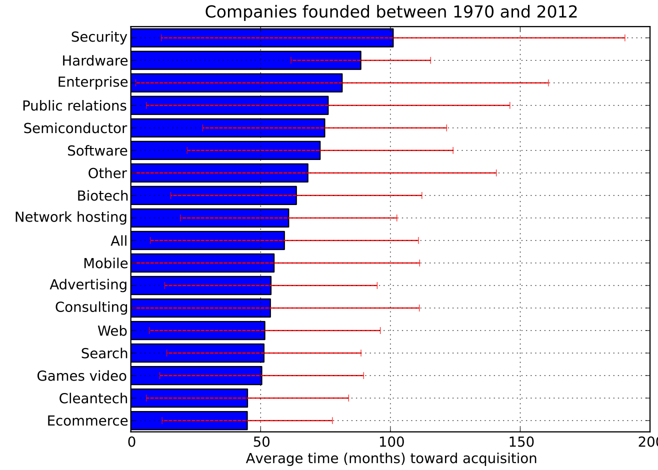

Figure 1: The average time toward acquisition among companies founded between 1970 and 2012 by category codes. The average time across all categories is 59.14 months. The red bars indicate the standard deviation, ranging from 26.89 for "hardware" and 89.29 for "security".

deviation of 74.21, suggesting a highly skewed article distribution over the companies. Interestingly, the top 10 companies with the most TechCrunch posts accumulated 13,874 articles in total, more than 1/3 of our total collection.

Experimental Result

A Summarization of Feature Values

Table 1 shows the summary statistics of the numerical features from companies in all categories. They were discretized prior to the model building step. As expected, successful companies have higher mean values for most funding-related features. We observed a similar pattern for features about founders. Features 13 and 14 suggest that a spree of spending by companies is more of a sign of deep pocket than being successful. Moreover, features 10 and 11 indicate that the quality of the products and resources matters more than the sheer scale of production.

Moreover, numbers for nominal features show that silicon valley is still the top choice for entrepreneurs and VCs/PEs, with 226 companies that got bought out, while the total number is 67 for New York City, Seattle and Boston.

Investments from Venture Capital and Private Equity Firms in the New Millennium

Since funding is an important aspect for companies, we investigated the VCs for informational purpose and ranked them by the number of successful companies they invested in since 2000. The top 5 include First Round Capital, Sequoia Capital, Benchmark Capital, Mayfield Fund and Lightspeed Venture Partners, with 17/73, 14/74, 10/47, 8/25 and 8/30 successful/total investments respectively. Note that our measure of success is an acquisition, not other forms like IPO. Also, we consider success as a binary value, whereas

Table 1: Statistics of numerical features for companies in all categories. The average is computed by removing entities with missing corresponding features from the denominator. The large standard deviations (std) for a few features are caused by skewed distributions for those features. All numerical features were discretized prior to model building.

		Acquired companies		Unacquired companies	
Feature ID	Feature name	Mean	Std	Mean	Std
Basic features					
1	#employees	$1,046.24$	$12,981.88$	560.01	$7,312.73$
2	Company age (months)	82.48	54.65	121.24	68.11
3	#milestones	0.08	0.94	0.15	0.9
4	#revisions on profile	4.0	11.05	6.25	13.77
5	#articles on techcrunch	1.01	9.39	0.57	22.64
6	#competitors	3.99	4.04	3.09	2.81
7	#competitors acquired	0.43	0.81	0.5	0.9
9	#offices	1.08	0.61	1.16	0.82
10	#products	2.05	3.42	2.26	3.12
11	#providers	1.3	0.75	1.39	1.11
Financial features					
12	#funding rounds	1.78	1.14	1.54	1.01
13	#investments by the company	0.01	0.13	0.04	0.76
14	#acquisitions by the company	0.05	0.35	0.1	1.36
15	#VCs and PEs investing	0.77	1.01	0.38	0.78
16	#finance person investing	0.16	0.63	0.06	0.39
17	#leaders with financial background	0.37	0.89	0.15	0.58
18	#investors per round	2.7	1.81	2.13	1.62
19	Funding per round (K$)	$11,708.08$	$20,980.64$	$10,896.42$	$33,730.96$
Managerial features					
20	#companies by founders	0.52	1.56	0.25	1.0
21	#successful companies by founders	0.12	0.52	0.03	0.27
22	Founder experience (months)	46.85	183.93	17.2	99.06

