Finding New Customers Using Unstructured and Structured Data

Prem Melville, Yan Liu, Richard Lawrence,
Ildar Khabibrakhmanov, Cezar Pendus, Timothy Bowden
IBM T.J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598
{pmelvil,liuya,ricklawr,ildar,cpendus,trbowden}@us.ibm.com

ABSTRACT

Identifying new customers is a critical task for any sales-oriented company. Of particular interest are companies that sell to other businesses, for which there is a wealth of structured information available through financial and firmographic databases. We demonstrate that the content of company web sites can often be an even richer source of information in identifying particular business alignments. We show how supervised learning can be used to build effective predictive models on unstructured web content as well as on structured firmographic data. We also explore methods to leverage the strengths of both sources by combining these data sources. Extensive empirical evaluation on a real-world marketing case study show promising results of our modeling efforts.

1. INTRODUCTION

Sales-oriented companies must continue to find new customers for their products and service offerings. For companies that sell to other businesses, this means identifying new companies with potential interest in purchasing the seller's offerings. Aside from specific for-purchase marketing databases, there are several sources of data relevant to this task. These include

1. Extensive financial information for publicly-traded companies (*e.g.* Standard and Poor's [7])
2. Firmographic data (*e.g.* location, industry, estimated company revenue and number of employees) for a large number of companies (*e.g.* D&B [1])
3. News feeds (*e.g.* Reuters [5])
4. Content extracted from the websites of a universe of potential customers.

Any of these sources of data can be joined with the seller's historical transactions as a basis for building probability-to-purchase models (*e.g.* [25]). For example, D&B firmographic information can be joined with past transactions to build customer targeting models [18] that estimate purchase probabilities based a labeled set of positive examples, *i.e.* previous purchasers of a specific product. But there can be instances where past transactional data are either unavailable or not immediately relevant to the specific task. For example, suppose we are interested in a slightly different business objective, namely identifying companies with whom we might wish to form a business partnership. Such a partnership could involve an agreement to sell each other's products and/or services. In this case, we may wish to find companies with a specific sales strategy that compliments our own business objectives. While firmographic data like D&B can be used to identify a broad pool of candidate companies, it does not contain specific information on a company's overall business alignment. It is clear that company websites are much more likely to contain the relevant information.

To illustrate the issues, we consider the following specific example. Let us assume we are interested in finding partners to sell a specific financial offering. We believe that companies interested in such an offering may also be interested in purchasing consulting services around Sarbanes-Oxley [6] compliance. One strategy to tap into this market quickly would be to enter into a co-marketing agreement with a company that sells Sarbanes-Oxley (SOX) consulting services. However, if we do a web search for "Sarbanes Oxley", we will find a lot of useful information on this topic, but relatively few companies that sell services related to it. Indeed, our objective is to find not only companies that specialize in SOX, but to find them within a specific firmographic window. We may wish to exclude both very small and very large companies, and hence limit the search only to companies with annual revenue between $100M and $1B. We may be interested only in a specific SIC [8] code covering Professional Services. This expanded search requires a fusion of structured (firmographic) and unstructured data (web content).

This scenario introduces some very interesting machine learning issues in the emerging area of analyzing combined structured and unstructured data. It may be possible to have experts inspect websites selected via a firmographic filter, and generate binary labels reflecting the degree of "fit" to the partner qualifications. In this case, we can build supervised models that learn these characteristics, and use the model to

KDD'07, August 12 – 15, 2007, San Jose, California, USA.

score other companies within this firmographic window. In the following section, we describe the data obtained in such a labeling exercise. In Sections 3 and 4, we describe models built using the web content and the firmographic data, respectively. We are particularly interested here in the relative predictive power obtained by combining these data sources – Section 5 describes our current efforts in this area.

2. THE COMPANY IDENTIFICATION TASK

We have been able to develop a labeled data set for supervised learning via the process summarized in Figure 1. As discussed in the previous section, the specific application is to identify companies with whom we may wish to partner in order to market a particular financial offering. The first step is to develop a universe of potential companies, based solely on firmographic data such as a company revenue and SIC industry classification. The revenue window is set to eliminate large companies since we are looking for mid-size companies with whom to partner. Using the US D&B table with approximately 15M company sites, this query yields a D&B subset of approximately 2400 companies.

