Semi-Supervised Text Mining For Dynamic Business Network Discovery

Wenping Zhang  
*Department of Information Systems, City University of Hong Kong, Hong Kong SAR, wzhang23@student.cityu.edu.hk*

Yi Cai  
*School of Software Engineering, South China University of Technology, China, ycai@scut.edu.cn*

Raymond Y.K. Lau  
*Department of Information Systems, City University of Hong Kong, Hong Kong SAR, raylau@cityu.edu.hk*

Stephen S. Liao  
*Department of Information Systems, City University of Hong Kong, Hong Kong SAR, issliao@cityu.edu.hk*

Ron C.W. Kwok  
*Department of Information Systems, City University of Hong Kong, Hong Kong SAR, isron@cityu.edu.hk*

Follow this and additional works at: [http://aisel.aisnet.org/pacis2012](http://aisel.aisnet.org/pacis2012)

**Recommended Citation**
[http://aisel.aisnet.org/pacis2012/138](http://aisel.aisnet.org/pacis2012/138)

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2012 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
SEMI-SUPERVISED TEXT MINING FOR
DYNAMIC BUSINESS NETWORK DISCOVERY

Wenping Zhang, Department of Information Systems, City University of Hong Kong, Hong Kong SAR, wzhang23@student.cityu.edu.hk

Yi Cai, School of Software Engineering, South China University of Technology, China, ycai@scut.edu.cn

Raymond Y.K. Lau, Department of Information Systems, City University of Hong Kong, Hong Kong SAR, raylau@cityu.edu.hk

Stephen S. Liao, Department of Information Systems, City University of Hong Kong, Hong Kong SAR, issliao@cityu.edu.hk

Ron C.W. Kwok, Department of Information Systems, City University of Hong Kong, Hong Kong SAR, isron@cityu.edu.hk

Abstract

Recently, much research effort has been devoted to the discovery and analysis of online social networks. However, relatively little research has been done for business network discovery and analysis. Although named entity recognition (NER) tools are available to identify basic entities in texts, there are still challenging research problems, such as co-reference resolution and the identification of abbreviations of organization names. Guided by the design science methodology, the main contribution of this paper is the design and development of a novel semi-supervised method for the identification of business entities (e.g., companies) and their relationships. Based on the automatically mined business networks, financial analysts can then predict the business prestige of companies for better financial investment decision making. Initial experiments show that the proposed business entity identification method is more effective than other baseline methods. Moreover, the proposed semi-supervised business relationship mining method is more effective than the state-of-the-art supervised machine learning classifier when a large number of manually labeled training examples are not available. The managerial implication is that business managers can apply the design artifacts to promptly identify potential business partners and competitors, and hence improve their strategic business decision making processes.

Keywords: Named Entity Recognition, Business Network Mining, Text Mining, Statistical Learning.
1 INTRODUCTION

In the era of Web 2.0, there has been an explosive growth of the volume of user-contributed contents on the Internet. This ever-increasing user-generated data has brought to organizations with unprecedented opportunities to exploit business intelligence (BI) and develop deep insights about their customers, business partners, and competitors (Chen 2010; Chen and Zimbra 2010). Nevertheless, the sheer volume of user-contributed unstructured data also poses new challenges to organizations. While manual analysis of the sheer volume of user-contributed data is not practical due to the problem of information overload (Lau et al. 2008; Yan et al. 2011), existing computational methods may not be able to effectively process the flood of unstructured data coming from the Internet. Accordingly, there is a pressing need to develop the next generation of computational methods and tools to convert the sheer volume of user-contributed unstructured data into actionable knowledge which can be easily leveraged by organizations to make timely and effective business decisions. For instance, recent research work has been devoted to the construction and analysis of social networks based on user-contributed data, and hence to improve marketing effectiveness and reduce marketing costs (Trusov et. al. 2010).

Although researchers have been active to examine the discovery and analysis of online social networks (Chakrabarti and Faloutsos 2006; Wilson and Banzhaf 2009; Yang and Ng 2008), relatively little work has been conducted for the discovery and analysis of business networks based on unstructured online data. Business networks can be seen as a special kind of social network which captures the linkage information among organizations rather than individual users. However, mining and constructing business networks poses new challenges to existing social network analysis (SNA) techniques. Unlike online social networks, explicit information about the connections among individual companies is generally not available on the Internet. Moreover, a business network of organizations is more volatile than a social network of individual users since there is not a forever friend (or an enemy) in the business world. In fact, a business partner of the current period may quickly turn into a competitor in the next period. Accordingly, it is far more difficult to construct and analyze a business network than conducting the same thing for a social network of individuals. Guided by the design science research methodology (Hevner et al. 2004; March and Storey 2008; Peffers et al. 2008), one of the main contributions of this paper is the design and development of a novel computational method (i.e., a designed artifact) for the automatic discovery and analysis of business networks based on the latent relationships embedded in online unstructured texts, such as financial news articles or investor comments. More specifically, a semi-supervised statistical learning method is proposed to reduce the burden of manually annotating a large number of training examples to train a supervised machine learning classifier. We believe that a semi-supervised computational method can improve the chance of the proposed design artefacts (i.e., computational method and the prototype system) to be adopted by the real-world business users.

