Mining Suppliers from Online News Documents

Chih-Ping Wei  
*National Taiwan University, cpwei@im.ntu.edu.tw*

Lien-Chin Chen  
*National Taiwan University, lcchen101@ntu.edu.tw*

Hung-Yao Chen  
*National Taiwan University, r99725039@ntu.edu.tw*

Chin-Sheng Yang  
*Yuan Ze University, csyang@saturn.yzu.edu.tw*

Follow this and additional works at: [http://aisel.aisnet.org/pacis2013](http://aisel.aisnet.org/pacis2013)

**Recommended Citation**  
Wei, Chih-Ping; Chen, Lien-Chin; Chen, Hung-Yao; and Yang, Chin-Sheng, "Mining Suppliers from Online News Documents" (2013). *PACIS 2013 Proceedings*. 261.  
[http://aisel.aisnet.org/pacis2013/261](http://aisel.aisnet.org/pacis2013/261)
MINING SUPPLIERS FROM ONLINE NEWS DOCUMENTS

Chih-Ping Wei, Department of Information Management, National Taiwan University, Taipei, Taiwan, R.O.C., cpwei@im.ntu.edu.tw
Lien-Chin Chen, Department of Information Management, National Taiwan University, Taipei, Taiwan, R.O.C., lcchen101@ntu.edu.tw
Hung-Yao Chen, Department of Information Management, National Taiwan University, Taipei, Taiwan, R.O.C., r99725039@ntu.edu.tw
Chin-Sheng Yang, Department of Information Management, College of Informatics, Yuan Ze University, Chungli, Taiwan, R.O.C. csyang@saturn.yzu.edu.tw

Abstract

Supplier intelligence represents an important type of competitive intelligence essential to organizational strategic planning and analysis, because the moves that the suppliers of a firm make will affect the business network’s health, which in turn will impact the performance of the firm. Central to supplier intelligence is to identify who the suppliers of a focal firm and its competitors are. Thus, an effective supplier mining technique that automatically identifies and discovers the suppliers of a firm and its competitors from publicly available information sources (news documents in this study) is desperately needed. In this study, we exploit text and link mining techniques to construct a supplier relationship mining (SRM) system to automatically discover supplier relationships concerning a focal company from online business news documents. Our empirical evaluation result suggests that our proposed SRM system outperforms its benchmark method in precision and recall.

Keywords: Supplier Relationship Mining, Supply Chain Management, Online News, Text Mining, Data Mining.
The business environment is increasingly volatile and turbulent. Firms must adapt to their environments proactively and swiftly in order to survive and prosper. Environmental scanning is the first link in the development of competitive intelligence that permits a firm to adapt to its environment and subsequently to make informed strategic decisions and develop effective responses to secure or improve its position in the future (Chen et al. 2002; Choo 1999; Jennings & Lumpkin 1992). Several prior empirical studies demonstrate a positive correlation between competitive intelligence endeavor (or, more specifically, environmental scanning effort) and organizational performance (Daft et al. 1988; Subramanian & Ishak 1998; Ahituv et al. 1998). So, frequently and broadly monitoring environment and competitors’ activities is crucial to the future of a successful firm.

Competitive intelligence refers to the continuous process of gathering, analyzing, and disseminating information and knowledge about a firm’s customers, competitors, suppliers, and political environment in order to support decision making process for sustaining and enhancing competitiveness of the firm (Anica-Popa & Cucui 2009; Teo & Choo 2001; Wright & Calof 2006). Generally, competitive intelligence can be classified into four major types: marketing intelligence, technological intelligence, competitor intelligence, and supplier intelligence (Fleisher 2008; Rouach & Santi 2001). Prior studies are largely devoted to marketing and technological intelligence (Castells et al. 2000; Fernandez et al. 1999; Fleisher 2008; Kline 2000; Tan & Ahmed 1999; Xu & Kaye 1995), but pay less research attention to the others, especially supplier intelligence. Moreover, competitor and supplier intelligence are fundamental to marketing and technological intelligence, because their analyses greatly depend on a firm’s understanding of its competitors and suppliers. So, the goal of this study is to focus on supplier intelligence for mining supplier relationships of a focal firm.

