Electronic Commerce Research and Applications 10 (2011) 418–427

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

journal homepage: www.elsevier.com/locate/ecra

Mining competitor relationships from online news: A network-based approach

Zhongming Ma [a,*], Gautam Pant [b,1], Olivia R.L. Sheng [b,2]

[a] Computer Information Systems Department, California State Polytechnic University, Pomona, 3801 West Temple Avenue, Pomona, CA 91768, United States
[b] Department of Operations and Information Systems, The University of Utah, 1645 East Campus Drive, Salt Lake City, UT 84112, United States

ARTICLE INFO

Article history:
Received 6 March 2010
Received in revised form 19 November 2010
Accepted 24 November 2010
Available online 30 November 2010

Keywords:
Web mining
Classification in networked data
Competitor discovery
Business news

ABSTRACT

Identifying competitors is important for businesses. We present an approach that uses graph-theoretic measures and machine learning techniques to infer competitor relationships on the basis of structure of an intercompany network derived from company citations (cooccurrence) in online news articles. We also estimate to what extent our approach complements the commercial company profile data sources, such as Hoover's and Mergent.

1. Introduction

Scanning the competitive environment of a company or a group of companies represents an essential facet of businesses. To gather information about competitor relationships, people resort to various options, such as asking business associates, reading news articles, searching the Web, attending business conventions, or looking through paid company profile resources such as Hoover's (http://www.hoovers.com) and Mergent (http://www.mergentonline.com). Although the company profiling resources have reduced search efforts and made some business relationship information easily accessible, due to their limited resources and/or differences in criteria, they can suffer from a scalability problem and provide incomplete information. For example, Hoover's considers Interchange Corp. a competitor of Google, whereas Mergent does not specify such a relationship. In contrast, Mergent includes Tercica Inc. as a competitor of GlaxoSmithKline plc, whereas Hoover's does not. Approaches that automatically discover important business relationships could complement and expand on existing, resource-intensive efforts.

Bao et al. (2008) observe that a company is more likely to cooccur with its competitors in web pages than with non-competitors. We recognize that simply because a company is mentioned in a news story about another company does not necessarily imply that the two companies are competitors. However, by aggregating and analyzing company citations from tens of thousands of news articles and by considering graph-theoretic properties of a network of companies derived from the citations, automated techniques could learn to recognize patterns for identifying competitors. A novelty of this research is in the use of structural attributes of a network derived from seemingly noisy data (company citations in news) to discover knowledge (i.e., competitor relationships), given the fact that news stories are generally not written to explicitly describe such relationships. Furthermore, our proposed approach is language neutral in that it does not apply natural language processing (NLP) methods on news beyond recognizing the tickers of companies. Our approach consists of the following four main steps.

1. Given a collection of news stories organized by company, we identify company citations in news stories. The first step is described in Sections 4.1 and 4.2.
2. We construct a directed, weighted intercompany network from the company citations, and identify four types of attributes from network structure which differ in their coverage of the intercompany network. After finding company pairs, we label some randomly selected pairs according to Hoover's and Mergent. Sections 4.3–4.5 explain this step.
3. We use the four types of attributes and competitors identified from Hoover's and Mergent to train classifiers that learn to infer competitor relationship between a pair of companies linked in the network, and evaluate the competitor classification performance on the basis of several metrics (e.g., precision, recall, false positive rate, F_1, etc.) with four different classifiers. In particular, noticing the issue of imbalanced data set, we tackle the problem with two different techniques, decision threshold adjustment (DTA) and undersampling-ensemble (UE). This step is presented in Section 5.

* Corresponding author. Tel.: +1 909 869 3242; fax: +1 909 869 3248.
E-mail addresses: zma@csupomona.edu (Z. Ma), gautam.pant@business.utah.edu (G. Pant), olivia.sheng@business.utah.edu (O.R.L. Sheng).
[1] Tel.: +1 801 585 3360.
[2] Tel.: +1 801 585 9071.

1567-4223/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.elerap.2010.11.006

4. Considering Hoover's and Mergent as gold standards (i.e., information sources that contain reliable data), we estimate the coverage (in terms of identifying pairs of competitors) of each gold standard in an (unknown) competitor space. In addition, we estimate the extension provided by our approach to each gold standard. The last step is covered in Section 6.

This study focuses on the following two research questions regarding the proposed approach:

1. How effectively can we discover competitor relationships between companies using four types of attributes derived from the intercompany network? Especially, given a big portion of our data set is considered imbalanced (i.e., number of non-competitor pairs $\gg$ number of competitor pairs), we apply special classification techniques to deal with the imbalance and report their classification performance.
2. To what extent can Hoover's and Mergent cover the set of all competitors, and to what extent does the proposed approach extend the competitor coverage by Hoover's and Mergent?

2. Literature review

Network structure has been shown to help classification in networked data. When classifying hypertext, Chakrabarti et al. (1998) find that using pages' hyperlink structure significantly improves classification as compared to using only text of the pages. Hogg (2010) demonstrates the benefits of network links in inferring preference correlations. In a systematic experimental study with twelve data sets, Macskassy and Provost (2007) compare various learning and inference techniques and find that a weighted-vote relational neighbor (wnRN) classification model often performs well, and the simple wnRN classifier makes predictions using the class labels of related neighbors and thus does not involve any learning or use any inherent attributes such as text of a page. Becchetti et al. (2008) propose a link-based technique for automatic detection of spamming web sites. They compute structural attributes, such as indegree, outdegree, PageRank and TrustedRank scores, of a URL-link network, feed those attributes to a decision tree classifier, and report classification performance (i.e., F_1) on the basis of different combinations of those network attributes. By analyzing network structure (e.g., k-core) and using structural attributes from a seller–buyer transaction network, Wang and Chiu (2008) improve online recommendation system on trusted auction sellers. Using structural information from linked bloggers, Bhagat et al. (2007) infer certain properties, such as age and location, of bloggers. In a different work, based on some of the structural attributes used in this paper, we predict the company revenue relation (CRR) that is derived from two companies' relative quantitative financial data (Ma et al. 2009b). In comparison to our preliminary study (Ma et al. 2009a) that presents the current approach of competitor discovery using the structural attributes, this paper contains a new attribute and more results on competitor classification and competitor extension.