The next step is to obtain the URL of the website for each company within this firmographic universe. Our D&B table does not contain this information, so we resort to submitting the company name to various search engines and processing the returned results. While this process is quite reliable for larger companies, it can return incorrect results for the set of relatively small companies under consideration here. We applied a set of heuristics to improve the accuracy of the company-to-URL mapping.

The immediate challenge is to assign a binary label to each company that reflects their perceived qualifications as a business partner. These labels were generated by a team of human experts with a detailed understanding of desirable characteristics of a successful partner. Each website within the 2400-company universe was inspected by this team, and positive labels were assigned to companies that appeared to be reasonable candidates, based solely on the experts' judgment. Note that there are no specific terms that were required to be on a site in order for it to be labeled as a positive. Rather, the experts would browse the site, and form an opinion based on a broad sense of the potential match. As a result of this exercise, 179 companies were labeled as positives, with the balance (2262) taken as negatives.

Once the data are available, our next step is to map the task to a machine learning problem. In essence, finding business partners is similar to a retrieval task or recommendation system, which can be treated as a ranking problem. However, since no reliable confidence score can be obtained for each company (even by human experts), we simply view the task as a classification problem with special properties, i.e unbalanced data and features from multiple sources.

3. WEB CONTENT MODELS

There exist many information sources where one can acquire the business profile of a company, such as Hoovers [4], Factiva [2], and Harte-Hanks [3]. However, with the increase of valuable information on the world wide web, we can gather a wealth of information just from the content of company websites. The rich information on a company's website often

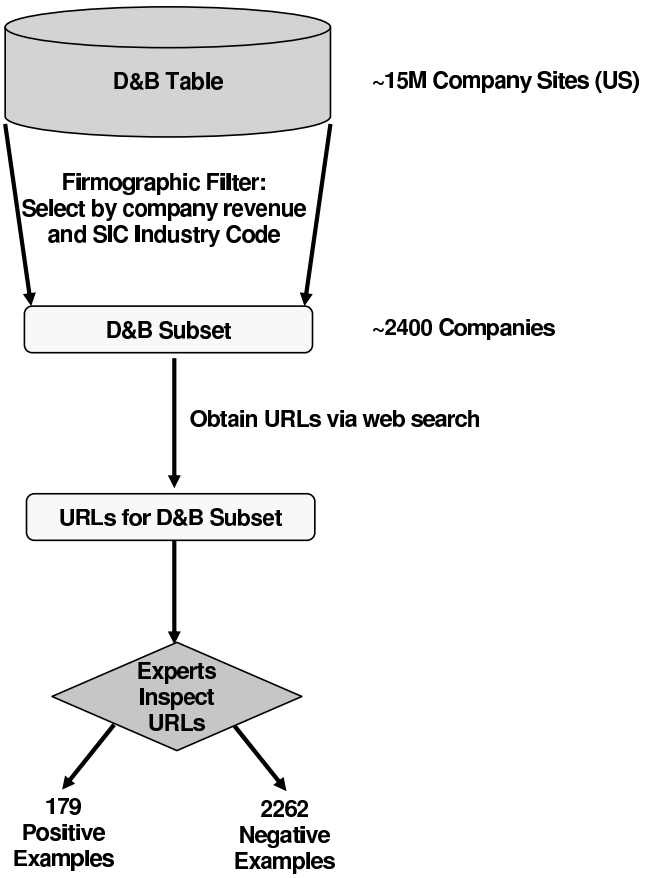

Figure 1: Construction of the labeled data set.

describes the services they support or the products they sell, as well as who their partners are. These are extremely valuable pieces of information that are difficult to acquire from alternative structured sources. In addition, recent changes in the strategies of a company are usually reflected immediately on their websites, while it may take longer for their entries to be updated in databases maintained by third parties.

3.1 Data Preparation

To extract the web content from companies' websites, we first need to find the URLs of the target companies by querying the company names on popular search engines, such as Google or Yahoo!. This seemingly simple task turns out to be rather challenging to automate due to the fact that

1. many of our targets are small or medium-sized companies, and therefore their websites are not ranked at the top in search results;
2. some companies share common words in their names or even share the same name, which makes it difficult to determine the correct URLs, even for humans;
3. some big companies have branches in multiple locations providing different business – in many cases, the search engine will return the URL of their parent company or a wrong branch.