A firm is not isolated; rather, it exists and operates in a business ecosystem that includes its suppliers, distributors, outsourcing firms and a host of other organizations and competes with the ecosystems of its competitors. Each firm in its business ecosystem, similar to an individual species in a biological ecosystem, ultimately shares the fate of the business network as a whole, regardless of the firm’s apparent strength (Iansiti & Levien 2004). The moves that the suppliers (or other organizations in the ecosystem) of a focal firm make will, to varying degrees, affect the business network’s health, which in turn will impact the performance of the focal firm (Iansiti & Levien 2004). Thus, the focal firm should concern the financial health of its suppliers, their cost and quality problems, and possible acquisition and/or alliances pertaining to suppliers (Herring 1999) and also need to be aware of the statuses and development of its competitors’ suppliers. As with competitor mining to competitor intelligence, supplier mining is a necessary precursor of supplier intelligence discovery and analysis. In this vein, an effective supplier mining technique that automatically identifies and discovers the suppliers of a focal firm from publicly available information sources (business news documents in this study) is desperately needed.

Several competitor mining techniques have been developed for automatically identifying firms’ competitors from business news documents and/or Web pages (Bao et al. 2008; Pant & Sheng 2009; Ma et al. 2010). Our analysis of the literature suggests that supplier mining has not been investigated by prior research. However, competitor relationships are fundamentally different from supplier relationships, existing competitor mining techniques cannot directly apply to supplier mining. In response, we will propose a supplier relationship mining (SRM) system that automatically discovers supplier relationships concerning a focal company from online business news documents and generate a supplier/customer graph of a focal company. The rest of the paper is organized as follows: Section 2 reviews the literature relevant to this study. Section 3 describes the detail design of our proposed SRM system. In Section 4, we depict our data collection and the evaluation results. Finally, we present the concluding remarks in Section 5.
2 LITERATURE REVIEW

To our best knowledge, supplier mining has not been investigated by prior research. Therefore, in this section, we review prior research on competitor mining and analyze their fundamental limitations to justify the motivation of our research study.

Bao et al. (2008) develop the CoMiner system that automatically extracts and then ranks likely competitors of a focal company. Specifically, given a focal firm, CoMiner first adopts the information extraction approach, predefines a set of linguistic patterns, sends each of these predefined linguistic patterns to a search engine to query the top 100 returned web pages. Then, CoMiner uses the collection of the returned web pages across all predefined linguistic patterns to extract and then rank likely competitors of the focal firm. The major limitations of CoMiner are as follows: 1) the construction of the set of linguistic patterns for a specific business relationship (i.e., competitors) is knowledge-intensive, and 2) the effectiveness of competitor mining greatly depends on whether the web pages retrieved on the basis of the predefined linguistic patterns truly contain the business relationship of interest (i.e., competitors). If many of the web pages retrieved do not pertain to the business relationship of interest, the subsequent competitor identification tasks are based on a very noisy data set and thus will significantly degrade the effectiveness of CoMiner. Competitor relationships are more often revealed (i.e., a firm and its competitors more often co-occur) in web pages and business news documents, but supplier relationships (i.e., a firm and its suppliers) are not. As a result, the aforementioned limitations of CoMiner constrain its practicability to our study.

Pant and Sheng (2009) assume that competitors are often associated with a similar set of companies. Thus, an overlap between web pages that link to (i.e., in-links) two companies’ web sites may be an indication of their substitutability and therefore a signal of potential competitor relationship. Similarly, an overlap between web pages that are linked from (i.e., out-links) two companies’ web sites may also signify that the two companies are competitors. Furthermore, because the content of the web site of a company provides a description of the company (e.g., its products and services), an overlap between the self-descriptions of companies can also serve as a signal of their substitutability (i.e., competitor relationship). Accordingly, besides the co-occurrences of two companies on online business news documents and search engine results, Pant and Sheng (2009) propose three web metrics (i.e., in-link similarity, out-link similarity, and text similarity) based on linkage structure and web site content (i.e., news count and search engine count) as additional predictors for identifying competitor relationships between companies. In terms of underlying algorithm for competitor mining, Pant and Sheng (2009) take the supervised learning approach and use C4.5 (Quinlan 1993) decision tree induction or logistic regression to train a competitor identification model from a set of competitor-pairs and non-competitor-pairs as training instances. Subsequently, a company-pair represented using the same set of input attributes is fed into the previously induced competitor identification model to predict whether the two companies are competitors or not.