Link prediction estimates the existence of a link between two nodes on the basis of attributes of the nodes and link structure of a network (Getoor and Diehl 2005). According to Getoor and Diehl (2005), one approach to prediction is entirely based on network structural properties. Using structural properties derived from network representing coauthorships among physicists, Liben-Nowell and Kleinberg (2007) predicts a coauthorship between two physicists who did not have such a relationship in past. They compute values for a variety of network structure attributes and compare them with those from random guess. Through trial and error with identified structural attributes Karamon et al. (2008) identify desired features and use decision tree classifier for link prediction.

The main difference between our current study and the abovementioned work is that they do not focus on discovering business relationships between companies. In addition, we predict competitor relationship not from a network constructed from given competitor relationships but from a citation-based intercompany network where a citation represents a company being mentioned in a news story belonging to another company. While a citation often does not indicate a direct business relationship, structural attributes derived from such a citation-based intercompany network can enhance the ability to infer competitor relationships from an otherwise noisy network.

Researchers in areas such as organizational behavior and sociology also have investigated the nature and implications of social networks created by business relationships. In his highly cited research, Granovetter (1985) argues that most economic behavior in model society is embedded in social relation networks. And many researchers analyze economic phenomena from social network structure. For example, using a commercial directory of biotechnology firms as their data source, Walker et al. (1997) demonstrate that network structure strongly influences the choices of a biotechnology startup in terms of establishing new relationships (licensing, joint venture, R&D partnership) with other companies. Uzzi (1999) investigates the effect of social relationships and networks on a firm's acquisition and cost of capital. Furthermore, Gulati and Gargiulo (1999) demonstrate that an existing interorganizational network structure influences the formation of new alliances, which eventually modifies the existing network. However, these prior studies use explicitly specified/given relationships, often from reliable data sources such as commercial or government databases, surveys, or interviews, to construct a social network. In contrast, we attempt to discover/predict business relationships from an intercompany network constructed by company citations which do not represent specific business relationships. And we want to examine whether such a network built from seemingly noisy citation-based links can still provide us interesting information.

Using ClearForest, a commercial text analytics software, to identify companies from Yahoo! News, Bernstein et al. (2002) construct an undirected and unweighted (binary weight) intercompany network. All companies are linked to each other if they cooccur in the same piece of news. They filter the network down to a few hundred companies, rank each company according to its connections with other companies, and report that some of the 30 top-ranked companies in the computer industry are also *Fortune* 1000 companies. In another study, with the same data set Bernstein et al. (2003) predict a company's industry sector using the sector information of its neighbors in an intercompany network. Without applying the ClearForest software the authors use stock tickers in news to identify companies (Bernstein et al. 2003). Hence, they take advantage of the ticker feature that is provided by Yahoo! Finance. In the current work, we also use tickers to identify companies. However, compared with Bernstein et al. (2002, 2003), we qualify links in the constructed network by both direction and weights. More importantly, we study a different problem: we employ a variety of graph-theoretic metrics to predict the competitor relationship between companies and estimate the competitor coverage of gold standards and the extension provided by our approach to each gold standard.

Bao et al. (2008) presents an NLP approach that can extract competitors from web search results (i.e., snippets) given a query, such as company name. Different from their work, our approach is language neutral (we do not resort to NLP techniques to analyze news content) when constructing an intercompany network and we resort to network structural attributes to infer competitor relationships.

3. Background knowledge: graph-theoretic attributes

We identify four types network structural attributes on the basis of the range of the network used to compute the attributes: dyad degree-based, node degree-based, node centrality-based, and structural equivalence-based. In another study (Ma et al. 2009b) we have presented the first three types of the following attributes. For the sake of completeness and readability, we introduce below all of the four types of attributes and describe how they are generated.

3.1. Two types of degree-based attributes

Fig. 1 shows a very small portion of the intercompany network that consists of five companies/nodes joined by 15 directed, weighted links. In this intercompany network, degree reflects the flow (inward, outward, or both) of citations from/into a node or between two nodes.

We first introduce a group of dyad (i.e., pairwise) degree-based attributes as follows:

- Weight of dyad indegree (WDID), such that WDID(n_i, n_j) is the weight of the link from n_j to n_i.
- Weight of dyad outdegree (WDOD), such that WDOD(n_i, n_j) is the weight of the link from n_i to n_j.
- Net weight of dyad (NWD), such that
$$\mathrm{NWD}(n_i, n_j) = \mathrm{WDOD}(n_i, n_j) - \mathrm{WDID}(n_i, n_j). \quad (1)$$
- Weight of dyad inoutdegree (WDIOD), such that
$$\mathrm{WDIOD}(n_i, n_j) = \mathrm{WDOD}(n_i, n_j) + \mathrm{WDID}(n_i, n_j). \quad (2)$$

In Fig. 1, WDID, WDOD, NWD, and WDIOD for the link of (YHOO, GOOG) is 478, 512, 34, and 990, respectively. Consistent with the observation of Bao et al. (2008) that a company is more likely to cooccur with its competitors than with non-competitors, we expect that a large WDID, WDOD, and/or WDIOD value may indicate a stronger relationship between the given pair of companies.

To take into account a node's neighbors, we also consider the following node degree-based attributes.

- Weight of node indegree (WNID), such that
$$\mathrm{WNID}(n_i) = \sum_{n_j \in NB_i} WDID(n_i, n_j), \quad (3)$$

where NB_i denotes all of n_i's neighbors, and the equation measures the flow of citations from all companies in the network to the focal company. We expect "important" companies to draw a greater number of total citations from other companies as indegree often represents prestige (Wasserman and Faust 1994) or authoritativeness (Kleinberg 1999).

- Weight of node outdegree (WNOD), such that
$$\mathrm{WNOD}(n_i) = \sum_{n_j \in NB_i} WDOD(n_i, n_j), \quad (4)$$
which measures the flow of citations from the focal company to all other companies in the network. The outdegree is often considered a simple measure of centrality (Wasserman and Faust 1994).
- Weight of node inoutdegree (WNIOD), or
$$\mathrm{WNIOD}(n_i) = \sum_{n_j \in NB_i} WDIOD(n_i, n_j), \quad (5)$$
which measures the overall flow of citations both to and from the focal company (n_i). In essence, this attribute measures the overall connectivity of the company and all its neighbor companies in the network, independent of the direction of those citations.

