The sheer volume of user-contributed data on the Internet has motivated organizations to explore the collective business intelligence (BI) for improving business decisions making. One common problem for BI extraction is to accurately identify the entities being referred to in user-contributed comments. 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. The main contribution of this paper is the illustration of a novel semi-supervised method for the identification of business entities (e.g., companies), and hence to automatically construct business networks. Based on the automatically discovered business networks, financial analysts can then predict the business prestige of companies for better financial investment decision making. Initial experiments show that the proposed NER method for business entity identification is more effective than other baseline methods. Moreover, the proposed semi-supervised business relationship extraction method is more effective than the state-of-the-art supervised machine learning classifiers when there are not many training examples available. Our research work contributes to advance the computational methods for the extraction of entities and their relationships from texts.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Text Mining; I.2.6 [Artificial Intelligence]: Learning; I.2.7 [Artificial Intelligence]: Natural Language Processing

General Terms
Algorithms, Performance, Experimentation

Keywords
Named Entity Recognition, Text Mining, Statistical Learning, Business Network Discovery

1. INTRODUCTION

The ubiquitous Web 2.0 applications have brought to organizations with unprecedented opportunities to exploit market intelligence and develop deep insights about their customers, business partners, and competitors [6, 7]. Due to the problem of information overload [13, 14], manual extraction of business intelligence from the flood of un-structured data coming from the Internet is not practical. Research effort has been devoted to the construction and analysis of social networks based on user-contributed data [25, 30], and hence to improve marketing effectiveness and reduce marketing costs [2, 17, 26]. Although some research work has been performed for the automatic construction and analysis of social networks [5, 25, 30], little work has been done for the discovery and analysis of business networks. Business networks can be seen as one kind of social network [17, 31].

The first step toward business network discovery is to identify the business entities referred to in the user-contributed free text. In fact, this is a very challenging task because abbreviations and pronouns are often used to refer to business entities in natural languages.

Figure 1: The Problem of Co-reference Resolution for Business Entity Identification

Figure 1 highlights the difficulties of business entity identification using the company “General Electric” as an example. As can be seen, “General Electric” is also referred to as GE (an abbreviation) or co-referenced by “it” or “the company”. To deal with the first problem, we apply a semi-supervised learning method to automatically extract various abbreviations statistically associated with the particular business entity name in a training corpus. In addition, we propose a heuristic rule-based approach to deal with co-references
One of the main contributions of our research reported in this paper is the development of a semi-supervised statistical learning method for the identification of abbreviations of business entities in financial comments. The second contribution of our research work is the development of a novel algorithm, called proximity-based backward searching (PBS) to resolve the co-references of business entities. Finally, by successfully identifying business entities, we show our novel computational method for the automatic discovery of business networks. The main advantage of our proposed computational methods is that minimal human intervention is involved. As a result, the proposed method has better chance to be applied to real-world applications requiring the functionality of business entity identification.

2. RELATED RESEARCH WORK

There are two common approaches to conduct named entity recognition, namely, rule-based approach [4, 32] and learning-based approach [21, 29]. For rule-based approach, manually pre-defined heuristic or linguistic rules are applied to identify specific types of entities such as people, organizations, places, etc. [4, 32]. However, the main weakness of this approach is that the pre-defined rules may not be able to cover a variety of situations, particularly when they are applied to different domains. On the other hand, learning-based approach utilizes training corpora to automatically build classifiers to identify entities in unseen documents [21, 29]. This approach is quite flexible since it can be applied to different domains as long as the classifiers are re-trained using labeled training examples of a new domain. Nevertheless, the main weakness of this approach is the requirement of manually labeling a large number of training examples. In this paper, we propose a statistical inference-based method to address the low recall problem in NER while avoiding the time consuming process of labeling training examples.

Business relationship mining techniques can be broadly classified into content-based or link-based approach. In a previous study, the contents of online financial news were scanned to estimate the co-occurrence statistics of a pair of stock tickers [2]. The co-occurrence statistics were then used to predict the possible relationships among companies. Moreover, the membership of a company pertaining to an industry sector was estimated [2]. The CoMiner system made use of NLP techniques and several pre-defined syntactic patterns (e.g., company-A “versus” company-B) to identify competitive company relationships based on a Web corpus [1, 16]. Each syntactic pattern was assigned a weight and the Point-wise Mutual Information (PMI) measure was used to estimate the strength of competitive relationships between two companies. However, relationship mining according to pre-defined syntactic patterns may encounter the low recall problem since natural languages used in financial news are very flexible. One main innovation of the business relationship discovery method proposed in this paper is the development of a statistical inference based method to automatically expand some seedling syntactic patterns to address the low recall problem.