The technique developed by Pant and Sheng (2009) follows the supervised learning approach and thus can avoid the first limitation of CoMiner (i.e., the need to construct a set of linguistic patterns for a specific business relationship). However, their technique also has several limitations. When web pages or links do not depict any formal relationships with the concerning companies, they represent serious noises when estimating in-link similarity and out-link similarity metrics and will undermine the effectiveness of this technique. In addition, the quality of news count and search engine count metrics for two given companies highly depends on whether the news documents and web pages retrieved discuss or contain the two companies being competitors. Imaginably, many news documents and web pages retrieved in effect involve other relationships other than competitor relationships between the two given companies. When this happens considerably, the news count and search engine count metrics may not be reliable indicators for identifying whether two companies are competitors or not. Finally, the technique developed by Pant and Sheng (2009) is for discovering competitor relationships and cannot directly be applied to mining supplier relationships between companies.
The competitor mining technique proposed by Ma et al. (2010) uses the co-occurrence of stock tickers of companies in business news documents to create connections between companies and then employs the structural properties of the resulting network to predict competitor relationships. Specifically, given a collection of business news documents organized by company, they first identify company citations in news documents and then construct a directed, weighted intercompany network from the company citations. Subsequently, four types of structural predictors (i.e., attributes) derivable from the network structure are employed for predicting competitor relationship purpose, including dyad degree-based, node degree-based, centrality-based, and structural equivalence (SE)-based. The centrality-based attributes are incorporated under the assumption that a more important company is more likely to have a relationship with a given company than is a less important one. Three centrality-based attributes, including PageRank (Brin & Page 1998), HITS (Kleinberg 1999), and betweenness centrality (Brandes 2001), commonly used in social network analysis are included because they can capture the importance of a node in the whole network. As with the study of Pant and Sheng (2009), the competitor mining technique by Ma et al. (2010) takes the supervised learning approach (using artificial neural network, Bayes network, C4.5 decision tree induction, and logistic regression as alternative underlying learning algorithm) and induces a competitor identification model from a training set of competitor-pairs and non-competitor-pairs. Given an intercompany network, the competitor identification model will predict for each pair of companies whether they are competitors or not. Because the competitor mining technique proposed by Ma et al. (2010) also follows the supervised learning approach and thus does not suffer from the CoMiner’s limitation inherent to the use of the information extraction approach. However, because all attributes employed are solely derived from the network structure of a given intercompany network (i.e., company citation network), this technique shares similar limitations as the technique proposed by Pant and Sheng (2009). For instance, the quality of the company citation network is highly affected by whether the corresponding news stories truly discuss or contain the two companies being competitors. If many news stories used to calculate the citation frequency from one company to another cover other relationships rather than competitor relationships, the corresponding citation frequency will be biased and in turn will affect the accuracy of the attribute values derived from the intercompany network. Finally, as with the two previous competitor mining techniques, the technique proposed by Ma et al. (2010) is for discovering competitor relationships only and is not applicable to mining supplier relationships between companies.

3 THE PROPOSED METHOD

The main purpose of our proposed method, namely Supplier Relationship Mining (SRM), is used to discover supplier relationships for a user-specific focal company on the basis of online business news documents. Figure 1 shows the overall process of SRM. There are three main phases in our proposed SRM system, including sentence classification (phase 1), direction classification (phase 2), and link assessment (phase 3). First, the sentence classification phase is used to classify whether a sentence contains a supplier relationship. Second, the goal of direction classification is to classify who is supplier and who is customer in a sentence containing supplier relationship. Third, the link assessment phase is applied to the generated supplier/customer graph produced by the previous phase (i.e., phase 2) for distinguishing true supplier links from false ones in the graph. Finally, a refined supplier/customer graph is reported to illustrate the supplier/customer relationships of the focal company. We will describe the details of these three phases in the following.