Table 2: Top 20 words (ranked by probabilities) from each of the 5 topics learned by LDA. Apparently, topic 3 coincides with mobile, and topic 5 relates closely to ads.

Topic No.	Top 20 words
1	million company companies business year startup capital funding technology online raised mobile investors game ventures energy market customers round data
2	facebook twitter users social people google search time site service news data web page friends information user company app post
3	google apple iphone mobile app android microsoft apps time phone web year search store device market ipad make today devices
4	users service music web site based social free app mobile startup services online people create company application features platform user
5	million video yahoo media content company advertising ad online network myspace networks aol youtube search ads sites year tv videos

these VCs and PEs would consider success in terms of return on investment (ROI).

Top Words in Topic Distributions

Before delving into the evaluation result, we visualized the most frequent 20 words from each of the 5 topics learned by the LDA model in Table 2, which is helpful for understanding how topic features are beneficial. Based on the top words, topic 1 is about startups and funding. Topic 2 is related to social networks and microblogs. Topic 3 strongly correlates to mobile devices and applications. The theme of topic 4 is obscure, but topic 5 is clearly about advertising.

Cross Validation Performance across Categories

In this section, we report the cross-validation performance of our approach in predicting M&A in Table 3. All statistics in those tables were achieved using Bayesian networks.

As shown in the table, our technique achieved a high TP (from 60% to almost 80%) for more than half of the categories, with acceptable FP. Particularly, we had a TP of 79.8% with 0% FP for the largest "other" category. That performance is much better than the 46.43% TP in the most recent work (Wei, Jiang, and Yang 2009), which has an FP of 1.61%. Moreover, with sufficient TechCrunch articles like "mobile" and "web", topic modeling improved the TP by a great margin (marked by $\star$ in the table), with some degradation on FP. The highest increase in TP occurred in the "web" category, with almost 20% enhancement. For most categories, TP changed only slightly or even remained intact with topic features. One reason is TechCrunch was founded in 2005 and not many tech articles, if any, were available for companies that got acquired prior to that in our corpus. Also, most categories do not align well with the subject of the 5 topics. Considering the summary metric AUC, topic features improved the overall performance significantly for "mobile", "web" and "computer", while influencing other categories only slightly.

Further examination on the topic distributions suggests that topic features were helpful in predicting some M&A. In Table 4, we list the means of the 5 topic features for the "advertising" category in the confusion matrix format. For this category, the majority of successful companies have a high value for topic feature 5, as indicated by the mean of 0.277. Recall from Table 2 that topic 5 is closely related to ads, and this signal is picked up well by the high values of the corresponding feature here. Topic feature 2 is helpful in classifying companies in this category as well, but is not as meaningful as topic 5 for this "advertising" category because topic 2 is more about social networks and microblogs. Similar analysis on the "mobile" category reveals that topic feature 3, which talks about mobile a lot, got more weight among successful companies.

Table 4: The breakdown on the mean value of each topic feature for the "advertising" category. The columns represent "true class label" $\rightarrow$ "predicted label". A high value for topic feature 5 for most successful companies, as manifested by the average mean of 0.277 in the table, contributed to the increase in TP from 56.8% to 68%. Table 2 shows that topic 5 corresponds strongly to ads, which is in perfect agreement with the high values for topic feature 5 in this category. Topic feature 2 helped improving TP as well, but was not as meaningful as topic feature 5 for this category.