Clearly, the task of automatically identifying URLs for a company itself desires more careful examination. We developed several heuristics to aid in correctly resolving company URLs. We skip the detailed discussion of heuristics here and assume that we have the correct URLs for each company.

Next, for each company in the data set described in Section 2, we crawl the company's website up to a depth of 4 and recompile all the pages into one big file. We pre-process the text by removing stop words, stemming the words into inflected forms (e.g. from the plural form to the singular form and from the past tense to the original form), and selecting features using the χ^2 scores, which is shown to be the best feature-selection method in previous empirical studies [35]. These processes result in a collection with a vocabulary of around 6000 words, which we convert into vectors using the bag-of-word representation with TF-IDF term weighting [12].

3.2 Data Analysis

It is not hard to imagine that there can be some irrelevant information on the company webpages which may affect our modeling results, such as spam advertisements, slogans, contact information, directions, etc.. After some examination, we found that the feature selection algorithm based on χ^2 tests is extremely helpful in reducing such noise in the data. Below is a list of top-ranked words using χ^2 scores:

sarban oxley FDICIA PCAOB outsourc quickbook CPA ERP fraud whitepap firm CFO forens llp financ client payrol sharehold consult COSO

These results are quite encouraging because all these terms have been identified as positively relevant by marketing experts. For example, one type of potential IBM partners are those companies that provide services and consulting on the "Sarbanes-Oxley Act". As a result, the terms such as "PCAOB" and "FDICIA" are relevant because the first refers to a private-sector, non-profit corporation, created by the Sarbanes-Oxley Act, to oversee the auditors of public companies and protect the interest of investigators, while the second term represents the Federal Deposit Insurance Corporation Improvement Act of 1991, which was passed before the Sarbanes-Oxley Act during the savings and loan crisis to strengthen the power of the Federal Deposit Insurance Corporation.

3.3 Experimental Evaluation

Given the web content, we cast the task of customer identification into one of text classification, i.e., given a text document representing a company, classify it as a positive or negative example of a potential customer. We can now use one of many text classification methods available to solve this problem. In particular, we compare the following approaches:

1. SVM-light [17] — an efficient and scalable implementation of Support Vector Machines for text classification.
2. Naïve Bayes using a multinomial text model[19].
3. K-Nearest Neighbor (KNN) [13], with the number of neighbors, k, set to 3.

We also ran versions of the above algorithms modified to deal with the high imbalance between the positive and negative class. SVM-light provides a straightforward mechanism for dealing with class imbalance by specifying a cost factor by which training errors on positive examples outweigh errors on negative examples. We set this cost factor to 10, and refer to this variant as SVM(c=10). As noted by Rennie et al. [24] and Frank and Bouckaert [15] naïve Bayes trained on imbalanced data produces predictions that are biased in favor of large classes. To overcome this, we re-weighted the instances in the training data so that a positive instance has 10 times the weight of a negative instance. We applied the same approach to KNN, which does not affect the choice of neighbors, but influences the relative contribution of positive and negative neighbors in determining a label. We refer to the re-balanced version of naïve Bayes and KNN as naïve Bayes (c=10) and KNN(c=10) respectively.

We compared all methods using 10 fold cross-validation and computed area under the ROC curve (AUC) as the performance metric. Table 1 summarizes the results in terms of AUC, and Figure 2 presents ROC curves comparing different classifiers. For clarity, the figure only shows the classifiers modified for dealing with imbalanced data.

The results show that accounting for the imbalance in data leads to classifiers with better or comparable performance. In particular, correcting for the skewed distributions in naïve Bayes significantly improves its performance, leading to the best classifier for this data. Given that random classification results in an AUC of 0.5, and a perfect classifier results in an AUC of 1, the naïve Bayes AUC score of 0.883 shows that the model is doing extremely well at ordering the instances in terms of likelihood of being a good customer. These results are very encouraging – in the following sections we explore modeling alternative information sources as well as the possibility of improving on the web content models by incorporating information from these sources.

Table 1: Comparing web content models.