A link-based approach was developed to extract competitor relations from online financial news [18]. Company names identification was conducted based on stock tickers appearing in financial news. A weighted directed graph approach was adopted. If company y was mentioned in a financial news cataloged for company x, a directed business relationship from x to y is assumed. Twelve graph-based features such as in-degree, out-degree, total-degree, PageRank, etc. were taken as features and fed into four supervised classifiers. A similar link-based approach was also applied to predict the binary company revenue relations (e.g., whether company x has higher revenue than company y) based on online financial news [17].

A hybrid approach of both hyperlinks and contents of Web pages were exploited to discover hidden competitor relationships [20]. Both network-based and content-based features were used by supervised classifiers for competitive business relationships discovery. Instead of assuming that co-occurring companies in Web pages imply competitive relationship, our proposed approach examines different kinds of company relationships according to relational keywords found at the sentence level. The CopeOpi system was proposed to extract opinions and implicit business relationships of some targeting entities based on the Web corpus [10, 11]. Business associations among the target entities are discovered based on the opinion-tracking plots (i.e., the pattern of opinion scores exhibited over time).

The CoNet system employed shallow natural language processing (NLP) techniques to identify commercial entities and relationships from online financial news [31]. Simple linguistic rules were applied to identify the abbreviations of company names and resolve the co-references of company names. Commercial relationship mining was performed according to a set of pre-defined relationship keywords. However, such an approach may suffer from low recall due to the limited number of pre-defined relationship keywords. Latent semantic indexing (LSI) was also examined to identify relationships among entities of interests [3]. 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 of the corresponding entity vectors in the LSI space provides a quantitative measure of the degree of contextual association between the entities. The ArnetMiner system employed a supervised machine learning method, conditional random fields (CRF), to identify researcher and their relationships based on annotated Web pages [25]. Genetic algorithm was applied to discover communication networks based on the contents of email messages of the Enron corpus [30]. OpenCalais is a commercially available Web Service that employs NLP and machine learning techniques for NER, entity relationship identification, and event tagging (e.g., business events). The Knowledge Base Population Track of the annual Text Analysis Conference (TAC) also explores novel methods for the identification of entity-linking [8].

3. THE COMPUTATIONAL MODELS

3.1 Business Entity Identification

The first step toward business relationship discovery is to identify the company names in the Web 2.0 documents (e.g., user-contributed financial comments). 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) related to a company name should be resolved.

http://www.opencalais.com/documentation/
The procedures for business entity identification can be summarized as follows:

1. Extract business full names and stock ticker labels from Web sources (e.g., Yahoo! Finance);
2. 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;
3. Apply general business entity identification rules to identify the remaining organization names;
4. Apply the proximity-based backward searching algorithm to resolve the co-references in each document.

The business full names and the stock tickers of companies extracted from Yahoo! Finance are passed to our NER module to compose the basic business name dictionary for preliminary name identification. The general business entity identification rules are also passed to our NER module for the identification of business entities. For instance, the tokens with title cases and preceding "Inc", "Co. Ltd.", "Corporation", etc. are automatically labeled as organization names. To automatically extract the abbreviations of business entities, a mutual information 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 could appear frequently in a corpus.

More specifically, we extend the BMI measure [15], a variant of the standard mutual information (MI) measure [24], to conduct statistical inference for the abbreviations of business entities. The BMI measure has been successfully applied to context-sensitive text mining for IR [12] and automatic domain concept extraction in ontology discovery [15]. 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 extended to develop the proposed abbreviation extractor:

$$ae(t_i, t_j) = \alpha \times Pr(t_i, t_j) \log_2 \left( \frac{Pr(t_i, t_j)}{Pr(t_i)Pr(t_j)} - (1 - \alpha) \times |Pr(t_i, -t_j)| \log_2 \left( \frac{Pr(t_i, -t_j)}{Pr(t_i)Pr(-t_j)} \right) + Pr(-t_i, t_j) \log_2 \left( \frac{Pr(-t_i, t_j)}{Pr(-t_i)Pr(t_j)} \right) \right)$$