In general, commercial firms always need to transact with other parties to produce goods or services, and hence generate profits for the continuous of the firms. During this production-survival process, firms have to interact with many parties (e.g., organizations). For instance, a commercial company needs to collaborate with suppliers to acquire the raw materials to manufacture products. In addition, it needs to collaborate with retailers to distribute and sell the finished products or services. When a company is in lack of financial capital, it needs to raise capital from other financial institutions (e.g., investor firms) to obtain the necessary financial resource to meet its operational expenses. Moreover, companies inevitably compete with other firms (i.e., the competitors) as long as a market is profitable. In order to gain bigger market share, sometimes companies may establish strategic alliance with other business partners to better flight with other competitors in a certain market. The aforementioned typical business activities reveal the natural formation of business networks. Typical business relationships include collaborative relationship, competitive relationship, supplier-consumer relationship, financial investment relationship, etc. This paper focuses on the automated mining of
collaborative and competitive business relationships based on unstructured online text collected from the Internet.

For existing business practice, most commercial firms allocate large amounts of human resources to collect and analyze market and competitor information (Ghoshal and Westney 1991). However, given the sheer volume of business data and the constantly changing business environment (e.g., an existing business partner could become a competitor the following day), it is not practical to solely rely on manual method to collect and analyze the huge volume of business network data. The managerial implication of our research work is that the proposed computational method demonstrates a feasible solution to address the important business problem of enabling commercial firms to effectively and efficiently identifying business networks and analyzing the prominent behaviour of such networks to gain competitive advantages.

2 RELATED RESEARCH

Linguistic rules or heuristic are often applied to identify specific types of entities such as people, organizations, places, etc. (Budi and Bressan 2007; Zhou and Su 2004). However, the main weakness of this approach is that the pre-defined rules may not be able to cover a wide variety of situations (i.e., low recall). On the other hand, supervised machine learning approach normally utilizes training corpora to automatically build classifiers to identify entities in unseen documents (Patrick et al. 2002; Vincent 2007). Nevertheless, the main weakness of supervised machine learning approach is the requirement of manually labeling a large number of training examples for classifier training. In this paper, we propose a semi-supervised statistical learning method to address the low recall problem in NER while avoiding the time-consuming process of manually annotating a large number of training examples.

Bernstein et al. (2003) proposed to scan the contents of online financial news articles to estimate the co-occurrence statistics of a pair of stock tickers. The co-occurrence statistics were then used to predict the possible relations among companies. In addition, the membership of a company falling in a particular industry sector could be also predicted. The CoMiner system made use of Natural Language Processing (NLP) techniques and several pre-defined lexical-syntactic patterns (e.g., company-A “against” company-B) to identify competitive company relations based on a Web corpus (Bao et al. 2008). Each lexical-syntactic pattern was assigned a weight and the point-wise mutual information (PMI) measure was used to estimate the strength of competition relation between two companies based on a Web corpus. Our proposed approach for business network mining can discover two different kinds of business relationships, e.g., competition relation and cooperation relation.

A network-based approach was developed to extract business competition relations from online financial news articles (Ma et al. 2009b). A weighted directed graph approach was adopted. For each financial news article about a company, the system tried to find other companies (identified by stock tickers) that also appeared in the same news article. If such a pair of companies (x, y) was found, it was assumed that there would be a directed link from x to y. Four graph-based features such as weight of in-degree, weight of out-degree, weight of total-degree, etc. were used by four supervised classifiers such as artificial neural network (ANN), Bayesian Network (BN), C 4.5 Decision Tree (DT), and Logistic Regression (LR). A similar network-based approach was also applied to predict company revenue relations based on online financial news articles, and it was found that both the decision tree based classifier and the logistic regression based classifier achieved comparable performance in various financial news corpora (Ma et al. 2009a).

A hybrid content-based and link-based approach was proposed to predict whether two companies were competitor and the specific direction of competition (Pant and Sheng 2009). In particular, five measures such as in-link similarity, out-link similarity, text similarity (similarity of respective home pages), news count (co-occurrence of the companies in financial news), and search engine count (co-occurrence of the companies in Web pages) were used to classify whether a pair of companies was
competitors or not using supervised machine learning techniques such as C4.5 decision tree or logistic regression. For our proposed computational method, we use a weighted undirected graph to represent a business network. Moreover, we use a semi-supervised statistical learning approach to mine business relationships from unstructured text because a large number of manually labelled training examples are too costly or even impossible to be constructed for the business domain.