Classifier	AUC
Naive Bayes	0.806
Naive Bayes(c=10)	**0.883**
SVM	0.796
SVM(c=10)	0.833
KNN	0.598
KNN(c=10)	0.597

4. FIRMOGRAPHICS MODELS

In the previous section we demonstrated how website content can be effectively used to identify companies that are likely to align well with particular marketing objectives. However, apart from content of company websites, we can also acquire firmographic information about companies through different sources. While web content can be effective in identifying specific sales strategies, firmographic data, such as size and revenue, can be used to identify a broader pool of candidates based on the viability of a sale or collaboration. Typically, firmographics do not contain much specific information about a company's detailed business alignments, however they may still provide valuable information that

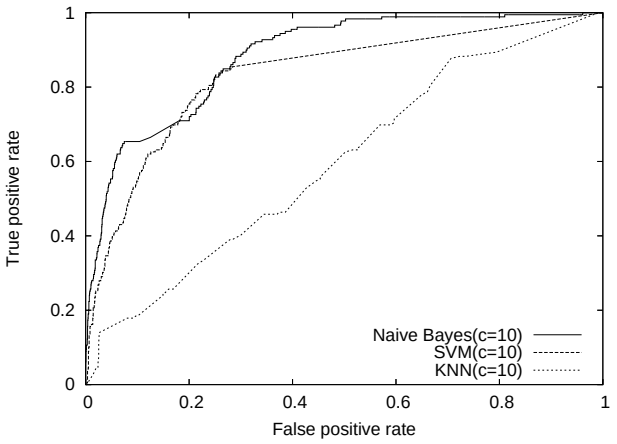

Figure 2: ROC curves comparing web content models.

could be critical in determining potential customers or partners. To verify this, we extracted firmographic data from Dun & Bradstreet (D&B) [1] and IBM Marketing Intelligence [28] which provides information on most businesses in the US and around the world. This data includes information such as:

1. Company size information: revenue and number of employees; along with information on dynamics of change in these figures in recent years.
2. Various levels of industrial classification: industrial sector, industry, sub-industry, etc.
3. Company location: city, state, country.
4. Company structure descriptors: legal status, location in corporate hierarchy (headquarters/subsidiary), etc.

Each company in this data is described by 231 features. We begin by eliminating features that are specific to a company, such as DUNS numbers and names. For ease of modeling, we also exclude categorical features that have too many distinct values, such as zip codes and phone area codes. We remove redundant features that are highly correlated with other features, such as city name and city code. We also eliminate features that are uncorrelated with the target class and are also unlikely to have a causal link with the class label, such as the indicator of a recent address change. Finally, we filter out features that have too many missing values. After these preprocessing steps we are left with 34 features, which are listed in Table 2 in decreasing order of information gain.

We observe that NAICS and SIC codes rank highest on the list of features – this is presumably because they provide the most information regarding business alignments. It is interesting to note that the size of company in terms of number of employees appears to be more important than sales. Furthermore, growth indicators, such as the increase in number of employees and sales in recent years, are far more indicative of a company's classification than the absolute national or global sales. This is consistent with the fact that for the offering under consideration in this data, we are looking for partners that are mid-sized business with the potential to grow.

4.1 Experimental Evaluation

Using the firmographic features listed in Table 2, along with the class labels as before, we compare the following three modeling techniques:

1. J48 [30] — a Java implementation of the C4.5 decision tree algorithm [23].
2. The naïve Bayes algorithm [20], using the Fayyad and Irani approach to discretization [14] of continuous features.
3. Boosted decision stumps, using AdaBoost [26] run for 100 boosting iterations.

To deal with the high imbalance in classes, we re-balance the training data by weighting positive instances 10 times higher than negative instances. Without this re-weighting, decision tree induction (J48) results in a trivial tree with a single leaf node classifying all instances as negative. Figure 3 shows the comparative performance of the different classifiers. The results are summarized in Table 3 in terms of area under the ROC curves. As before, all results were averaged using 10 fold cross-validation.

Boosted decision stumps emerge as being clearly the best approach for this data. These models based solely on firmographics perform surprisingly well at identifying potential customers. However, in absolute terms the firmographics by themselves are not as effective as using the web content (as can be seen by comparing Tables 1 and 3).

Firmographic data provides information that helps identify higher-level characteristics of potential customers, e.g. mid-sized businesses that have been steadily growing. By exploiting industry categorization, the firmographic models can also identify business alignments at a coarse level. For example, the first decision stump in the AdaBoost model learns to classify companies with a NAICS code of 541211 as a positive. This NAICS code corresponds to offices of Certified Public Accountants, which comprises establishments of accountants that are certified to audit the accounting records of public and private organizations and to attest to compliance with generally accepted accounting practices [9]. As described before, in order to market the specific financial service offering for which our data set was created, we are very interested in firms that provide such accounting services. However, in order to be able to further refine our search among all companies within these broad industry classifications, it is crucial that we know the specific services they offer – this is the information we extract from company websites, as done in Section 3.