(1)

where abbreviation extractor $ae(t_i, t_j)$ is a function to "infer" the statistical association between two terms $t_i$ and $t_j$. For our application, one of the terms is the seeding business full name. The parameter $\alpha \in [0, 1]$ was used to adjust the relative weight of positive and negative evidence respectively [15]. $Pr(t_i, t_j)$ is the joint probability that both terms appear in a text window, and $Pr(t_i)$ is the probability that a term $t_i$ appears in a text window. According to previous studies in information retrieval and ontology discovery, a text window of 5 to 10 tokens is effective [12, 15]. This text window introduces a constraint (i.e., a proximity factor) to the inference process of business abbreviations 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 tend to appear in adjacent sentences rather than occurring within the same sentence. Therefore, we set the window size of two paragraphs for this NER application. The probability $Pr(t_i)$ is estimated based on $|w|$ where $|w|$ is the number of text windows containing the term $t$ and $|w|$ is the total number of text windows constructed from an unlabeled training corpus. Similarly, $Pr(t_i, t_j)$ is the fraction of the number of windows containing both terms out of the total number of text windows. Negation such as $Pr(t_i, -t_j)$ is interpreted in the way that $t_j$ but not $t_i$ appears in a text window. After computing the abbreviation extraction scores of the tokens with specific POS (e.g., no valid POS found) and specific properties (e.g., with title case and followed by "Inc", "Co. Ltd.", "Corporation", etc.), the top $n$ tokens with the highest abbreviation extraction scores are selected as the abbreviations of the business full names. For co-reference resolution, the proximity-based backward searching algorithm depicted in Figure 2 is applied.

**Algorithm: Proximity-Based Search**

Input: $d_n$, a document (i.e., financial news)
Output: $d_{as}$, a document with co-references resolved

Main Procedure:
1. $d_{as} = d_n$; /* initialize the returned object
2. $\alpha \leftarrow \text{CountTokens}(d_{as})$; /* count # of tokens in document
3. FOR $|d| > 1$ TO $0$
4. $i \leftarrow \text{ReadToken}();$ $d_{as}();$
5. IF $\text{Poss}(t_i) = "\text{company}"$ /* the POS of the token is pronoun
6. $j = 1; t = 1$
7. WHILE $j < \text{CountTokens}(d_{as})$ /* loop backward in document
8. $d_{as}();$
9. IF Type($t_i$) = "STRING" /* not recognized entity
10. $|t| = t_i$
11. ELSE
12. IF Type($t_i$) = "organization" /* type of entity
13. $d_{as}(); j = t_i;$ /* replace co-reference
14. ENDIF
15. $i = 0$ /* quit backward search
16. ENDIF
17. END
18. END
19. RETURN $d_{as}$;

**Figure 2: The Proximity-based Backward Searching Algorithm**

Within each document $d_n$, all the business entities are first identified based on the company table (containing full names, stock tickers, and abbreviations). In addition, generic business entity identification rules (e.g., title cases followed by "Co. Ltd.") are also applied to identify business entities. The proximity-based backward search algorithm then locates the first ambiguous pronoun or token (e.g., "company") $t_1 \in d_n$. Then, the algorithm works backward to search for the nearest business entity already identified by the previous NER procedure. The proximity is globally defined in advance; for our experiment reported in this paper, the proximity of two paragraphs are defined. If a nearest business entity is found and it is located within the two paragraphs boundary, 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.

3.2 Business Relationship Mining

One unique feature of the proposed business relationship mining method is that it is based on an un-supervised statistical inference technique. The main advantage is that manually labeled training examples are not required for the proposed method, and hence it improves the chance of deploying the proposed methodology to support real-world applications. Similar to the ideas proposed by Turney and Littman [28] for automatic opinion indicators expansion, we develop a statistical inference method to automatically extract a set of domain-specific relationship indicators based on some seeding relationship indicators. In particular, 10 collaboration indicators (e.g., cooperate, ally, collaborate, joint, own, partner, coordinate, engage, partnership, agree) and other 10 competition indicators (e.g., compete, challenge, against, vie, contend, fight, contest, dispute, battle, accuse) were used as the seeding relationship indicators. The synonyms of these indicators are also extracted from WordNet [19] automatically. The initial set of relationship indicators is used by the statistical inference module to generate the relationship lexicon. Then, business relationship mining is conducted based on the system generated relationship lexicon. When compared to the method proposed by Turney and Littman [28], the improvements of our method includes using a shallow NLP method to identify specific POS for relationship indicators mining, and the application of WordNet to filter noisy relationship indicators generated by the mining process. Above all an enhanced mutual information based metric (Eq.1) is applied to identify additional relationship indicators statistically correlated with the seeding relationship indicators.