CopeOpi system was proposed to extract opinions and implicit business relationships of some targeting entities based on the Web corpus (Ku et al. 2009; Ku et al. 2006). Business associations among the target entities are discovered based on the opinion-tracking plots (i.e., the pattern of opinion scores exhibited over time). However, their method assumed that two companies exhibiting similar opinion patterns were related (Ku et al. 2009; Ku et al. 2006). Our proposed business relationship mining method supports a more fine-grained relationship mining process conducted at the sentence level rather than at the document level. In addition, we do not make the unrealistic assumption that two companies showing similar opinion patterns over time are related. The CoNet system employed shallow natural language processing (NLP) techniques to identify commercial entities and relationships from online financial news (Xia et al. 2010). Basic linguistic rules were applied to identify the abbreviations of company names and resolve the co-references of company names.

Latent semantic indexing (LSI) was examined to identify relationships among entities of interests (Bradford 2006). In particular, each entity was represented by an entity vector and mapped to the LSI space. For any two entities appearing in a document collection, the proximity (computed in terms of cosine similarity) of the corresponding entity vectors in the LSI space provides a quantitative measure of the degree of contextual association between the entities. Instead of using a computationally expensive LSI method, the computational method proposed in this paper employs a more computationally friendly statistical learning approach for the discovery of hidden business relationships.

3 THE COMPUTATIONAL METHODS

3.1 Business Entity Identification

The first step toward business relationship mining is to identify the company names in a financial news article or investor comment. There are two main challenges of business name tagging. First, various abbreviations of a company name (e.g., “General Electronic” as “GE”) should be identified. Second, the co-references (e.g., the company, it, etc.) related to a business name should be resolved. By using “General Electronic” as an example, Figure 1 shows the variations of the references related to “General Electronic”. For instance, it is referred to as GE (i.e., an abbreviation), the company, or simply the pronoun “it”. An effective business entity identification mechanism should be able to resolve all these references, and link them to the same business name, that is, “General Electronic”.

![Figure 1. The Challenges of Business Entity Identification.](image-url)
The proposed procedures for business entity identification can be summarized as the following main steps:

- Extract business full names and stock ticker labels from Web sources (e.g., Yahoo! Finance);
- Apply a general NER tool to identify general organization names;
- Apply a mutual information based statistical learning method to extract abbreviations frequently co-occurring with the business full names or stock tickers in a training corpus;
- Apply the automatically learned business abbreviations to resolve short business names in a document;
- Apply the proximity-based co-reference resolution algorithm to convert the co-references to business full names;

The business full names and the stock tickers of companies extracted from Yahoo! Finance are passed to an existing NER system to expand its organization name dictionary for preliminary business name identification. In particular, we employ the named entity recognition module of GATE (Maynard et al. 2001) for basic business name identification. GATE is a general information engineering service which also contains a text tokenizer, a sentence splitter, a part-of-speech tagger, a morphological analyzer, and a VP chunker (Cunningham 2002). The NER rules of GATE are extended to consider the specific requirements of business name identification. For instance, the tokens followed by “Inc”, “Co”, “Ltd”, etc. are likely to be business names.

To automatically extract the abbreviations of business entities, a point-wise mutual information (PMI) based statistical learning method is proposed. The basic intuition of the proposed semi-supervised learning method is that a business name and its abbreviations tend to appear in adjacent sentences (as shown in Figure 1), and this kind of co-occurrence may appear frequently in a corpus. More specifically, we calibrate the Balance Mutual Information (BMI) measure (Lau et al. 2009), a variant of the point-wise mutual information measure (Shannon 1948), to conduct statistical learning for the abbreviations of business entities. The BMI measure has been successfully applied to context-sensitive text mining for information retrieval (IR) (Lau 2003) and automatic domain concept extraction in ontology discovery (Lau et al. 2009). The distinct advantage of the BMI measure is that it can take into account both positive and negative evidence presented in text to infer the strength of association between two tokens. The BMI measure is calibrated to develop the proposed abbreviation extractor (AE) function as follows:

\[
AE(t_i, t_j) = \alpha \times \left( \frac{Pr(t_i, t_j) \log_2 \left( \frac{Pr(t_i, t_j)}{Pr(t_i) Pr(t_j) + \gamma} \right) + Pr(\neg t_i, \neg t_j) \log_2 \left( \frac{Pr(\neg t_i, \neg t_j)}{Pr(\neg t_i) Pr(\neg t_j) + \gamma} \right)}{Pr(t_i) Pr(t_j) + \gamma} \right) - (1 - \alpha) \times \left( \frac{Pr(t_i, \neg t_j) \log_2 \left( \frac{Pr(t_i, \neg t_j)}{Pr(t_i) Pr(\neg t_j) + \gamma} \right) + Pr(\neg t_i, t_j) \log_2 \left( \frac{Pr(\neg t_i, t_j)}{Pr(\neg t_i) Pr(t_j) + \gamma} \right)}{Pr(\neg t_i) Pr(t_j) + \gamma} \right)
\]