5. COMBINING INFORMATION FROM MULTIPLE SOURCES

In Sections 3 and 4 we evaluated models built using only web content and firmographics, respectively. The fact that

Table 2: Firmographic features used for modeling, ordered by decreasing information gain.

Feature	Description
NAICS CD	6 digit No. American Industrial Classification Sys
SIC	Standard Industrial Classification code
OFFICE SUPPLY RANK	Based on wholesale buying index. Score 1-100
ST PROVINCE	Name of state or province in which site is located
EMPL RANGE CD	Establishment employee size code
EMPL	Establishment employee size
PC ESTIMATED QTY	Estimates number of PCs at a site
EMPL 5YR PCT	Percent growth in employees (5 year)
CUST PROSPECT CD	Customer or Prospect indicator
WEB PRESENCE CD	Indicates probability of having a Web presence
SALES RNGE CD	Indicates the sales volume range
EMPL 3YR PCT	Percent growth in employees (3 year)
URL STATUS CD	Indicates status of URL for business
NETWORK PC RNGE CD	Estimated number of Nodes or Network connected PCs
STRUCTURE CD	Code for type of business at location
SLE 3YR PCT	Percent growth in sales (3 year)
TECH DEMAND CD	Estimated demand for Technology and Office products
IT BUDGET CD	A ranking of businesses by their likely IT spend
YEAR OWNER CHANGED	Year new owner acquired firm
PTB UNIX SERVER CD	Propensity to buy UNIX servers
YEAR STARTED	Year the Company was Established
SUBSIDIARY CD	Indicates if business is a subsidiary
PTB OTHR SERVER CD	Propensity to buy other servers
WAN PRESENCE CD	Estimated probability of presence of WAN
OFFICE SUPPLY TIER	Ranking of potential to purchase office supplies
PTB WIN SERVER CD	Propensity to buy Windows servers
PUBLIC COMPNY INDC	Indicates if Company is publicly held
SMALL BUSINES INDC	Indicates if enterprise is a small business
NTWK PRESENCE CD	Likely presence of network indicator
WOMEN OWN INDC	Indicates if business is controlled by women
GU SALES US CRCY	Sales for Global Ultimate in whole US dollars
SLE 5YR PCT	Percent growth in sales (5 year)
SALES US CURRENCY	Sales expressed in whole U.S. dollars

Table 3: Comparing firmographics models.

Classifier	AUC
AdaBoost	**0.749**
Naive Bayes	0.692
J48	0.566

it is possible to build effective customer-identification models using each source independently raises the question of whether we can build an even better model by combining these information sources. Although web content is a richer source of information for the task at hand, it is also more susceptible to noise. Automatically mapping company names to their correct URLs in itself is a non-trivial task, and is not 100% accurate. Even with the correct URLs, we end up with lot of noisy (irrelevant) information from company websites such as advertisements, slogans, contact information, etc. On the other hand, firmographic data is more reliable since it is structured and comes directly from database lookups. Hence, web-content and firmographics can be viewed as complimentary sources of information, and by combining them we may be able to leverage the strengths of both sources.

Common approaches to combining information sources vary from early fusion [27], which merges the feature sets extracted from different sources, to late fusion, which combines the output of classifiers built on each features set separately [31]. Following these approaches we explore the following fusion methods:

1. Boosting decision stumps applied to training instances created by merging the web content and firmographic feature sets, which we refer to as AdaBoost (early fusion).

2. Vote-Avg: Build separate classifiers on the web and firmographic features, and average the class probability estimates output by both. We tried two variants of this method using naïve Bayes and SVMs for the web model. In both cases we use AdaBoost for the firmographic models. We refer to the two variants as Vote-Avg (Naive Bayes+AdaBoost) and Vote-Avg (SVM+AdaBoost) respectively.