For business relationship mining, our system first utilizes the proposed NER method to identify the valid lexical patterns such as (organization, relation, organization) and (relation, organization, organization), or (organization, organization, relation). A proximity threshold $\omega_{prox}$ is applied to measure the proximity between a relationship indicator and a company name. If the distances (by words) between a relationship indicator and the respective companies are all below the proximity threshold $\omega_{prox}$, a lexical pattern is considered a valid pattern. Then, each valid lexical pattern is classified as competitive or collaborative based on the relationship indicators captured in the relationship lexicon. The functions $Coll(x, y) = \frac{Freq_{coll}(x,y)}{Freq(x,y)}$ and $Comp(x, y) = \frac{Freq_{comp}(x,y)}{Freq(x,y)}$ are used to derive the collaborative and the competitive scores for each pair of companies $(x, y)$ identified from the financial news corpus. $Freq_{coll}(x,y)$ is the occurrence frequency of a collaborative relationship involving $x$ and $y$. $Freq_{comp}(x,y)$ is the occurrence frequency of a competitive relationship involving $x$ and $y$. $Freq(x,y)$ is the occurrence frequency of any recognized relationships involving $x$ and $y$. To prune noisy business relationships, only the business relationship with $Coll(x, y) \geq Comp(x, y)$ score greater than a relationship threshold $\omega_{gap}$ will be extracted by our system. If $Coll(x, y) - Comp(x, y) \geq \omega_{gap}$ ($Comp(x, y) - Coll(x, y) > \omega_{gap}$) is true, the pair of companies are considered collaborative (competitive). The threshold $\omega_{gap}$ is used to filter significant collaborative (competitive) relationships. The set of significant relationships pertaining to each industry can then be used to build a business network for that industry.

We use Pajek\(^2\), a shareware for graph plotting, to visualize the discovered business networks. Figure 3 shows a sample business network discovered by our system. The top financial institutions included in the 2010 Forbes 2,000 list (Banking industry) were used as the seeds, and the news collected from Reuters for the period from 2006 to 2009 were used as the source. Solid lines indicate collaborative relationships and dash lines indicate competitive relationships. It seems that the global banking sector operates in a collaborative rather than a competitive manner in general. Such a network helps a financial analyst or business manager quickly identify which company is the leader and which company is under keen competition in a specific business sector.

![Figure 3: The Business Network Discovered Based on the Forbes Banking Industry](http://pajek.imfm.si/doku.php?id=pajek)

4. EXPERIMENTS AND RESULTS

Based on the financial news and company data collected from Reuters and Yahoo! Finance, the effectiveness of the proposed computational models for NER and business relationship mining were evaluated. Other baseline methods were also applied to carry out the same tasks. One of the main difficulties of evaluating the proposed computational models is the construction of a test dataset of reasonable size. It is quite labor-intensive to manually annotated companies and business relationships. A subset of companies from the 2010 Forbes 2,000 companies was used as our test cases. A total of 474,716 financial news were crawled from Reuters for the period from January 1, 2005 to December 31, 2009. A subset of the 2009 financial documents was manually annotated to build the evaluation dataset. For the evaluation of the proposed NER method, 314 financial news from among the downloaded financial news 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. For the evaluation of the proposed business relationship mining method, 2,074 sentences were annotated by other two human annotators. Similar to the annotation of business entities, when both annotators agreed on a business relationship, the annotated relationship would be added to the evaluation dataset. There were 839 collaborative relationships and 535 competitive relationships in the dataset. The evaluation measures such as Precision, Recall, and F-measure commonly used in information retrieval and opinion mining research were applied to our experiments [22].