where \( AE(t_i, t_j) \) is the abbreviation extractor function applied to two terms \( t_i \) and \( t_j \). \( Pr(t_i, t_j) \) is the joint probability that both terms appear in a virtual text window, and \( Pr(t_i) \) is the probability that a term \( t_i \) appears in a text window. The probability \( Pr(t_i) \) is estimated according to \( \frac{w_i}{w} \), where \( w_i \) is the number of virtual text windows containing the term \( t \) and \( |w| \) the total number of virtual text windows constructed from a text corpus. Similarly, \( Pr(t_i, \neg t_j) \) is the fraction of the number of virtual text windows containing both terms out of the total number of virtual text windows constructed. The interpretation of \( Pr(t_i, \neg t_j) \) is that the probability of \( t_i \) present and \( t_j \) absent in a virtual text window.
So, if two terms often co-occur in a text corpus, it contributes to raise the AE score due to the statistical association. In contrast, if two terms are often not to be observed together in a text corpus, it will reduce the AE score. For the business abbreviation finding application, one of the terms is the business full name which is collected from public financial sites.

The parameter $\alpha$ is used to adjust the relative weight of positive and negative evidence for estimating the statistical association between two terms. We empirically established this parameter in the current study. First, a set of 10 companies was selected as the targets. Second, Eq.(1) instantiated with various values of $\alpha$ was applied to a financial corpus of 10K documents related to the chosen companies. Third, the abbreviations with top AE scores were manually verified. Finally, the parameter value of $\alpha = 0.66$ which led to the best abbreviation finding results in this empirical study was adopted for the remaining experiments reported in this paper. The parameter $\gamma$ is the Laplace smoothing parameter.

![Algorithm](image)

Figure 2. The Proximity-based Co-reference Resolution Algorithm.

During text mining, a virtual text window is formed by first locating the targeted business full name (or stock ticker number) and then extracting the $n$ terms to the left and right of the business full name according to the window size ($n$) specified as a system parameter. This text mining process is repeated for the entire text corpus (e.g., a collection of financial news articles). According to previous studies in
information retrieval and ontology discovery, a virtual text window of 5 to 10 tokens is effective (Lau 2003; Lau et al. 2009). This virtual text window introduces a constraint (i.e., a proximity factor) to the statistical learning process of business abbreviations finding such that only the tokens adjacent to the business full names or stock tickers are considered. Such a constraint aims at reducing the noise of the text mining process. However, for business abbreviation extraction, the text window should be much larger since a business full name and its abbreviation may appear in adjacent sentences rather than occurring within the same sentence. Therefore, we set the window size of two paragraphs for this particular NER application. After computing the abbreviation extraction scores of the tokens with specific part-of-speech (POS) (e.g., no valid POS found) and specific properties (e.g., a token with title case), the top \( n \) tokens with the highest abbreviation extraction scores are selected as the potential abbreviations of the business full names.

For processing the co-references of a business entity, the proximity-based co-reference resolution algorithm depicted in Figure 2 is applied. Within each document \( d_{in} \), all the business entities are first identified according to a pre-composed company table (containing full names, stock tickers, and abbreviations). The proposed proximity-based co-reference resolution algorithm then locates the first ambiguous token \( t_1 \in d_{in} \) (e.g., ``company'', ``it'', etc.) by invoking the ReadToken function. Then, the algorithm works backward to search for the nearest business entity already identified by the previous NER procedure. For our implementation, we used GATE (Maynard et al. 2001) to conduct the basic NER process. The proximity for co-reference resolution is globally defined in advance; for our experiment reported in this paper, the proximity of two paragraphs was defined. If a nearest business entity is found and it is located within the boundary of two paragraphs, the pronoun or ambiguous token is replaced by the full name of the nearest business entity. The algorithm continues to process the next ambiguous pronoun in the document until no more ambiguous pronoun is found or the end of the document is encountered. Finally, the algorithm returns \( d_{out} \) that is the document with co-references resolved.

### 3.2 Business Relationship Mining

The aim of dynamic business relationship mining is to identify and extract the associations among business entities from a text corpus (e.g., a collection of financial news articles) of a particular period. The extracted business relationships can be formally represented by a binary tuple \( BN = (G, tp) \) whereas \( G = (V, E) \) is a graph. The property of such a graph can be further analyzed to identify the unique properties of a business network (e.g., the prestige of companies of an industrial sector).