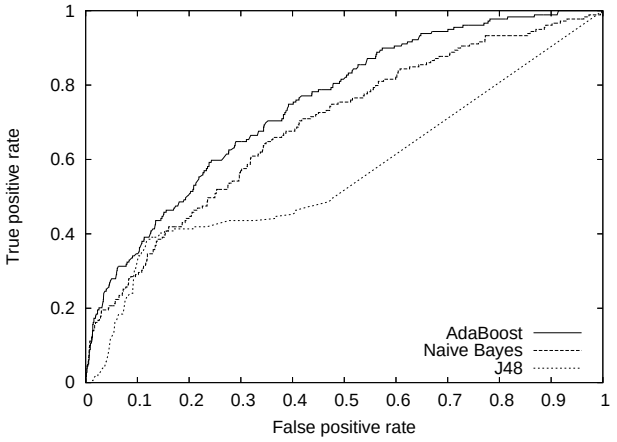

Figure 3: ROC curves comparing firmographics models.

3. Vote-Prod: Follow the same procedure as Vote-Avg, except the class probability estimates produced by individual models are multiplied and renormalized, instead of being averaged. Here again we used the same base model combinations are in Vote-Avg, and refer to them as Vote-Prod (Naive Bayes+AdaBoost) and Vote-Prod (SVM+AdaBoost).

The methods described here can be viewed as *weak* fusion approaches. Below we describe a *strongly-coupled* fusion approach, where the classifiers trained on separate information sources can influence the inductions on each others.

5.1 Transductive Co-training

Given the properties of the data (*i.e.* there are few labeled positive examples, and we have two sources of information available, each of which provides redundant but complementary information), it is natural to apply the co-training algorithm [11]. Originally co-training was developed for semi-supervised learning, which makes use of the unlabeled examples with two distinct sets of features. Following the same idea, we use co-training in the transductive setting, i.e., taking the test set into account during induction and trying to minimize misclassifications of just those particular examples. The details of transductive co-training is shown in Algorithm 1.

The basic assumption of the co-training algorithm is that either set of the features should be sufficient for learning if we had enough labeled data. In our application, this assumption has been obviously violated since there is a lot of noisy or irrelevant information in the web content, while the D&B features do not provide enough information to satisfy the condition. Therefore we do not expect dramatic performance improvement compared with the classifiers using individual sets of features.

As in the case with late fusion, we tried two variants of co-training: naïve Bayes and SVMs for the web model, with AdaBoost for the firmographic models. We refer to these as Co-training (Naive Bayes+Adaboost) and Co-training (SVM+Adaboost), respectively. We use $p = 2$ and $n = 20$ in Algorithm 1, in keeping with the low ratio of positives and negatives in the data.

Algorithm 1 Transductive Co-training

Given:
L : a set of training examples
T: a set of testing examples
$I_{\max}$: maximum iterations

1. Loop until $T = \emptyset$ or for $I_{\max}$ iterations
2. Use L to train a classifier h_1 that uses only the web content features
3. Use L to train a classifier h_2 that uses only the firmographic features
4. Allow h_1 to label p most-confident positive and n negative examples from T
5. Allow h_2 to label p most-confident positive and n negative examples from T
6. Add these self-labeled examples to L

5.2 Experimental Evaluation

The results of all the fusion approaches are summarized in Table 4, and Figure 4 presents ROC curves of different combination techniques. For the sake of clarity we only present ROC curves for three approaches — one each demonstrating early fusion, late fusion and transductive co-training.

Table 4: Comparing methods to combine multiple information sources.

Classifier	AUC
AdaBoost(early fusion)	0.867
Vote-Avg(Naive Bayes+AdaBoost)	0.883
Vote-Prod(Naive Bayes+AdaBoost)	**0.887**
Vote-Avg(SVM+AdaBoost)	0.842
Vote-Prod(SVM+AdaBoost)	0.843
Co-training(Naive Bayes+Adaboost)	0.874
Co-training(SVM+Adaboost)	0.828

Compared with previous approaches using SVMs for the web model (AUC = 0.833, Table 1) and AdaBoost for the firmographic model (AUC = 0.749, Table 3), both early fusion using AdaBoost, as well as both variants of late fusion through voting are successful in improving on the individual models. However, when we use naïve Bayes for the web model (AUC = 0.883, Table 1), only voting with taking products of probabilities (AUC = 0.887) performs better than using only the web content. Furthermore, the added benefit in performance is fairly small.