3.1 Sentence Classification (Phase 1)

The goal of sentence classification is to learn a sentence classification model by using a set of training sentences, each of which is preclassified into two classes: positive (i.e., containing information about supplier relationship) and negative (not containing information about supplier relationship). First, we apply the bag-of-words model to extract features from sentences. Second, for feature selection, the information gain (IG) measure is used to filter out terms with no or less discriminating power and
select top $K_{SC}$ terms to represent each sentence. Let $c_1$ and $c_2$ denote the two classes of sentences (i.e., with or without information on supplier relationship, respectively). The information gain (IG) of term $t$ is then defined as follows:

$$IG(t) = - \sum_{i=1}^{2} P(c_i) \log P(c_i) + P(t) \sum_{i=1}^{2} P(c_i|t) \log P(c_i|t) + P(\bar{t}) \sum_{i=1}^{2} P(c_i|\bar{t}) \log P(c_i|\bar{t})$$

Third, by using the selected $K_{SC}$ terms, a $tf-idf$ vector is calculated for each sentence. Let $D_t$ denotes the set of sentences, each of which contains two or more companies in the sentence, $tf_{ij}$ be the term frequency of term $i$ in sentence $j$, and $df_i$ be the number of sentences that term $i$ occurs in the training set. The formula of $tfidf_{ij}$ is defined as $tfidf_{ij} = tf_{ij} \times \log \frac{N(D_t)}{df_i}$. So, each sentence $j$ will be represented as $<tfidf_{1j}, tfidf_{2j}, \ldots, tfidf_{KSCj}>$. Finally, the SVM technique is applied on these training sentences to construct a sentence classification model (for classifying whether a sentence contains information about supplier relationship or not).

![Figure 1](image.png)  
*Figure 1. Overall process of the proposed SRM system.*
3.2 Direction Classification (Phase 2)

Based on the prediction result of sentence classification, we need to specify the role (i.e., supplier or customer) of each tagged company in the sentence. Let $com_i$ and $com_j$ are two tagged companies in a sentence that has been predicted as having supplier relationship by the sentence classification phase, the direction $\{com_i\} \rightarrow \{com_j\}$ means that $com_i$ is the supplier of $com_j$ (or $com_j$ is the customer of $com_i$). In the following, we use two examples of news sentences to explain in more details.

News sentence 1: “Finally, the Chinese character version of Chinanews.com reported this morning that Apple’s most important EMS provider, Hon Hai Precision, had an explosion at its Chengdu facility.”

News sentence 2: “Taiwanese-owned Foxconn is the world’s biggest contract electronics manufacturer with customers including Apple Inc., Sony Corp. and Hewlett-Packard Co.”

In the news sentence 1, it describes a supplier relationship that Apple Inc. has a supplier Hon Hai Precision. So, the direction of the sentence 1 is $\{Apple\} \rightarrow \{Hon Hai Precision\}$ defined as a right-side supplier direction. The news sentence 2 describes a supplier relationship that Foxconn is the supplier of Apple, Sony and HP. The direction in the sentence 2 is $\{Foxconn\} \rightarrow \{Apple Inc., Sony Corp., Hewlett-Packard Co.\}$ which is defined as a left-side supplier direction.

In order to construct the direction classification model, company names in a sentence are first separated into two groups according to some criteria (e.g., when company names are connected by a comma or a conjunction, we will put them into the same group). Second, the processes of feature extraction, feature selection and sentence representation are identical to those in the sentence classification phase. Note that, in information gain measurement, the classes of $c_1$ and $c_2$ in this direction classification are right-side and left-side supplier directions, respectively. And, in feature selection, we select top $K_{DC}$ terms to represent a sentence. Finally, the SVM is also applied to learn the supplier relationship direction prediction model, and direction type is predicted for each of supplier relationships. Therefore, a supplier/customer graph can be constructed according to these prediction results. In this directed graph, each node represents a company and each directed edge indicates the predicted direction type. And, the number on directed edges shows the number of sentences predicted this supplier/customer relationship.