Topic ID	$no \rightarrow no$	$no \rightarrow yes$	$yes \rightarrow no$	$yes \rightarrow yes$
1	0.2	0.203	0.198	0.172
2	0.201	0.157	0.207	0.163
3	0.199	0.137	0.194	0.185
4	0.2	0.157	0.192	0.203
5	0.2	0.346	0.209	0.277

Error Analysis on Misclassifications

A bit more scrutiny on the errors reveals that the variance in the performance across categories is mainly due to the various degrees of sparsity in the CrunchBase data and the uneven number of the TechCrunch articles.

In our feature set, five funding-related features utilize the "funding_rounds" attribute in the CrunchBase company profiles, and another four features resort to the "relationships" attribute. Accordingly, the absence of those two is extremely detrimental to our machine learning technique. For example, the CrunchBase profile for Microsoft has no funding information. Categories "legal" and "education" almost have no positive instances, with 0% TP accordingly, and we do not conduct further analysis on them. Particularly, for categories where the successful companies have more non-empty content than the unsuccessful ones (in terms of percentage) for at least one of those two attributes, the TP tends to be higher, usually over 56% except for "mobile". For "mobile", learning 15 topics yielded a higher TP of 63.7% with an FP of 14.9%. Actually, the "mobile" category, together with "games video", "network hosting", "search", "web", has the most TechCrunch articles in our corpus. For categories including "other", "public relations" and "software", the discriminative topic features compensated for the missing attributes to some extent, leading to superior performance.

For "cleantech", "enterprise" and "hardware", more successful companies lack content for the aforementioned two attributes, and the themes of the 5 topics correlate remotely with these categories, which explains their less desirable performance in Table 3. Our result also suggests the necessity of category-wise evaluation for the M&A prediction task.

Examining Features by Gain Ratio

In addition to building machine learning models, we also examined the efficacy of our individual features by the gain ratio metric (Quinlan 1993), which is widely adopted to compute the correlation between features and nominal class labels using information theoretic techniques. Actually, there are a wide range of different measures for evaluating the predictive value of features, and they are all relatively correlated with one another. We conducted this analysis in each individual category, since companies in different domains have special properties.

The conclusion based on our data set is that "#revisions on profile" was the best feature across all categories except for "mobile" where it ranked No.2, while the performance of other features was not unanimous across categories partially due to data sparsity. Out of the 59,631 companies in our corpus, 94.1% had at least one revision on their CrunchBase profile by web users, a percentage much higher than the 8.5% of the companies having TechCrunch articles. The high ranking of "#revisions on profile" and its extensive user and target base indicated the power of this social-like feature for our classification task. When data sparsity was less of a concern, funding and founder related features typically outperformed other factual features such as #products and #providers. Specifically, for categories with enough TechCrunch articles including "advertising", "games video", "mobile", "search", "web", topic features ranked high, mostly among the best 10. Founder-related features were more effective in sectors with shorter time toward an acquisition like "cleantech", "games video", "web", "mobile" and "advertising". Funding-related features were discriminative in categories with less missing attributes in the CrunchBase profiles, as explained in the error analysis section above. With these findings, we believe that our technique can keep improving as we collect more text articles from more channels to learn topic models, while at the same time, CrunchBase and other data sources become more complete and accurate.

Table 3: M&A prediction performance using Bayesian networks. The aggregate "computer" category includes "ecommerce", "enterprise", "games video", "mobile", "network hosting", "search", "security", "software", "web". The aggregate "hardware-related" category includes "hardware", "semiconductor". TP improves significantly when sufficient TechCrunch articles exist for the corresponding companies, highlighted by $\star$ after the category names.