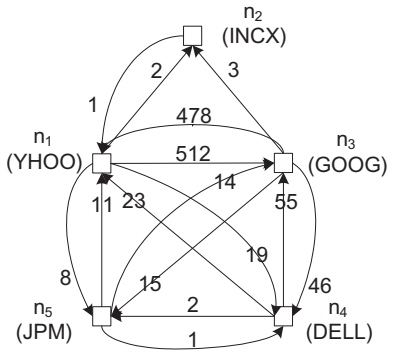

Fig. 1. Directed and weighted graph derived from news citations. *Notes*: DELL: Dell Inc., INCX: Interchange Corp., GOOG: Google Inc., JPM: JP Morgan Chase & Co., YHOO: Yahoo! Inc.

3.2. Centrality-based attributes

In addition to dyad and node degree-based measurements, we use a network analysis package, JUNG (O'Madadhain et al. 2006), to compute scores on the basis of three different centrality/importance measure schemas: PageRank (Brin and Page 1998), HITS (Kleinberg 1999), and betweenness centrality (Brandes 2001). These schemas extend beyond immediate neighbors to compute the importance or centrality of a given node across the whole network. The PageRank algorithm computes a popularity score for each Web page on the basis of the probability that a "random surfer" will visit the page (Brin and Page 1998), whereas the HITS algorithm as implemented by O'Madadhain et al. (2006) generates an authority score for each page. Both HITS and PageRank compute principal eigenvectors of matrices derived from graph representations of the Web (Kleinberg 1999), so our use of them in a graph whose nodes refer to companies differs from their original use. Furthermore, as another node centrality measurement, betweenness measures the extent to which a node lies between the shortest paths of other nodes in the graph (Freeman 1979). These global centrality attributes use the same underlying intuition as that for the node degree-based attributes but could be more informative because they consider the entire network instead of focusing on immediate neighbors. We expect that a more important company is more likely to have a relationship with a given company than is a less important one, thus we use those node global centrality attributes and node degree-based attributes because they represent the importance (e.g., prestige or centrality) of a node in the whole network.

3.3. Structural equivalence (SE)-based attributes

Lorrain and White (1971) identify two nodes as structurally equivalent if they have the same links to and from other nodes in the network. Because it is unlikely that two nodes will be exactly structurally equivalent in our intercompany network, we use a similarity metric to measure their degree of structural equivalence (SE). To represent the intercompany network, we use a weighted adjacency $N \times N$ matrix, where N is the number of nodes. The SE between two nodes is the normalized dot product (i.e., cosine similarity) of the two corresponding rows in the matrix, in which an element can be WDID, WDOD, or WDIOD and therefore produce WDID-, WDOD-, or WDIOD-based SE similarity. Intuitively, the WDID-based SE similarity between company A and company B captures the overlap between companies whose news stories cite A and companies whose news stories cite B (analogous to co-citation (Small 1973)); the WDOD-based SE similarity reflects the overlap between companies that news stories of both A and B cite

(analogous to bibliometric coupling (Kessler 1963)). A high overlap between the neighbors of two nodes in our intercompany network may reflect overlap in their businesses or markets, which could indicate a competitor relationship.

The four types of attributes represent a wide variety of network properties suggested in the social network analysis and graph-theoretic literature and they differ in amount of network covered by them. In Table 1, we summarize these attributes by type and range of network covered.

4. Data processing

4.1. Raw data

The raw data set consists of eight months (July 2005–February 2006) of business news for all companies on Yahoo! Finance. We include all companies across all nine industry sectors (basic materials, conglomerates, consumer goods, financial, healthcare, industrial goods, services, technology, and utilities) used in Yahoo! Finance.

4.2. Preliminary data processing

Yahoo! Finance organizes business news from yahoo.com, forbes.com, thestreet.com, businessweek.com, and other sources by company. Taking advantage of this organizing mechanism, we programmatically fetch news stories for each company during the eight-month period. We observe that very often Yahoo! organizes the same piece of news under different companies if the news contains stock tickers for those companies; we treat such a news story as belonging to each of the companies that Yahoo! identifies for the story. For example if a news article mentions companies A (identified by its ticker) twice and company B once and it is organized under each of the companies, from this piece of news we derive WDID, WDOD, and WDIOD as 2, 1, and 3, respectively, for the pair (A, B).

4.3. Node and link identification

A news story in Yahoo! Finance mentions a company according to its stock ticker. This feature makes company identification easy. Thus we resort to these stock tickers to identify the companies mentioned in a given news story. It restricts our current analysis to publicly traded companies, but note that our general approach can be extended to any type of company as long as there is a way to recognize the companies in news stories. If a news story belonging to company n_i mentions another company n_j, we identify a directed link from n_i to n_j, denoted as (n_i, n_j). If company n_j appears several times in the same piece of news, each citation adds to the accumulated weight for that directed link. We aggregate the citation count for the directed company link across all news stories but do not count self-references. That is, we ignore citations to company n_i if they appear in a news story belonging to n_i. Our final data set consists of 6428 companies and 60,532 news stories.

In the following we introduce two data sets that we use to evaluate competitor classification performance. The first data set represents the entire set of company pairs in the network, and the second one represents the imbalanced part of the network.

Table 1
Four types of network attributes.

Attribute type	Attributes	Range of network covered
Dyad degree-based	WDID, WDOD, WDIOD	A given node and only one directly connected node
Node degree-based	WNID, WNOD, WNIOD	A given node and all directly connected nodes
Node centrality-based	Pagerank, hits, betweenness	Whole network
SE-based	WDID-, WDOD-, WDIOD- based SE similarity	Any two nodes and their directly connected nodes in the whole network

4.4. Instance selection and labeling of 840 selected pairs

We first use NWD (WDOD – WDID, a net flow of citations between a pair of companies) to identify all distinct company pairs in the network without considering direction. To obtain distinct company pairs we include only pairs with non-negative NWD values, and for any link (n_i, n_j) with a NWD value of 0, we ignore the opposite link (n_j, n_i). In other words, all distinct company pairs in the intercompany network that have any citations between them are identified. For the entire intercompany network, we identify a total of 87,340 company pairs. Next, we sort the company pairs by their WDIOD values, which range from 1 to 990, in descending order, because WDIOD captures the total volume of citations between two companies in news. In terms of WDIOD values, the data set is skewed/imbalanced; most company pairs have small WDIOD values. Recall that according to Bao et al. (2008), a company is more likely to cooccur with its competitors in web pages than with non-competitors. Similarly, we conjecture that more citations between two companies in news stories should increase the likelihood that two companies have a business relationship. We group company pairs with the same or similar WDIOD values (which represent the combined citations between two focal companies) by dividing those pairs into baskets, such that links with different WDIOD values do not appear in the same basket unless the basket contains fewer than 200 pairs. This procedure results in 21 baskets each of which is associated with the same or similar WDIOD values. Then we randomly choose 40 company pairs from each basket to form a sample basket, and we name the resulting 840 selected company pairs *data set 840*, which we will later use to examine the classification performance for company pairs in the individual baskets that inherently have different WDIOD values.