3.3 Link Assessment (Phase 3)

The purpose of link assessment phase is to construct a classification model for refining whether the predicted direction of link is true or not by using the strength of directed edges and the relationships of customers’ customers and the suppliers’ suppliers of the focal company. In other words, if we want to refine that whether company $B$ is the supplier of company $A$ or not, i.e. $B \rightarrow A$, all of the predicted supplier and customer relationships of company $A$ and $B$ are taken into account in this link assessment phase. Specifically, we define 11 structural predictors (i.e., variables) to represent each directed link (e.g., $B \rightarrow A$ and company $A$ is the focal company) as follows:

- $V_1$: the number of sentences predicted as company $A$ having a supplier company $B$.
- $V_2$: the number of sentences predicted as company $A$ being a supplier of company $B$.
- $V_3$: $V_1 - V_2$.
- $V_4$: \( \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|} \), where $S(A)$ is the set of suppliers of company $A$ in the graph.
- $V_5$: \( \frac{\hat{S}(A) \cdot \hat{S}(B)}{\sqrt{\hat{S}(A)^2 \cdot \hat{S}(B)^2}} \), where $\hat{S}(A)$ is the vector of supplier links in the graph.
- $V_6$: \( \frac{|C(A) \cap C(B)|}{|C(A) \cup C(B)|} \), where $C(A)$ is the set of customers of company $A$ in the graph.
\[V_7: \frac{\bar{c}(A) \cdot \bar{c}(B)}{\sqrt{\bar{c}(A)^2 + \bar{c}(B)^2}}, \text{ where } \bar{c}(A) \text{ is the vector of customer links in the graph.}\]

\[V_8: \frac{|c(B) \cap c(S(A))|}{|c(B) \cup c(S(A))|} \sum_{x \in S(A)} \frac{|c(B) \cap c(c(x))|}{|c(B) \cup c(c(x))|} \]

\[V_9: \frac{|S(B) \cap S(C(A))|}{|S(B) \cup S(C(A))|} \sum_{x \in C(A)} \frac{|S(B) \cap S(c(x))|}{|S(B) \cup S(c(x))|} \]

\[V_{10}: \frac{|S(B) \cap S(C(A))|}{|S(B) \cup S(C(A))|} \sum_{x \in C(A)} \frac{|S(B) \cap S(c(x))|}{|S(B) \cup S(c(x))|} \]

Figure 2 shows some examples to illustrate the 11 variables defined above. For instance, for the focal company \(A\), we need to determine the 11 variables for the directed link \(B \rightarrow A\) (meaning that company \(B\) is a supplier of company \(A\)). \(V_1, V_2\), and \(V_3\) for \(B \rightarrow A\) are 12, 5, and 7, respectively. \(S(A)\) (i.e., the suppliers of \(A\)) is \(\{S_1, S_2, S_3\}\) and \(S(B)\) is \(\{S_1, S_3\}\) for \(B \rightarrow A\). Thus, \(S(A) \cup S(B) = \{S_1, S_2, S_3\}\), and \(S(A) \cap S(B) = \{S_1, S_3\}\). As the result, \(V_4\) for \(B \rightarrow A\) is \(2/3 = 0.667\). With respect to \(V_5\) for \(B \rightarrow A\), \(\hat{S}(A)\) and \(\hat{S}(B)\) are \(<2, 3, 6>\) and \(<3, 0, 4>\), respectively, and \(V_5\) is \(30/35 = 0.8571\). To calculate \(V_6\), \(C(A)\) (i.e., the customers of \(A\)) is \(\{C_1, C_2, C_3\}\), \(C(B)\) is \(\{C_1, C_2, C_3\}\), and \(V_6\) for \(B \rightarrow A\) is \(2/3 = 0.667\). For \(V_7\), \(\bar{c}(A)\) is \(<3, 4, 0>\), \(\bar{c}(B)\) is \(<2, 3, 6>\), and, as the result, \(V_7\) for \(B \rightarrow A\) is \(18/35 = 0.5143\). In addition, because \(C(B)\) is \(\{C_1, C_2, C_3\}\) and \(C(S(A))\) (i.e., the customers of \(A\)'s suppliers) is \(\{C_2, C_3, C_4\}\), \(V_8\) is \(1/2 = 0.5\) for \(B \rightarrow A\). For \(V_9\), \(\frac{c(B) \cap c(C_1)}{c(B) \cup c(C_1)}\) is calculated for each supplier \(x\) of \(A\), where \(x \in \{S_1, S_2, S_3\}\). So, \(\frac{c(B) \cap c(S_1)}{c(B) \cup c(S_1)}\), \(\frac{c(B) \cap c(S_2)}{c(B) \cup c(S_2)}\), and \(\frac{c(B) \cap c(S_3)}{c(B) \cup c(S_3)}\) are 1/4, 0, and 1/3, respectively. Accordingly, \(V_9\) for \(B \rightarrow A\) is \(7/36 = 0.1944\). Moreover, \(S(B)\) is \(\{S_1, S_3\}\) and \(S(C(A))\) (i.e., the suppliers of \(A\)'s customers) is \(\{S_1\}\). As the result, \(V_{10}\) is \(1/2 = 0.5\) for \(B \rightarrow A\). For every supplier \(x\) of \(A\), we calculate \(\frac{|S(B) \cap S(x)|}{|S(B) \cup S(x)|}\) and \(V_{11}\) for the link of \(B \rightarrow A\) then becomes \(1/4 = 0.25\).