**Definition 1** (Business Network): A business network \( BN \) is a binary tuple \( BN = (G, tp) \) whereas \( G = (V, E) \) is a weighted undirected graph and \( tp \) is the time point when the graph \( G \) is valid. The set \( V \) represents the set of nodes (companies) and the set \( E \) denotes the set of edges (business relationships) of \( G \). Each edge \( e_{xy} \in E \) is an unordered pair of nodes \( (x, y) \) which indicates a connection between nodes \( x \in V \) and \( y \in V \). The weight of an edge (strength of business connection) is denoted \( w(e_{xy}) \). The type of edge (business relationship type) is denoted \( t(e_{xy}) \).

One unique feature of the proposed business relationship mining algorithm is that it is based on a semi-supervised statistical learning technique and the notion of context vectors (Jing and Tzoukermann 1999; Schütze 1999). A context vector \( v_x \) for the business entity \( x \) is formally defined by \( v_x = (t_1, w_1), (t_2, w_2), \ldots, (t_n, w_n) \) where \( (t_i, w_i) \) represents the descriptive term and its weight for the entity \( x \). A context vector is formed by extracting the adjacent terms of a targeted entity throughout a text corpus (Jing and Tzoukermann 1999; Schütze 1999). From an implementation perspective, virtual text windows of a pre-defined size surrounding some entities are formed in a text
corpus (Jing and Tzoukermann 1999; Schütze 1999). The statistically associated terms of these entities are then identified to build the corresponding context vectors (Lau et al. 2009).

For our application, context vectors provide an expressive and computationally friendly description of the business entities. In particular, we employ a statistical learning method, namely the Google Similarity Distance (GSD), which has been successfully applied for estimating the semantic similarity of concepts in large corpora (Cilibrasi and Vitanyi 2007), to build context vectors for business relationship mining. Essentially, a context vector is used to represent the set of terms and their weights measured in terms of normalized GSD; this set of weighted terms provides a context-aware description about a business entity. For our particular constructions of context vectors, the name of a business entity is always included as a constituent term of its context vector, and the maximal weight of 1 is assigned. To identify the latent relationships among a set of business entities in this entity-term space, we adopt the cosine similarity measure (Salton and McGill 1983) to estimate the strength of latent business relationships. The rational of using the cosine similarity measure is that a high cosine score will be derived if two business entities have strong association (e.g., the names of the related business entities are often included as the constituent terms of the context vector). For instance, if “IBM” and “HP” are two associated companies, their context vectors are likely to include the names of the other side with a relative high term weight.

<table>
<thead>
<tr>
<th>Company</th>
<th>Context Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>ibm(1.000) corp(0.948) software(0.937) inc(0.936) technology(0.934) service(0.928) sun(0.922) hp(0.921) dell(0.917) apple(0.917) data(0.917) oracle(0.917) earnings(0.916) hewllettpackard(0.915) microsoft(0.915) growth(0.914) computer(0.911) business(0.901) storage(0.901) firm(0.901) international(0.901) machine(0.900) …………….</td>
</tr>
<tr>
<td>HP</td>
<td>hp(1.000) hewllettpackard(1.000) printer(0.966) software(0.966) corp(0.958) company(0.932) technology(0.924) ibm(0.918) business(0.915) international(0.915) ag(0.915) service(0.914) sales(0.913) results(0.913) nasdaq(0.911) sun(0.911) business(0.902) server(0.900) …………….</td>
</tr>
</tbody>
</table>

Table 1. The context vectors of IBM and HP

Table 1 shows the context vectors of IBM and HP, respectively. The common terms of the two context vectors are highlighted. These context vectors are mined based on our financial news corpus (200K news articles) crawled from Reuters.com. Only the top weighted terms of a context vector are shown in Table 1. As a whole, the main advantage of the proposed statistical learning method is that manually labeled training examples are not required for business relationship mining, and hence it improves the chance of the proposed business relationship mining algorithm to be adopted by real-world business applications.

Since the proposed algorithm strives to identify two different types of business relationships, namely, cooperative and competitive relationships, the final step of business relationship mining is to label each business relationship according to a system-generated relationship lexicon. Similar to the ideas proposed by Turney and Littman (Turney and Littman 2003) for automatic opinion lexicon expansion, we apply the normalized GSD to conduct statistical inference so as to extract a set of domain-specific relationship indicators according to some seeding relationship indicators. More specifically, 10 collaboration indicators (e.g., cooperate, ally, collaborate, joint, own, partner, coordinate, engage, partnership, and agree) and another 10 competition indicators (e.g., compete, challenge, against, vie, contend, fight, contest, dispute, battle, and accuse) were used as the seeding relationship indicators. The synonyms of these indicators are also extracted from WordNet (Miller et al. 1990) automatically. The initial set of relationship indicators is used by the statistical learning module which is underpinned by normalized GSD to generate a comprehensive relationship lexicon. Finally, each textual context (i.e., sentences) containing two business entities identified via cosine similarity at the early stage will be used to disambiguate the type of business relationships (e.g., cooperative vs. competitive) of the related business entities.
When compared to the method proposed by Turney and Littman (2003), the improvement of our statistical inference method includes the application of WordNet to automatically build the initial relationship lexicon and filter the noisy relationship indicators generated by the statistical inference process. For instance, a potential relationship indicator statistically inferred for the “collaborative” business relationship type will be checked for contradictory semantic meaning in WordNet. If contradictory meaning (e.g., competitive) is not found in WordNet, then the statistically inferred business relationship indicator will be added to the relationship lexicon; otherwise it will be discarded. A similar process is applied to automatically filter potential relationship indicators for the “competitive” business relationship type.