We manually determine whether each of the 840 company pairs in the 21 sample baskets is a competitor pair using the Hoover's and Mergent sources. We could not automatically derive this data from Hoover's and Mergent (through a Web agent or crawler) because the two sources (at the time of this data collection) restrict such access to their proprietary data. If we find a competitor relationship between the two companies according to either Hoover's or Mergent, we assign the pair a class label of 1 (positive instance); otherwise, it receives a class label of 0 (negative instance). Compared with the first 17 sample baskets, the last four (sample baskets 18–21) are more imbalanced in that they contain no more than 10% of positive instances.

4.5. Instance selection and labeling of imbalanced data set

In an imbalanced data set, most instances occur in one class, whereas the minority is labeled as the other class, and the latter typically is the more important class (Kotsiantis et al. 2006). As prior research (e.g. Weiss and Provost 2003) show that typical classification methods fail to detect the minority in an imbalanced data set and they generate poor precision and recall (e.g., close to 0%) for the positives (i.e., the minority class). In our case, the positives are the competitor pairs. The main reason for this poor performance is that the classifiers, by default, maximize accuracy and therefore give more weight to majority classes than minority ones (Kotsiantis et al. 2006). For example, for a data set where only 1% of the instances have a positive label, simply assigning every instance a

negative label and not detecting any positives achieves an accuracy of 99%. To handle the imbalanced data set problem, we first create a larger data set, *imbalanced data set 2000*, by proportionally (according to basket size) sampling a total of 2000 company pairs from the four imbalanced baskets (18, 19, 20, and 21) that have the lowest ratio of positives ($\leqslant$10%). As before, we manually label the 2000 company pairs using Hoover's and Mergent.

Moreover we also combine the 17 (more balanced) sample baskets (1–17) in data set 840 and all the 2000 company pairs in imbalanced data set 2000 in order to calculate estimated overall performance for the whole 87,340 company pairs. For convenience, we call this combination of the two data sets *combined data set 2680*, which contains 18 sample baskets, and the imbalanced data set 2000 provides the eighteenth sample basket.

5. Competitor discovery

In this section we present classification results for data set 840 and imbalanced data set 2000, and estimate overall performance on the basis of results from combined data set 2680.

5.1. Evaluation metrics

Table 2 is the confusion matrix containing the actual and classified classes for a classification problem with two class labels. *TP* refers to the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* represents the number of false negatives. Using the confusion matrix, we employ the following common metrics for evaluating and comparing classification performance: precision, TP rate (recall), FP rate $\left(\frac{FP}{TN+FP}\right)$, accuracy, and F_1 (Salton 1971).

One of the most common metrics to evaluate classifiers for an imbalanced data set is the receiver operating characteristics (ROC) curve (Kotsiantis et al. 2006), a two-dimensional curve with *TP rate* (recall) on the *y*-axis and *FP rate* on the *x*-axis. Thus, a ROC curve can address an important tradeoff—namely, the number of correctly identified positives or true positives increases at the expense of introducing additional false positives. The area under ROC, which is called *AUC*, also offers an evaluation metric.

5.2. Competitor classification with data set 840

Using the publicly available Weka API (Witten and Frank 2005), we employ four classification methods: artificial neural network (ANN), Bayes net (BN), C4.5 decision tree (DT), and logistic regression (LR) to classify company pairs. Models based on ANN, BN, and DT are common classifiers in data mining, and LR frequently appears in business research to address problems with a binary class label (as in our competitor classification problem). We employ four popular classification methods so as to compare their performance and to allow a user to decide a proper one based on his or her specification. For each sample basket of data set 840, except for basket 21, which does not contain any competitor pairs (we address this basket, together with three other baskets as the imbalanced data set 2000, in the following section), in Fig. 2 we report the average precision, recall, and accuracy generated by 10-fold cross validation for the ANN method.

Table 2
Confusion matrix.

Actual class label	Classified class label	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

In Fig. 2, we also include the prior distribution of positives in each sample basket for comparison. The precision curve is almost always above the prior probability, except for the last two sample baskets with the lowest prior distributions (5.0% and 2.5%). Though for most baskets ANN's classification performance is reasonably good, it weakens when WDIOD values are very small (in last few baskets). This result highlights the inherent challenge of accurately classifying the minority class for imbalanced data sets (the last few baskets although their accuracy remains high). The other three classification methods (BN, DT, and LR) show similar performance patterns but poorer performance overall.

5.3. Competitor classification with imbalanced data set 2000

5.3.1. Background on handling imbalanced data sets

The solutions for handling imbalanced data sets for classification problems appear in two broad categories—data-oriented and algorithmic. Several data-oriented solutions use different resampling approaches, such as undersampling majority, oversampling minority, or oversampling minority by creating a synthetic minority (Chawla et al. 2002), which changes the prior distribution of the original data set (Kotsiantis et al. 2006) before learning from the data set. Another approach at the data level segments the whole data into disjoint regions, such that the data in certain region(s) are no longer imbalanced (Weiss 2004).

Some popular solutions at the algorithmic level include the following:

- Decision threshold adjustment (DTA), which, given a (normalized) probability of an instance being positive (or negative), changes the probability threshold used to determine the class label of the instance (Provost 2000).
- Cost-sensitive learning (CSL), which assigns fixed and unequal costs to different misclassifications, such as cost(false negative) > cost(false positive), to minimize the misclassifications of positives (Pazzani et al. 1994).
- Recognition-based learning (RBL), which, unlike a two-class classification method that learns rules for both positive and negative classes, is a one-class learning method and learns only rules that classify the minority (Weiss 2004, Kotsiantis et al. 2006).