Taken independently, the web content and firmographic information both lead to useful models for our specific task. However, it also appears that the predictive power realized via the firmographic data can be achieved independently using only the web content. On the other hand, there are also the cases where firmographics helps to correct the ordering of instances of the web models, hence giving rise to the

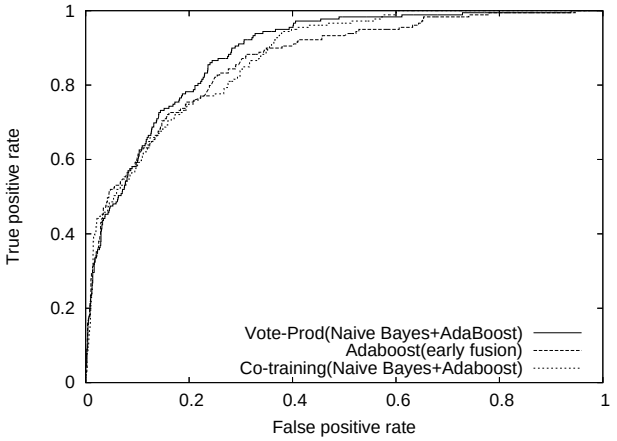

Figure 4: ROC curves comparing combination techniques.

increase, though small, in the area under the ROC curve. Given the high AUC we obtain with the naïve Bayes models on the web content, it is also possible that we are experiencing a ceiling effect, and there is little room for further improvement.

The models built through co-training improve on the firmographic models, but still under-perform the naïve Bayes web models. Blum and Mitchell [11] prove that co-training can learn from unlabeled data starting with only weak classifiers. However, their theoretical guarantees rely on two fundamental assumptions. The first assumption is that the distribution of instances is *compatible* with the target function, i.e., for most instances, the target functions over each feature set predict the same label. In our domain, this means that the class of a company should be identifiable using either the web data or the firmographic data alone. This is clearly not the case, especially since the firmographic data do not provide as much detail in terms of sales strategies of a company. The second assumption is that the two feature sets are conditionally independent of each other, given the class of the instance. This implies that words on a companies web page are not related to the industry classification, and other such firmographic features. This assumption too seems unlikely to hold in practice. Though co-training often works despite the violation of its underlying assumptions [11, 21], it appears that, for this particular domain, it is not as effective, at least in the transductive setting. Given a large pool of unlabeled examples, which is the more common setting for co-training, we may observe better performance.

6. RELATED WORK

In addition to our task in marketing intelligence, there have been many applications in machine learning that share the challenges of combining the information from multiple sources. These include multimedia content analysis from images and text, protein-protein interaction prediction using micro-array data, function annotation as well as sequence information, and sensor networks by combining the data from multiple sensors.

Up to now, the strategies to make use of the information from diverse sources can be summarized as two general approaches, i.e. early fusion, which merges the feature vectors extracted from different data modalities, and late fusion, which combines the output of classifiers built on each single sources [16, 34, 31]. It remains an open question as to which fusion strategy is more appropriate for a certain task, and several comparison studies are discussed for applications in different domains [27, 34]. One extension of the early fusion approach is to derive the latent semantic representation of the data by jointly modeling the low-level features from multiple sources. Possible approaches range from simple methods, such as principal component analysis (PCA), independent component analysis (ICA), Fisher linear discriminant (FLD) and kernel methods, to more sophisticated modeling using graphical models, such as Bayesian model for gene function prediction [29], correspondence latent Dirichlet allocation (Corr-LDA) [10] and dual-wing harmonium models [32] for multimedia applications. These methods have been demonstrated to be more effective than naively joining the low-level features into common feature space. On the other hand, the algorithms in the late fusion approach vary from the equal-weight combination of the sub-classifiers or sub-models, to varied weight with weights learned from cross-validation, and to more adaptive methods with weights depending on the specific testing example [33, 22].

7. CONCLUSION

This paper presents the task of customer identification for companies that sell to other businesses. We formulate this task as a supervised learning problem, and present a case study on an expert-created data set for identifying companies with whom we may wish to partner in order to market a particular financial offering. We analyze the web pages of candidate companies and find that they provide a rich source of information. We demonstrate how we can build highly effective customer-identification models using only this freely available unstructured web content. We also show that, alternatively, we can build models for the same task using data from more structured firmographic information sources. Using firmographics alone can lead to good models, based on coarse-grained characteristics such as industry classifications and dynamics of revenue and employee sizes. However, web content models are more effective because of the richer information available in terms of a company's services, products and sales strategies. Finally, we have explored several approaches to combining the unstructured web content with the structured firmographics data. The results show that by voting classifiers built on the two sources separately, we can get an improvement, albeit small, in the model performance. More sophisticated feature-fusion approaches, such as dual-wing harmonium models [32] may yield better results and provide an avenue for future work.