The normalized GSD is the main computational apparatus of the proposed business relationship mining algorithm. Since the original GSD measure is applied to estimate semantic distance (Cilibrasi and Vitanyi 2007), we calibrate the GSD measure to develop the normalized GSD which is used to estimate semantic relatedness rather than distance. The normalized GSD (NGSD) is defined below:

\[
NGSD(t_i, t_j) = 1 - \left( \frac{GSD(t_i, t_j) - GSD_{\text{min}}}{GSD_{\text{max}} - GSD_{\text{min}}} \right)
\]  

where \( |w_j| \) represents the number of virtual text windows containing the term \( t_j \) in the text corpus. \(|w_{i,j}|\) refers to the number of virtual text windows containing both terms \( t_i \) and \( t_j \). The term \(|w + 1|\) represents the total number of virtual text windows extracted from the text corpus plus 1. For our application, one of the terms is a business entity name and the size of the virtual text windows is 10 because a window size between 5 to 10 seems effective for text mining applications (Lau 2003; Lau et al. 2009). \( GSD_{\text{max}} \) (\( GSD_{\text{min}} \)) is the global maximal (minimal) of the GSD scores. It should be noted that the proposed NGSD is applied to generate context vectors of business entities and expand the relationship lexicon for business relationship type disambiguation, respectively.

The cosine similarity score between two context vectors (representing two business entities) is defined by (Salton and McGill 1983):

\[
w(e_{xy}) \approx \cosSim(v_x, v_y) = \frac{\sum_{i=1}^{T} w_i^x \times w_i^y}{\sqrt{\sum_{i=1}^{T} (w_i^x)^2} \times \sqrt{\sum_{i=1}^{T} (w_i^y)^2}}
\]

where \( w_i^x \) and \( w_i^y \) are the weights of the corresponding terms of the two vectors \( v_x \) and \( v_y \), respectively. \( T \) is the set of unique terms of the text corpus. The proposed semi-supervised business relationship mining algorithm (SemiBRM) is depicted in Figure 4. The algorithm first applies the NGSD measure to expand the seeding relationship indicators and constructs the context vectors of the set of business entities \( B \) identified at the business entity identification stage.

Cosine similarity scores are then computed to identify the subset of \( B \) with strong latent business associations. The type of business relationship is then disambiguated by examining the sentences containing a pair of companies with strong association and a specific type of relationship indicators. The frequency of competitive relationship \( \text{Freq}_{\text{com}}(x, y) \) (collaborative relationship \( \text{Freq}_{\text{col}}(x, y) \)) of an arbitrary pair of
companies \((x, y) \in BA\) is used as the basis to identify strong business association \((SBA)\) and prune the weak business relationships. After business relationship pruning, the relationship weight and the relationship type of each pair \((x, y)\) is updated. Finally, the mined business network \(G\) pertaining to a particular period is returned by the algorithm.

**Algorithm** `SemiBRM(D, B, SR, freq, gap freq, win, N)`

**Input:**
- \(D\) /* a text corpus for business relationship mining*
- \(B\) /* a set of business entity names*
- \(SR\) /* a set of seeding relationship indicators*

**Output:**
- \(G\) /* a mined business relationship graph of a time point*

**Main Procedure:**
1. Apply NGSD to expand the seeding relationship lexicon \(SR\) to identify a set of competitive (collaborative) relationship indicators \(Com\) \((Col)\);
2. Extract the virtual text windows of the set of business entities \(B\) from the corpus \(D\) according to the pre-defined window size \(win\);
3. Apply NGSD to build the context vectors for the set of business entities \(B\);
4. Apply cosine similarity measure \(CosSim\) to compute the cosine scores of \(B\);
5. For the top \(N\) pairs of business associations \(BA\) ranked according to \(CosSim\), disambiguate their business relationship type;
6. For each sentence of \(D\) containing a pair of business entities \((x, y) \in BA\), apply
   
   \[Com\] to compute \(Com(x, y) = \frac{Freq_{com}(x, y)}{Freq(x, y)}\), where \(Freq_{com}(x, y)\) is the number of sentences containing \((x, y)\) and a competitive relationship indicator; \(Freq(x, y)\) is the number of sentences containing \((x, y)\);
7. For each sentence of \(D\) containing a pair of business entities \((x, y) \in BA\), apply
   