We employ several of these techniques to address our imbalanced data set. Specifically, we divide the whole data set into 21 baskets on the basis of WDIOD, and many of these turn out to be more "balanced" than the entire data set, so it matches the segment data approach (Weiss 2004) for handling imbalanced data sets. For the few imbalanced baskets, we sample more instances to form the imbalanced data set 2000. Next we apply two different approaches, DTA approach and an undersampling-ensemble (UE) method, to address the imbalanced data set problem. We choose DTA due to its simplicity and select UE as a representative of resampling approach. We do not choose the CSL approach, mostly because we do not know the right ratio for the cost of FN versus the cost of FP in the context of our competitor classification problem. However, we consider DTA and CSL to be very similar, in that they both create a bias toward positive classifications. For imbalanced data set 2000, we report various performance metrics suited for an imbalanced data set, including F_1, precision, TP rate, FP rate, ROC, AUC, and accuracy. Next we briefly review the two approaches (DTA and UE) for dealing with classification of imbalanced data.

5.3.2. The DTA approach

With this approach, we simply adjust the decision threshold used by a classifier to determine whether to classify an instance as positive or negative, given its (normalized) probability of being

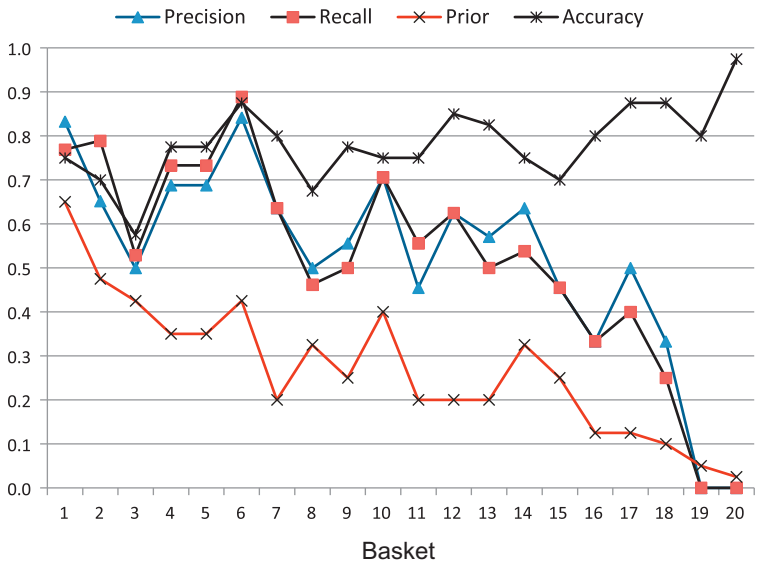

Fig. 2. Classification results for Data Set 840 by ANN and the prior distribution.

positive. For example, given that Pr(x is positive) = 0.3, the instance x is labeled negative when the decision threshold is 0.5. However, when the threshold is adjusted to 0.2, x is classified as positive.

For training and testing, we follow strict tuning procedures suggested by Salzberg (1997) and describe our steps as follows. We randomly select 1500 instances as a training set from the imbalanced data set 2000 and the remaining 500 as the testing set. For each classification method, we use 10-fold cross validation and tune the input parameters to observe the best performance on the F_1 measure with just the training set. Finally, we apply each trained classifier with its respective "best" parameter setting to the testing set for evaluation purposes. Moreover, to determine robustness, we randomly divide the 2000 company pairs into four disjoint sets of equal size, which form four different pairs of training and testing sets ($C_4^3 = 4$). We then apply the training–tuning–testing procedures to the four pairs of training and testing sets and report the average results (see the Eq. (6)). In each case, training and parameter tuning relies solely on the training data set, whereas our evaluation uses only the testing data set. During training, for ANN we tune the learning rate from 0.1 to 1.0 and momentum from 0.1 to 0.3; for BN, we choose K2 (Cooper and Herskovitz 1992) and TAN (Friedman et al. 1997) as algorithms for the search network structure; for DT, we change the minimum leaf size from 2 to 10; and we require no parameter tuning for LR. For all other parameters, we accept the default from Weka. We apply the same tuning procedures throughout the study whenever we use parameter tuning.

5.3.3. The UE approach

Due to space concerns, we do not review the UE approach here. Interested readers can refer to Chan and Stolfo (1998). Fig. 3 illustrates the key idea for the UE approach. The final classifier is based on majority vote by count (MVC) or majority vote by probability (MVP).

During the training phase, from the initial ratio of positives in the subsets, we tune the parameters for each classifier (except for LR) and record its performance in an output file. We repeat this procedure with different ratios of positives, which change from 0.05 to 0.60 with a step size of 0.05. From all output files, on the basis of the best performance on the F_1 measure, we determine a set of best parameters for a classifier and a best ratio of positives. Finally, we apply the trained classifiers with their best parameter

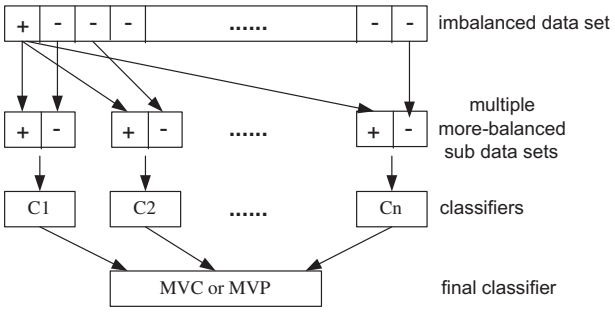

Fig. 3. Generating more-balanced sub data sets for an ensemble classifier.

settings and best ratios of positives to the testing set for evaluation. Similarly, we divide the 2000 company pairs into four disjoint sets of equal size, generate results separately for the four pairs of training and testing sets, and report the average results.

5.3.4. Classification performance for imbalanced data set 2000

In Table 3, we report the precision, TP rate (recall), FP rate, F_1, accuracy, and AUC of the testing sets for each classification method using the DTA approach. Each number in bold indicates the best performance for a measurement across the four classification models. Because we have four pairs of training (1500 instances) and testing (500 instances) sets (as described in Section 5.3.2), we generate and report performance across the four testing sets. The following is the equation for precision.

$$precision = \frac{\sum_{i=1}^{4} TP_i}{\sum_{i=1}^{4} (TP_i + FP_i)}. \quad (6)$$

Here the subscript i, a number between 1 and 4, denotes the four disjoint testing sets. For brevity, the equations for other performance measurements, which are similar to (6), are not included here.

Table 3 also contains results for the same data set with and without sector information. A simple local information of node (i.e., industry sector) improves the classification performance for imbalanced data set 2000 across the four classifiers; for example, the maximum F_1 measures (both produced by ANN) increase by 63%. With sector information, we do not observe a significant

Table 3
Classification performance of imbalanced data set 2000 by DTA approach.