For the first experiment, we examined the effectiveness of the proposed NER method (AE). The first baseline (BASE1) was the NER model developed based on the business full name and the generic business entity identification rules. In other words, automatic abbreviation extraction of business names (Eq.1) and co-reference resolution were not supported. The second baseline (BASE2) incorporate the proposed proximity-based backward searching co-reference resolution algorithm, but automatic automatic abbreviation extraction of business names was not supported in BASE1. The parameter $\alpha = 0.57$ was empirically established based on a subset of the evaluation dataset. The experimental results are reported in Table 1. It is obvious that the proposed NER method with semi-supervised abbreviation extraction and proximity-based backward searching co-reference resolution achieves the best recall while maintaining a comparable precision with the other baseline methods. AE outperforms BASE2 by 9.7% in terms of recall and 4.5% in terms of F-measure. As expected, the performance of the first baseline method is the worst in business entity identification because of the lack of capability of recognizing the abbreviations of business entities and resolving co-references in free text.

Table 1: Comparative Performance of Business Entity Identification

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>0.898</td>
<td>0.839</td>
<td>0.868</td>
<td>0.885</td>
</tr>
<tr>
<td>BASE2</td>
<td>0.819</td>
<td>0.842</td>
<td>0.829</td>
<td>0.869</td>
</tr>
<tr>
<td>BASE1</td>
<td>0.631</td>
<td>0.881</td>
<td>0.735</td>
<td>0.852</td>
</tr>
</tbody>
</table>

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 statistical inference method (SI) (Eq.1) was used to construct the relationship lexicon for business relationship identification. One of the baseline systems (BASIC) used the 20 seeding relationship indicators alone. Both the SI and the BASIC methods employed the proposed business abbreviation and co-reference resolution methods. The second baseline method employed a state-of-the-art supervised classifier, the conditional random fields (CRF) classifier [9, 23] to label the sequences of company relationships. CRF is based on a discriminative probabilistic model (an undirected graph model) with each vertex representing a random variable, and the edge indicating the dependency between two random variables. The ultimate goal is to estimate the probability distribution of each vertex based on some training data. A publicly available Java-based implementation of the CRF classifier was used in this experiment. Finally, the SVM-struct supervised classifier for sequence labeling was also applied [27].

Table 2: Comparative Performance of Business Relationship Classification

<table>
<thead>
<tr>
<th>Method</th>
<th>Collaborative Business Relationships</th>
<th>Competitive Business Relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>Precision</td>
<td>F-measure</td>
</tr>
<tr>
<td>SI</td>
<td>0.701</td>
<td>0.638</td>
</tr>
<tr>
<td>CRF</td>
<td>0.626</td>
<td>0.638</td>
</tr>
<tr>
<td>SVM-struct</td>
<td>0.618</td>
<td>0.633</td>
</tr>
<tr>
<td>BASIC</td>
<td>0.622</td>
<td>0.639</td>
</tr>
</tbody>
</table>

For each kind of relationship classification, 70% of the annotated positive examples were used in the training set and the reaming 30% of the positive examples were used in the test set. The same number of negative examples was added to the training set and the test set, respectively. The results of our experiment were tabulated in Table 2. For both the collaborative and the competitive business relationship identification tasks, it is clear that the SI method outperforms the other baseline methods. More specifically, the SI method outperforms CRF and SVM-struct in terms of F-measure by 5.7% and 6.9%, respectively for the collaborative relationship classification task. Surprisingly, both CRF and SVM-struct did not perform very well given a relatively small training set. However, the training processes for CRF and SVM-struct are realistic 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 inference method can be successfully applied to build a relationship lexicon, which eventually improves the performance of business relationship identification. The additional advantage of the proposed method is that labeled training examples are not required for business relationship mining.

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 contributions of this paper include: (1) the development of a novel semi-supervised business entity recognition method; (2) the development of a novel semi-supervised business network discovery method; (3) the empirical evaluation of the proposed computational methods based on real-world datasets. Experimental results confirm that the proposed business entity identification and business network discovery methods are effective. More specifically, the proposed business relationship discovery method outperforms state-of-the-art supervised machine learning classifiers for the respective classification tasks. The distinct advantage of the proposed computational methods is that

manually labeled training examples are not required. Accordingly, it facilitates the application of the proposed computational methods to real-world applications. Future work involves evaluating the effectiveness of the proposed methods using a larger dataset and comparing the performance of the proposed methods with other well-known systems. The properties of the discovered business networks will be examined to develop effective business prestige metrics to predict business performance.

6. REFERENCES