   \[Col\] to compute \(Col(x, y) = \frac{Freq_{col}(x, y)}{Freq(x, y)}\);
8. Strong association \(SBA = \{(x, y) : Com(x, y) > freq \lor Col(x, y) > freq\}\);
9. The set of competitive companies \(Comp\) is defined by:
   \[Comp = \{(x, y) \in SBA : (Com(x, y) − Col(x, y)) > gap\}\]
10. The set of collaborative companies \(Coll\) is defined by:
    \[Coll = \{(x, y) \in SBA : (Col(x, y) − Com(x, y)) > gap\}\]
11. \(V = Comp \cup Coll\), \(E = \{e_{xy} : e_{xy} \in Comp \lor e_{xy} \in Coll\}\);
12. For each \(e_{xy} \in E\), \(w(e_{xy}) = \begin{cases} Com(x, y) & \text{if} \ e_{xy} \in Comp \\ Col(x, y) & \text{if} \ e_{xy} \in Coll \end{cases}\)
13. For each \(e_{xy} \in E\), \(t(e_{xy}) = \begin{cases} "competitive" & \text{if} \ e_{xy} \in Comp \\ "collaborative" & \text{if} \ e_{xy} \in Coll \end{cases}\)
14. Return \(G\).

**Figure 4.** The Proposed Semi-supervised Business Relationship Mining Algorithm

Figure 3 shows a business network pertaining to the “Technology and Equipment” Forbes industry by using Pajek, a shareware for graph plotting. The seeding companies are extracted from the “Technology and Equipment” industry, and solid (dash) lines represent collaborative (competitive)
relationships. “Apple Inc” is identified as the hub of this business sector mined based on a subset of our financial news corpus pertaining to 2009.

Figure 3. The Business Network of the Technology & Equipment Industry

4  EXPERIMENTS AND RESULTS

Based on the financial news articles and company data collected from Reuters and Yahoo! Finance, the effectiveness of the proposed computational methods for business entity identification and business relationship mining were evaluated. Other baseline methods were also applied to carry out the same tasks. The financial news corpus being used for our experiments included 200,114 news articles crawled from Reuters.com. These news articles covered the period from January 1, 2006 to December 31, 2009. A subset of this financial news corpus was manually annotated to create the benchmark evaluation dataset. For the evaluation of the proposed NER method, 314 financial news articles were annotated by two human annotators. When both human annotators agreed on a business entity (i.e., a company), the particular token would be annotated and added to the evaluation dataset. There were 1,532 annotated business entities and all of them belonged to the Forbes 2,000 companies (2010). For the evaluation of the proposed business relationship mining method, 766 financial news articles were manually annotated by the same human annotators. This evaluation data set consisted of 516 collaborative relationships and 497 competitive relationships. Similar to the task of annotating business entities, when both annotators agreed on a business relationship, the annotated relationship would be added to the benchmark evaluation dataset. The evaluation measures such as Precision, Recall, F-measure, and Accuracy commonly used in information retrieval research were applied to our experiments (Salton and McGill 1983).

For the first experiment, we examined the effectiveness of the proposed business entity identification method (AE). The experimental system implemented the proposed AE method which included automatic business abbreviations extraction and co-reference resolution. The first baseline system is the NER module of (GATE) which was extended by the full names of the Forbes 2,000 companies. In other words, automatic abbreviation extraction of business names through (Eq. 1) and co-reference resolution were not supported in the GATE baseline system. The second baseline system
(COREFERENCE) implemented the proposed proximity-based co-reference resolution algorithm. However, automatic abbreviation extraction of business names was not supported in the Co-reference baseline system. The parameter $\alpha = 0.66$ was empirically established and applied to this experiment.

<table>
<thead>
<tr>
<th>System</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>0.8969</td>
<td>0.8378</td>
<td>0.8663</td>
<td>0.8840</td>
</tr>
<tr>
<td>COREFERENCE</td>
<td>0.8185</td>
<td>0.8416</td>
<td>0.8299</td>
<td>0.8696</td>
</tr>
<tr>
<td>GATE</td>
<td>0.6456</td>
<td>0.9007</td>
<td>0.7521</td>
<td>0.8535</td>
</tr>
</tbody>
</table>

*Table 2. Comparative Performance of Business Entity Identification*

The experimental results are reported in Table 2. It is obvious that the experimental AE system the best recall while maintaining a comparable precision with the other baseline systems. The AE system outperforms the COREFERENCE system by 9.6% in terms of recall and 4.4% in terms of F-measure. The main reason of this performance improvement is brought by the extra functionality of the AE system which can recognize business entities by their abbreviations referred to in financial news articles. Since there are many business abbreviations used in financial news articles, the performance improvement achieved by the AE experimental system is obvious. Nevertheless, the precision of the AE system is not significantly higher than other baseline systems. After a careful examination of the identification results, we found that there are a number of tokens mistakenly identified as business abbreviations. It reveals that the proposed automatic business abbreviation extraction method is not perfect although it does lead to considerable performance improvement as a whole (e.g., 4.4% improvement of F-measure). Our experimental result also shows that the performance of the GATE baseline system which does not support business abbreviation extraction and co-reference resolution is the worst for the business entity identification task.