Category code	#successful	#all	TP(%)		FP(%)		Area under ROC	
			0 topics	5 topics	0 topics	5 topics	0 topics	5 topics
Advertising ($\star$)	169	1,983	56.8	68.0	2.4	8.3	0.846	0.843
Biotech	312	2,464	62.2	62.2	0.0	0.0	0.878	0.878
Cleantech	65	1,002	50.8	50.8	0.0	0.0	0.684	0.684
Consulting	95	1,994	73.7	73.7	0.0	0.0	0.843	0.843
Ecommerce	140	2,297	60.0	60.0	0.0	0.0	0.836	0.836
Education	1	47	0.0	0.0	0.0	0.0	0.043	0.043
Enterprise	212	1,392	55.7	55.7	0.2	0.2	0.784	0.784
Games video	226	1,930	55.8	57.5	3.0	3.3	0.795	0.793
Hardware	127	1,276	51.2	51.2	0.1	0.1	0.749	0.749
Legal	2	185	0.0	0.0	0.0	0.0	0.089	0.089
Mobile ($\star$)	204	1,970	44.1	51.5	1.8	4.8	0.81	0.824
Network hosting	129	1,084	57.4	57.4	0.4	0.4	0.792	0.792
Other	1,897	25,156	79.8	79.8	0.0	0.0	0.942	0.945
Public relations	152	1,505	64.5	64.5	0.0	0.0	0.813	0.813
Search ($\star$)	49	637	51.0	61.2	1.2	6.3	0.866	0.863
Security	80	473	57.5	57.5	0.0	0.0	0.811	0.811
Semiconductor	119	574	62.2	62.2	0.0	0.0	0.806	0.806
Software	976	7,776	61.4	62.6	0.0	0.4	0.88	0.894
Web ($\star$)	652	5,886	58.3	78.4	2.4	17.3	0.845	0.851
	Performance under combined categories							
Computer ($\star$)	2,668	23,445	59.9	70.9	2.2	10.6	0.882	0.888
Hardware-related	246	1,850	56.5	56.5	0.0	0.0	0.857	0.857

Evaluation with Other Learning Algorithms

We also evaluated our approach using the Support Vector Machines (SVM) and Logistic Regression (LR). Limited by space, we refrain from reporting all the statistics. However, the finding here is that both SVM and LR were significantly outperformed by BN in terms of TP, with SVM using the radial basis function as the kernel. Although their FP is better than that of BN, BN still comes out the winner with respect to the overall performance. For example, the TP and FP on companies in the aggregate "computer" category for SVM and LR without topic features are 39.6%/0.1% and 2.8%/0.3% respectively, while BN achieved 59.9%/2.2%. This observation is not surprising due to the correlation among our features and the absence of a linear separator in the feature space that SVM and LR typically learn.

Discussion

Sparsity of the CrunchBase Corpus

As analyzed in the section about errors, despite its large magnitude, the CrunchBase corpus is sparse with many missing attributes in the profiles. It is especially so for old companies even like Microsoft. One explanation for that is, TechCrunch is only six years old and maintaining such a huge corpus takes time. Although web users around the world contributed to a large number of edits to the CrunchBase profiles, the power-law-like principles apply here as well, with most revisions made to popular entities (hot companies, influential investors, etc.) and attributes (funding and so on). However, our approach still achieved good performance even under data with such sparsity, and we believe it will keep improving as the company profiles on CrunchBase become more complete.

Further Improvement on Our Approach

We found that a small number of successful companies tend to have higher values for a certain features than other successful ones, so is the case for some unsuccessful companies, which may cause false negatives and false positives potentially. One way to alleviate this problem is to adopt more features to bridge such discrepancies among the entities with the same class label. In our technique, we did not use traditional features such as price to earning ratio, return on average asset, etc., as reviewed in (Wei, Jiang, and Yang 2009), which might help our M&A prediction task. However, values of those features may not be readily accessible, especially for new startups, which again emphasizes the originality and necessity of building topic models over text articles, which come in abundance and can be obtained with ease. In addition, superstar companies that went public on the IPO are also in our data, with negative class labels for the M&A prediction task, which are also likely to be misclassified. For such cases, one trick to augment our approach is treat IPO as acquisition as well in the ground truth.