8. REFERENCES

[1] *Dun and Bradstreet (D&B).* http://www.dnb.com.
[2] *Factiva Dow Jones & Company.* http://www.factiva.com.
[3] *Harte-Hanks Inc.* http://www.harte-hanks.com.
[4] *Hoover's, Inc.* http://www.hoovers.com.
[5] *Reuters.* http://www.reuters.com/.
[6] *Sarbanes-Oxley Act.* http://www.soxlaw.com/.

[7] *Standard & Poor's*. http://www.standardandpoors.com.

[8] *Standard Industrial Classification (SIC)*. http://www.sec.gov/info/edgar/siccodes.htm.

[9] *U.S. Census Bureau*. http://www.census.gov.

[10] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual Int'l ACM SIGIR Conf. on Research and development in information retrieval*, pages 127–134, New York, NY, USA, 2003. ACM Press.

[11] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI, 1998.

[12] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval*. Springer-Verlag, 1994.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, second edition, 2001.

[14] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 1022–1027, Chambéry, France, 1993. Morgan Kaufmann.

[15] E. Frank and R. R. Bouckaert. Naive bayes for text classification with unbalanced classes. In *Proc 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 503–510, 2006.

[16] G. Iyengar and H. J. Nock. Discriminative model fusion for semantic concept detection and annotation in video. In *Proc. of the 11th ACM Int'l Conf. on Multimedia*, pages 255–258, New York, NY, USA, 2003. ACM Press.

[17] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the Tenth European Conference on Machine Learning (ECML-98)*, pages 137–142, Berlin, 1998. Springer-Verlag.

[18] R. Lawrence, C. Perlich, S. Rosset, J. Arroyo, M. Callahan, M. Collins, A. Ershov, S. Feinzig, I. Khabibrakhmanov, S. Mahatma, M. Niemaszyk, and S. Weiss. Analytics-driven solutions for customer targeting and sales force allocation. *IBM Systems Journal*, 2007.

[19] A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. In *Papers from the AAAI-98 Workshop on Text Categorization*, pages 41–48, Madison, WI, July 1998.

[20] T. Mitchell. *Machine Learning*. McGraw-Hill, New York, NY, 1997.

[21] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM-2000)*, pages 86–93, 2000.

[22] W. S. Noble and A. Ben-Hur. Integrating information for protein function prediction. *Bioinformatics - From Genomes to Therapies*, 3, 2007.

[23] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo,CA, 1993.

[24] J. Rennie, L. Shih, J. Teevan, and D. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pages 616–623, 2003.

[25] S. Rosset and R. Lawrence. Data enhanced predictive modeling for sales targeting. In *Proceedings of SIAM Conference On Data Mining*, 2006.

[26] R. E. Schapire. Theoretical views of boosting and applications. In *Proceedings of the Tenth International Conference on Algorithmic Learning Theory*, pages 13–25, 1999.

[27] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.

[28] Z. Su, J. Jiang, T. Liu, G. Xie, and Y. Pan. Market intelligence portal: an entity-based system for managing market intelligence. *IBM Systems Journal*, 43(3), 2004.

[29] O. Troyanskaya, K. Dolinski, A. Owen, R. Altman, and D. Botstein. A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proc Natl Acad Sci*, 100, 2003.

[30] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, 1999.

[31] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *Proc. of the 12th annual ACM Int'l Conf. on Multimedia*, pages 572–579, 2004.

[32] E. Xing, R. Yan, and A. Hauptmann. Mining associated text and images with dual-wing harmoniums. In *Proceedings of the 21th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-05)*. AUAI press, 2005.

[33] R. Yan. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. PhD thesis, Carnegiel Mellon University, 2006.

[34] R. Yan, J. Yang, and A. G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *Proc. of the 12th ACM Int'l Conf. on Multimedia*, pages 548–555. ACM Press, 2004.

[35] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann Publishers, San Francisco, US, 1997.