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) prediction has been an interesting and challenging research topic in the past a few decades. However, past work has only adopted numerical features in building models, and yet the valuable textual information from the great variety of social media sites has not been touched at all. To fully explore this information, we used the profiles and news articles for companies and people on TechCrunch, the leading and 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 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 company categories with a small number of missing attributes in the CrunchBase profiles.

Introduction

Merger and acquisition (M&A) is an important business strategy and a challenging research task. Although quite a few techniques for M&A prediction have been proposed, 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 acquisitions (Wei, Jiang, and Yang 2009). Second, prior work has employed numerical operationalizations of financial, managerial, and technological variables in predictive models, while ignoring the valuable textual data that is available as a rich resource from social media sites. 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%. Our major 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 the first in utilizing one of the premier sources for tech news and startups nowadays, i.e., TechCrunch.

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 features include market to book value ratio, cash flow, management inefficiency, industry variations, etc.

In addition to the financial and managerial view, 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). In (Wei, Jiang, and Yang 2009), Wei et al. 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.

Lastly, researchers have also studied business failures and bankruptcies. Among them, the first (Altman 1968; Beaver 1966) 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 techniques dominated the domain of bankruptcy prediction, yielding a few representative works (Olson, Delen, and Meng 2012; Cho, Hong, and Ha 2010; shik Shin, Lee, and jung Kim 2005; Wilson and Sharda 1994).

TechCrunch and CrunchBase

In our paper, we used the public people and company profiles as well as tech news articles from TechCrunch and CrunchBase. TechCrunch (Arrington 2005), founded in 2005, is a popular technology publication, dedicated to profiling startups, reviewing new products and breaking tech news daily. CrunchBase is TechCrunch's open database with information about startups, investors, trends, milestones, etc. It relies on the web community to edit most pages.

Algorithmic Method

For this task, we designed two types of features, including 22 factual features based on the CrunchBase profiles and a varied number of topic features using TechCrunch articles.

Terminology: We use "successful" to denote companies that got acquired, and "unsuccessful" to represent other companies including failed ones.

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*.

Features 2, 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 company profile on CrunchBase. Features 6, 9, 10 and 11 come directly from the company profile. Feature 11 captures entities providing services, data, hardware, etc. to the target company.

Financial Features Strong financial backing is generally considered critical to the success of a company, and as such, we designed eight features to capture the finance factors.

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 (VC) and private equity (PE) 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*.

Based on a list of 92 VCs and 266 PEs, feature 15 refers to the funding information in the CrunchBase profiles for value extraction. Feature 16 inspects persons with financial experience that ever invested in the target company. Feature 17 examines the "relationships" field of the company profile for persons with experience in financial organizations.

Managerial Features The conventional wisdom is that the experience of founders have an invaluable impact on a company, and here, we evaluate companies along that line. 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

The central idea is to treat the news for each company as a finite mixture over an underlying set of topics, each of which is in turn characterized by a distribution over words, and build models via such topic distributions using machine learning techniques. With such a representative feature space, our approach is more robust than the bag-of-words paradigm. Specifically, we adopt the latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) to build the composite topical features, as shown in Algorithm 1. Since our TechCrunch news corpus is of a small volume, 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

Experiment Setup

Category-wise Evaluation

We adopted *true positive rate* and *false positive rate* as the main evaluation metrics. We also used the area under the ROC curve (AUC), a summary statistic portraying the trade-off between TP and FP. Moreover, we conducted evaluation by company categories, which is obtained from the "category_code" field of the company profile. Some categories are similar in nature, and so we also evaluated our technique after combining them (Table 2).

Profiles on CrunchBase and Ground Truth Labels

Since the ground truth for new companies are unavailable, we only used those founded between 1970 and 2007, and those with missing founding date, leading to $105,795$ person profiles and $59,631$ company profiles in our corpus for evaluation. 94.1% of the companies in our corpus have at least one revision on their profiles, with $359,986$ edits by the web users in total. We checked the "acquisition" field of each company's profile, and extracted $5,915$ class labels.

TechCrunch News Articles

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. Among those articles, $36,642$ were posted prior to the acquisition of the corresponding companies. In particular, the $5,075$ companies have an average of 7.22 articles per company, with a standard deviation of 74.21, suggesting a highly skewed 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.

Table 1: Top 10 words from each topic learned by LDA. Topic 3 coincides with mobile, and topic 5 relates closely to ads.

Topic No.	Top 10 words
1	million company companies business year startup capital funding technology online
2	facebook twitter users social people google search time site service
3	google apple iphone mobile app android microsoft apps time phone
4	users service music web site based social free app mobile
5	million video yahoo media content company advertising ad online network

Experimental Result

Top Words in Topic Distributions

In Table 1, we visualized the 10 most frequent words from each of the 5 topics. Suggested by 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, and topic 5 is about advertising.

Cross Validation Performance across Categories

We report the performance of our approach using Bayesian networks evaluated by 10-fold cross validation in Table 2. Our technique achieved a high TP (from 60% to almost 80%) for more than half of the categories, with acceptable FP. Moreover, for categories with sufficient TechCrunch articles like "mobile" and "web", topic modeling improved the TP by a great margin. 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 were available for companies that got acquired prior to that. We give further evidence of the effectiveness of topic modeling in Table 3, which shows the mean values of the topic features for the "advertising" category.

Table 3: The breakdown on the mean value of each topic feature for the "advertising" category. The columns represent "true class label" → "predicted label". A high value for topic feature 5 for most successful companies, as manifested by the average mean of 0.277, contributed to the increase in TP from 56.8% to 68%. Topic 5 corresponds strongly to ads, which is in perfect agreement with the high values for topic feature 5 here. 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

The variance in the performance across categories is mainly due to the various degrees of sparsity in the TechCrunch data. In our feature set, five funding-related features utilize the "funding_rounds" attribute in the CrunchBase company profiles, and another four resort to the "relationships" attribute. For categories where more successful companies have non-empty content than the unsuccessful ones (percentage) for at least one of those two attributes, the TP tends to be higher, usually over 56% except for "mobile".

Examining Features by Gain Ratio

We also examined the efficacy of our features by the gain ratio metric (Quinlan 1993) for each category. The conclusion 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 due to data sparsity. When data sparsity was less of a concern, funding and founder related features typically outperformed other factual features such as #products and #providers.

Discussion

Despite its large magnitude, the CrunchBase corpus is sparse with many missing attributes. The power-law principles suggest that web users are more willing to edit popular entities and attributes. However, our approach still achieved good performance, and we believe it will keep improving as the company profiles on CrunchBase become more complete.

In this work, we did not use traditional features such as price to earning ratio, return on average asset, etc., which might help our M&A prediction task and yet are not as readily accessible for most companies as the news articles. In addition, companies that went public on the IPO were assigned negative labels, which are also likely to be misclassified. For such cases, one possible solution is to treat IPO as acquisition in the ground truth. Furthermore, we can enhance our topic features by harnessing other popular social sites like Twitter, Quora and Wikipedia in addition to TechCrunch.

Conclusions

In this paper, we proposed to attack M&A 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, 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 a small number of 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. 2005. `http://techcrunch.com/`.

Table 2: 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

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.

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.

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.

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.

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