As to the topic features, the quality and quantity of the training text corpus matter, and there are other popular social sites that we could harness to enhance topic modeling, such as Twitter, Quora and Wikipedia, which has way more data and may be suitable for this task.

Conclusions

In this paper, we proposed to attack acquisition prediction by exploring topic features based on tech news together with a set of other features of our design, providing a novel framework that exploits text news in addition to the numerical features for this task. In evaluation, we crawled the profiles on CrunchBase for various entities such as companies, people, and conducted experiments of the largest scale in the literature. Our approach achieved a high TP between 60% to 79.8% with a reasonable FP mostly between 0% and 8.3% over categories with less missing attributes in the CrunchBase profiles.

References

Altman, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23:589–609.

Arrington, M. 2005a. `http://techcrunch.com/`.

Arrington, M. 2005b. `http://crunchbase.com/`.

Banko, M., and Brill, E. 2001. Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. In *Proceedings of the First International Conference on Human Language Technology Research (HLT'01)*, 1–5.

Beaver, W. H. 1966. Financial ratios as predictors of failure. *Journal of Accounting Research* 4:71–111.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Cho, S.; Hong, H.; and Ha, B.-C. 2010. A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the mahalanobis distance: for bankruptcy prediction. *Decision Support Systems* 52(2):3482–3488.

Cortes, C., and Mohri, M. 2003. Auc optimization vs. error rate minimization. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS'2003)*.

CrunchBase. 2011. `http://www.crunchbase.com/help/api`.

Deakin, E. B. 1972. A discriminant analysis of predictors of business failure. *Journal of Accounting Research* 10(1):167–179.

Gugler, K., and Konrad, K. A. 2002. Merger Target Selection and Financial Structure. University of Berlin.

Hyytinen, A., P. M., and Ali-Yrkko, J. 2005. Does patenting increase the probability of being acquired? evidence from cross-border and domestic acquisitions. *Applied Financial Economics* 15(14):1007–1017.

Karels, G. V., and Prakash, A. 1987. Multivariate normality and forecasting of business bankruptcy. *Journal of Business Finance and Accounting* 14(4):573–593.

Meador, A. L.; Church, P. H.; and Rayburn, L. G. 1996. Development of prediction models for horizontal and vertical mergers. *Journal of Financial and Strategic Decisions* 9(1):11–23.

Norvig, P. 2008. Statistical learning as the ultimate agile development tool. In *ACM 17th Conference on Information and Knowledge Management (CIKM'08)*.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM'10)*.

Olson, D. L.; Delen, D.; and Meng, Y. 2012. Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems* 52(2):464–473.

Pasiouras, F., and Gaganis, C. 2007. Financial characteristics of banks involved in acquisitions: Evidence from asia. *Applied Financial Economics* 17(4):329–341.

Quinlan, J. R. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers.

Ragothaman, S., N. B., and Ramakrishna, K. 2002. Predicting corporate acquisitions: An application of uncertain reasoning using rule induction. University of South Dakota, Vermillion, SD 57069.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web (WWW'10)*, 851–860.

shik Shin, K.; Lee, T. S.; and jung Kim, H. 2005. An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications* 28(1):127–135.

Slowinski, R.; Zopounidis, C.; and Dimitras, A. 1997. Prediction of company acquisition in greece by means of the rough set approach. *European Journal of Operational Research* 100(1):1 – 15.

Wei, C.-P.; Jiang, Y.-S.; and Yang, C.-S. 2009. Patent analysis for supporting merger and acquisition (m&a) prediction: A data mining approach. *Lecture Notes in Business Information Processing* 22(6):187–200.

Wikipedia. 2011a. List of private equity firms. `http://en.wikipedia.org/wiki/List_of_private_equity_firms`.

Wikipedia. 2011b. List of venture capital firms. `http://en.wikipedia.org/wiki/List_of_venture_capital_firms`.

Wilson, R., and Sharda, R. 1994. Bankruptcy prediction using neural networks. *Decision Support Systems* 11(5):545–557.