Avg. performance	Without sector information				With sector information[a]			
	ANN	BN	DT	LR	ANN	BN	DT	LR
Precision	0.268	0.125	0.185	0.322	0.372	0.262	0.283	0.380
Recall	0.220	0.240	0.230	0.190	0.420	0.430	0.360	0.380
False positive rate	0.032	0.088	0.053	0.021	0.037	0.064	0.048	0.033
F_1	0.242	0.164	0.205	0.239	0.394	0.326	0.317	0.380
Accuracy	0.931	0.878	0.911	0.940	0.936	0.911	0.923	0.938
AUC	0.736	0.672	0.610	0.723	0.858	0.853	0.741	0.834

[a] Company's sector from Yahoo! Finance is included as an attribute.

difference in the F_1 measure across the 20 baskets in data set 840 (two-tailed t-test, $p = 0.827$). We think other relevant local information, such as product category, can be helpful to competitor discover too. However, not only is identifying such local information typically difficult, but also it is not the focus of this study. Thus we limit to only sector which is relative easy to obtain. We note that data set 840 with an average of 28.75% positives is much more balanced than imbalanced data set 2000, which has only 5.0% positives. Hence, competitor discovery in data set 840 is an easier problem and hence sector information is not as valuable as it is in case of imbalanced data set 2000. However, for this imbalanced data set 2000 the performance of using sector alone and ignoring all four types of network attributes is inferior (e.g., $F_1 = 0.199$) to using network attributes with sector ($F_1 = 0.394$) and without sector ($F_1 = 0.242$). We compare the F_1 measure because results in Table 3 are generated on the basis of model parameters that provide the best F1 value.

The UE approach with MVC and MVP produces similar results as those in Table 3. For example, with MVC, the maximum values of the F_1 measures are 0.381 and 0.204 with and without sector information, respectively. Although the UE approach is more complex than the simple DTA approach, in that it requires an undersampling of majority class to form multiple smaller data sets and adjusting ratios of positives in these small data sets, the two methods show similar classification performance. Thus, in next section when estimating the extent to which our approach extends the gold standards, we use the results from the DTA approach.

5.4. Estimated overall classification performance on the basis of combined data set 2680

Our classification performance measurements thus far compute values for each sample basket. Because sample baskets consist of randomly selected links from the original (larger) baskets, these performance results represent the performance on the original baskets. However, we also want to estimate the classification performance for all of the baskets combined, or the whole data set with its 87,340 company pairs. This estimation requires that we extrapolate the performance observed in the sample baskets to the entire original basket. So we estimate overall precision, TP rate (recall), FP rate, accuracy, and F_1 using combined data set 2680. For the 17 sample baskets from data set 840, the classification results are based on 10-fold cross validation, whereas for the eighteenth sample basket, we combine and use the results generated from the four disjoint testing sets (each with 500 instances). The estimated overall precision is computed through the following equation:

$$\text{estimated overall precision} = \frac{\sum_{i=1}^{18} TP_i \times \frac{B_i}{S_i}}{\sum_{i=1}^{18} (TP_i + FP_i) \times \frac{B_i}{S_i}}. \quad (7)$$

where B_i is the size of basket i, and S_i is the size of sample basket i. Please note that for different i, $\frac{B_i}{S_i}$ has a different value. The equations for other performance metrics are similar and not included for brevity.

We estimate the overall classification performance by extending performance measurements for a sample basket to the corresponding full basket and then combining the measures across the 18 baskets in combined data set 2680. For example, if the sample basket S_i, which represents the original basket B_i, contains m instances that are classified as positives by a classification model, we expect the original basket B_i to contain $m \times \frac{B_i}{S_i}$ instances that would be classified as positives by the same model. Our measurements, such as that shown in Eq. (7), estimate the overall classification performance for the whole data set of 87,340 company pairs by considering different basket sizes. Hence, such measurements are insensitive to how the whole data set is partitioned, and the resulting estimation indicates the performance of an ensemble of 18 classifiers (one for each basket), all using a given classification method. The estimated overall prior probability for positives is 11.8% (approximately 1 in 9 company pairs in the original data set is a competitor pair). In contrast with this low estimated prior, Table 4 shows that our competitor discovery approach can achieve reasonably good estimated classification performance. ANN achieves the best performance on more metrics than the other three methods, but unlike the three methods (ANN, DT, and BN), LR does not require any parameter turning and produces comparably good results. We highlight the best performance value for each measurement in Table 4.

6. Competitor extension

In Introduction, we use an anecdote to illustrate that the gold standards can be incomplete. In this section we suggest metrics to estimate (1) the coverage of competitor pairs by a gold standard and (2) the extent to which our approach extends each gold standard.

6.1. Estimating the coverage of a gold standard

We require the following notation from Fig. 4 to describe the estimation procedure:

C: (unknown) complete set of competitor pairs,
H: set of competitor pairs covered by Hoover's,
M: set of competitor pairs covered by Mergent, and
$J_{HM} = H \cap M$, intersection of H and M.

Following the ideas discussed in Le Cren (1965) to estimate wildlife populations and in Lawrence and Giles (1998) to estimate the coverage of search engines, we assume H and M are independent subsets of C and thus estimate the extent to which H covers C, according to how much of H covers M (i.e., J_{HM}) and the size of M. We therefore estimate the coverage of the entire competitor set C by Hoover's (Cov(H)) and Mergent (Cov(M)) as follows:

Table 4
Estimated overall performances for the whole data set.