For the second experiment, we examined the effectiveness of the proposed business relationship mining method. This task involves 3 classes (i.e., collaborative, competitive, neither). We assessed the performance of the collaborative relationship classification task and the competitive relationship classification task separately. The proposed semi-supervised business relationship mining algorithm (SemiBRM) was implemented as the experimental system. The parameters $\sigma_{\text{freq}} = 0.55$, $\sigma_{\text{gap}} = 0.12$, $N=3,000$ were empirically established and applied to this experiment. One of the baseline systems (NOEXPAND) used the 20 seeding relationship indicators alone for business relationship mining. Both the SemiBRM and the NOEXPAND methods employed the proposed business abbreviation and co-reference resolution methods. The second baseline method employed a state-of-the-art supervised machine learning classifier, called (SVM-struct) which had been successfully applied to sequence labelling of texts (Tschantaridis et al. 2005). For each business relationship classification task, 70% of the annotated positive examples were used in the training set and the remaining 30% of the positive examples were used in the test set. The same number of negative examples (i.e., 50% positive and 50% negative) was added to the training set and the test set, respectively. This 70-30% split of the evaluation dataset was repeated ten times to produce ten test sets. The experimental and baseline systems were then applied to ten test sets to produce the average performance figures.

The results of our experiment were tabulated in Table 3. For both the collaborative and the competitive business relationship identification tasks, it is clear that the SemiBRM system outperforms the other baseline systems. More specifically, for the collaborative relationship classification task, the SemiBRM system outperforms SVM-struct by 6.0% in terms of F-measure. Surprisingly, the SVM-struct system does not perform much better than the NOEXPAND system which uses relatively little business relationship knowledge for the mining task. The reason may be that the SVM-struct classifier cannot establish near optimal feature weights to separate the two classes given a relatively small training set. For the competitive relationship classification task, the SemiBRM system outperforms SVM-struct by 5.4% in terms of F-measure. In practice, it is difficult to apply a
supervised machine learning classifier like SVM-struct to business relationship mining because a large number of labeled business relationships are rarely available in the real-world. The experimental results show that the proposed semi-supervised statistical learning method is effective. The main reason for the SemiBRM algorithm achieving better precision and recall than that of other baseline methods is that relevant relationship indicators can automatically be learnt from a large training corpus based on a small number of seeding relationship indicators. The resulting rich set of relationship indicators can then be utilized to improve the system’s performance on business relationship classification. The additional advantage of the SemiBRM algorithm is that labeled training examples are not required for real-world business relationship mining applications.

### Comparative Performance of Business Relationship Classification

<table>
<thead>
<tr>
<th>System</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemiBRM</td>
<td>0.6574</td>
<td>0.6044</td>
<td>0.6298</td>
<td>0.7610</td>
</tr>
<tr>
<td>SVM-struct</td>
<td>0.5896</td>
<td>0.5992</td>
<td>0.5944</td>
<td>0.7331</td>
</tr>
<tr>
<td>NOEXPAND</td>
<td>0.5857</td>
<td>0.6025</td>
<td>0.5939</td>
<td>0.7231</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemiBRM</td>
<td>0.6250</td>
<td>0.5814</td>
<td>0.6024</td>
<td>0.7438</td>
</tr>
<tr>
<td>SVM-struct</td>
<td>0.5625</td>
<td>0.5806</td>
<td>0.5714</td>
<td>0.7156</td>
</tr>
<tr>
<td>NOEXPAND</td>
<td>0.5563</td>
<td>0.5779</td>
<td>0.5669</td>
<td>0.7125</td>
</tr>
</tbody>
</table>

Table 3. Comparative Performance of Business Relationship Classification

### 5 CONCLUSIONS AND FUTURE WORK

Although some research work was performed for the automatic construction and analysis of social networks, little work has been done for the discovery and analysis of business networks. The main contribution of this paper is the design and development of a novel semi-supervised business relationship mining method and its instantiation. Experimental results confirm that the proposed business network mining method is more effective than a state-of-the-art supervised machine learning method. The distinct advantage of the proposed computational method is that manually labeled training examples are not required. Accordingly, it facilitates the application of the proposed method to real-world applications. Future work involves evaluating the effectiveness of the proposed method using a larger benchmark dataset and comparing the performance of the proposed method with other existing systems. The properties of the discovered business networks will be examined to develop effective business prestige metrics to predict business performance.

### References