	Without sector information					With sector information				
	Precision	Recall	FP rate	F_1	Accuracy	Precision	Recall	FP rate	F_1	Accuracy
ANN	0.419	0.378	0.046	0.397	0.907	0.450	0.513	0.055	0.479	0.910
BN	0.238	0.354	0.095	0.284	0.863	0.388	0.514	0.071	0.442	0.895
DT	0.341	0.374	0.064	0.357	0.891	0.432	0.457	0.053	0.444	0.907
LR	0.388	0.330	0.046	0.357	0.904	0.382	0.437	0.062	0.407	0.897

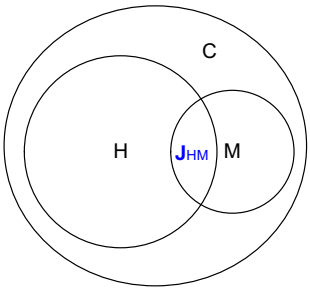

Fig. 4. Competitor covered by two gold standards.

$$\mathrm{Cov}(H) = \frac{|J_{HM}|}{|M|}, \tag{8}$$

$$\mathrm{Cov}(M) = \frac{|J_{HM}|}{|H|}. \tag{9}$$

If H and M are not completely independent, the value of J_{HM} (their intersection) is expected to be larger than when they are independent. Then our coverage estimation provides an upper bound on true coverage.We previously labeled the positive instances according to Hoover's and Mergent for each sample basket, which enables us to compute the number of competitor pairs identified by Hoover's (H_i) and Mergent (M_i) separately, as well as the intersection of Hoover's and Mergent (J_{HiMi}) for the ith sample basket. Similar to our approach used in Eq. (7), we estimate the number of positives (for Hoover's, Mergent, and their intersection) in each original basket by multiplying the number of positives in the sample basket by the ratio of the basket size to the sample basket size. Then, using the following Eqs. (10) and (11), we calculate the coverage of Hoover's and Mergent as follows:

$$Cov(H) = \frac{|J_{HM}|}{|M|} = \frac{|\sum_{i=1}^{18} J_{HiMi} \times \frac{B_i}{S_i}|}{|\sum_{i=1}^{18} M_i \times \frac{B_i}{S_i}|}. \tag{10}$$

$$Cov(M) = \frac{|J_{HM}|}{|H|} = \frac{|\sum_{i=1}^{18} J_{HiMi} \times \frac{B_i}{S_i}|}{|\sum_{i=1}^{18} H_i \times \frac{B_i}{S_i}|}. \tag{11}$$

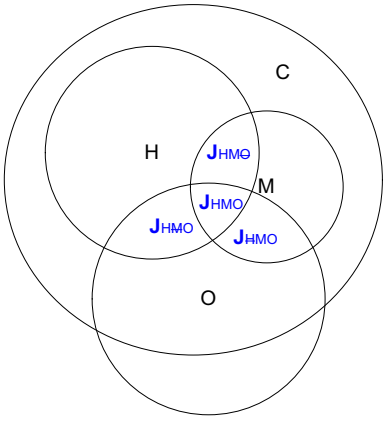

Fig. 5. Competitor covered by two gold standards and our approach.

We find that the estimated coverage of Hoover's and Mergent is 46.0% and 24.9%, respectively. So we estimate that both company profile resources individually cover less than 50% (note that 46% and 24.9% are the upper bounds as explained earlier) of all competitor pairs. The results quantify and confirm our initial anecdote about incompleteness of these industry-strength data sources.

6.1.1. Estimating the extension of the proposed approach to a gold standard

We now present a procedure to estimate how much our automated approach might extend a gold standard (i.e., identify competitor pairs that are not covered by the gold standard). Our estimation procedure uses the following notation in Fig. 5:

O: the set of competitor pairs classified by our approach,

$\bar{H} = C - H$,
$\bar{M} = C - M$,
$\bar{O} = C - O \cap C$,
$J_{HMO} = H \cap M \cap O$,
$J_{\bar{H}MO} = \bar{H} \cap M \cap O$,
$J_{H\bar{M}O} = H \cap \bar{M} \cap O$, and
$J_{HM\bar{M}} = H \cap M \cap \bar{M}$.

Thus, J_{HMO} is a subset of competitor pairs that our approach classifies as positive and that Mergent confirms as positive but that Hoover's does not identify as competitors. Given that competitor pairs in Mergent are a subset of all competitor pairs, we estimate the extent to which our approach extends Hoover's (Ext(O, H)) as follows:

$$\mathrm{Ext}(O,H) = \frac{|J_{\bar{H}MO}|}{|M|}. \tag{12}$$

Similarly, we estimate the extent to which our approach extends Mergent (Ext(O, M)) as follows:

$$\mathrm{Ext}(O,M) = \frac{|J_{\bar{H}MO}|}{|H|}. \tag{13}$$

As researchers, we do not judge/label whether two given companies are competitors or not. Instead, we resort to real-world commercial company profile resources (i.e., Hoover's and Mergent), trust their identified competitors, and thus call them gold standards. The extension of our approach to one gold standard depends on how much the other confirms our output as we are not in the position to judge competitors. Since the entire universe (C) of competitor pairs is unknown, we can only estimate the extensions through samples of the universe and we consider H and M samples of C. Moreover, it may be possible to combine the two data sources (Hoover's + Mergent) and use a third data source for estimating extension by our approach over Hoover's + Mergent. The methodology for estimating the extensions would remain the same. One could incrementally apply the methodology to a combination of even larger number of data sources if they are available. On the basis of Eqs. (12) and (13), we compute the extension of our approach to each gold standard using results from combined data set 2680 with the following equations:

Table 5
Extensions to a gold standard.

	Without sector information (%)				With sector information (%)			
	ANN	BN	DT	LR	ANN	BN	DT	LR
Ext(O, H)	5.9	7.3	9.80	5.0	12.1	11.3	10.1	10.5
Ext(O, M)	28.7	23.4	30.50	24.3	33.8	37.1	35.8	32.9

$$\mathrm{Ext}(O,H) = \frac{|J_{\bar{H}MO}|}{|M|} = \frac{|\sum_{i=1}^{18} J_{\bar{H}MO} \times \frac{B_i}{S_i}|}{|\sum_{i=1}^{18} M_i \times \frac{B_i}{S_i}|}. \quad (14)$$

$$\mathrm{Ext}(O,M) = \frac{|J_{\bar{H}MO}|}{|H|} = \frac{|\sum_{i=1}^{18} J_{\bar{H}MO} \times \frac{B_i}{S_i}|}{|\sum_{i=1}^{18} H_i \times \frac{B_i}{S_i}|}. \quad (15)$$

Table 5 shows the estimation of how much our approach extends the knowledge available from each of the gold standards, for the different classification methods (with and without sector information). Using the sector information and any classification method, our approach extends Hoover's and Mergent by more than 10% and 32%, respectively. These extension values are based on classification results generated from a set of input parameters and classification methods. The results in Table 5 are associated with estimated overall performance in Table 4. For example, for ANN the extensions of our approach to Hoover's (12.1%) and Mergent (33.8%) are associated with precision, recall, and FP rate of 0.450, 0.513, and 0.055, respectively.

7. Conclusions

We propose and evaluate an approach that exploits company citations in online news to create an intercompany network whose structural attributes are used to infer competitor relationships between companies. As noted earlier the company citations in news may not necessarily represent competitor relationships. However, we find that such a citation-based network carries latent information and the structural properties can be used to infer competitor relationships. Our evaluations prompt three broad observations. First, the intercompany network captures signals about competitor relationships. Second, the structural attributes, when combined in various types of classification models, infer competitor relationships. For imbalanced portions of the data, we require more advanced modeling techniques (e.g., data segmentation, DTA) to achieve reasonable performance. Third, we quantify the degree to which two commercial data sources are incomplete in their coverage of competitors and estimate the extent to which our approach extends them while still maintaining adequate performance. Our approach, especially as an initial filtering step before further manual examinations, can be used by an individual company to find its emerging competitors and competitors of its clients or suppliers. The suggested approach can be used by a financial analyst to identify a large group of potential competitors in a sector. A company profile resource such as Hoover's and Mergent can also use this approach to identify what it could be missing and to greatly reduce its manual efforts.

It would be interesting to examine our approach with news stories written in another language. Also it would be worthwhile to explore whether the intercompany network can predict future competitor relationships. Finally, our approach may be extended to discover other business relationships.

References

Bao, S., Li, R., Yu, Y., and Cao, Y. Competitor mining with the Web. *IEEE Transactions on Knowledge and Data Engineering*, 20, 10, 2008, 1297–1310.

Becchetti, L., Castillo, C., Donato, E., Baeza-Yates, R., and Leonardi, S. Link analysis for Web spam detection. *ACM Transactions on the Web*, 2, 1, 2008 (Article 2).

Bernstein, A., Clearwater, S., Hill, S., and Provost, F. Discovering knowledge from relational data extracted from business news. In *Proceedings of the KDD 2002 Workshop on Multi-Relational Data Mining*, Edmonton, Alberta, Canada, 2002, 7–20.

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

Bhagat, S., Rozenbaum, I., and Cormode, G. Applying link-based classification to label blogs. In *Proceedings of Joint 9th WEBKDD and 1st SNA-KDD Workshop*, San Jose, CA, 2007.

Brandes, U. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25, 2, 2001, 163–177.

Brin, S., and Page, L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30, 1–7, 1998, 107–117.

Chakrabarti, S., Dom, B., and Indyk P. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, Seattle, WA, 1998, 307–318.

Chan, P., and Stolfo, S. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In Agrawal, Stolorz, and Piatetsky-Shapiro (eds.), *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, CA, 1998, 164–168.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 2002, 321–357.

Cooper, G., and Herskovitz, E. A. Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, 4, 1992, 309–347.

Freeman, L. C. Centrality in social networks: Conceptual clarification. *Social Networks*, 1, 1979, 215–239.

Friedman, N., Geiger, D., and Goldszmidt, M. Bayesian network classifiers. *Machine Learning*, 29, 2–3, 1997, 131–163.

Getoor, L., and Diehl, C. P. Link mining: A survey. *ACM SIGKDD Explorations Newsletter*, 7, 2, 2005, 3–12.

Granovetter, M. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91, 3, 1985, 481–510.

Gulati, R., and Gargiulo, M. Where do interorganizational networks come from? *American Journal of Sociology*, 104, 5, 1999, 1439–1493.

Hogg, T. Inferring preference correlations from social networks. *Electronic Commerce Research and Applications*, 9, 2010, 29–37.

Karamon, J., Matsuo, Y., and Ishizuka, M. Generating useful network-based features for analyzing social networks. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 2008, 1162–1168.

Kessler, M. M. Bibliographic coupling between scientific papers. *American Documentation*, 24, 1963, 123–131.

Kleinberg, J. Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46, 5, 1999, 604–632.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30, 1, 2006.

Lawrence, S., and Giles, C. L. Searching the World Wide Web. *Science*, 280, 3, 1998, 98–100.

Le Cren, E. D. A note on the history of mark-recapture population estimates. *Journal of Animal Ecology*, 34, 2, 1965, 453–454.

Liben-Nowell, D., and Kleinberg, J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58, 7, 2007, 1019–1031.

Lorrain, F., and White, H. G. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1, 1971, 49–80.

Ma, Z., Pant, G., and Sheng, O.R.L. A network-based approach to mining competitor relationships from online news. In *Proceedings of the 30th International Conference on Information Systems*, Phoenix, USA, December 2009a.

Ma, Z., Sheng, O. R. L., and Pant, G. Discovering company revenue relations from news: A network approach. *Decision Support Systems*, 47, 4, 2009, 408–414.

Macskassy, S., and Provost, F. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning*, 8, 2007, 935–983.

O'Madadhain, J., Fisher, D., White, S. and Boey, Y.B. JUNG: The Java Universal Network/Graph Framework (ver. 1.7.4), 2006. <http://jung.sourceforge.net>.

Pazzani, M., Merz, C., and Murphy, P. reducing misclassification costs. In *Proceedings of the 11th International Conference on Machine Learning*, New Brunswick, NJ, 1994, 217–225.

Provost, F. Machine Learning from Imbalanced Data Sets 101. Invited paper, in Workshop on Learning from Imbalanced Data Sets, AAAI. Texas, USA, 2000.

Salton, G. *The SMART Retrieval System: Experiments in Automatic Document Proceeding*. Prentice Hall, Englewood Cliffs, NJ, 1971.
Salzberg, S. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1, 1997, 317–327.
Small, H. Co-citation in the scientific literature: A New measurement of the relationship between two documents. *Journal of the American Society of Information Science*, 24, 4, 1973, 265–269.
Uzzi, B. Embeddedness in the making of financial capital: How social relations and networks benefit firms seeking financing. *American Sociological Review*, 64, 1999, 481–505.
Walker, G., Kogut, B., and Shan, W. Social capital, structural holes and the formation of an industry network. *Organization Science*, 8, 2, 1997, 109–125.
Wang, J. C., and Chiu, C. C. Recommending trusted online auction sellers using social network analysis. *Expert Systems with Applications*, 34, 3, 2008, 1666–1679.
Wasserman, S., and Faust, K. *Social Network Analysis: Methods and Applications*, 1st edition. Cambridge University Press, 1994.
Weiss, G. M. Mining with rarity: A unifying framework. *SIGKDD Explorations*, 6, 1, 2004, 7–19.
Weiss, G. M., and Provost, F. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19, 2003, 315–354.
Witten, I. H., and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. Morgan Kaufmann, San Francisco, CA, 2005.