![Figure 2. An example of the supplier/customer graph of a focal company A.](image)

In order to confirm whether links in a supplier/customer graph represents true supplier relationships, a supplier link classification model is constructed on the basis of a set of training instances covering
links of true supplier relationships and links of non-supplier relationships. Each of these training instances is represented by the 11 variables defined previously and is annotated as either in the true class (i.e., true supplier relationship) or in the false class (i.e., non-supplier relationship). Subsequently, the C4.5 algorithm (Quinlan 1993) is employed to learn a supplier link classification model (for link assessment purpose).

Balancing asymmetric data is also a critical problem for training a classification model. In study, the number of true instances is certainly much less than that of false instances. A bagging approach is used to balance the number of true and false instances for training classification model. Specifically, given a training dataset, we construct 31 training subsets, each of which contains all true instances (let the number of true instances be \( n_T \)) and \( \beta \times n_T \) randomly selected false instances. Subsequently, each training subset is used to induce a classification model. The 31 classification models will then be used via the voting mechanism to predict an outcome for a link. Finally, our proposed SRM system generates the refined supplier/customer graph for company \( A \) (i.e., the focal company).

4 EMPirical Evaluation

4.1 Data Collection

Our news documents were collected from Reuters (http://www.reuters.com/finance) from 2011/01/01 to 2011/12/31. We restricted news documents pertaining to the Financial Times Global 500 companies in 2011. As a result, our news corpus contains 60,445 news documents and a total of 4,737 companies are mentioned in these news documents. We randomly selected 3,000 news documents to construct a training set for sentence classification (phase 1) and direction classification (phase 2). Out of 54,425 sentences in these 3,000 news documents, 8,669 sentences contain two or more companies. To prepare a training set for these two phases, six business school students were recruited to manually tag these 8,669 sentences regarding whether or not they contain supplier relationships and, if so, who the roles of the companies play in each supplier relationship. Since the rare occurrence of supplier relationships in news documents, only 93 sentences were tagged with directional supplier relationships. For sentence classification, the 93 sentences serve as positive examples and we also need negative examples (i.e., sentences that do not describe supplier relationships). Specifically, for each of the 93 sentences tagged with supplier relationships, one sentence was chosen from the remaining 8,576 (= 8,669 - 93) sentences where the cosine similarity with the focal sentence is highest but the similarity is smaller than 0.8. Finally, 186 sentences containing 93 positive and 93 negative examples were selected as the training set of phase 1.

For the ground truth of the suppliers for companies, four companies have provided their supplier list on the Web, including Apple, Dell, HP, and Intel. There are 156, 131, 90, and 48 suppliers for Apple, Dell, HP, and Intel, respectively. We then filtered out the suppliers that are not mentioned in our news documents. Finally, our evaluation includes 44, 19, 21, and 10 suppliers for Apple, Dell, HP, and Intel, respectively. In this study, we employ the supplier lists of Dell, HP, and Intel to prepare a training dataset for the link assessment phase (i.e., phase 3) and use the supplier list of Apple for testing purpose.

4.2 Performance Benchmark and Tuning Result

The parameters of top \( K_{SC} \) and \( K_{DC} \) representative terms in the feature selection step of phase 1 and phase 2 are evaluated by using the manually tagged sentences. In the evaluation result, the parameter

---

settings of $K_{SC} = 40$ and $K_{DC} = 70$ reach the best performance with $F1 = 0.54$ (precision = 0.46 and recall = 0.66) and $F1 = 0.67$ (precision = 0.84 and recall = 0.56) for phase 1 and phase 2, respectively.

Because there does not exist any supplier mining technique in the literature, we thus develop a benchmark method as follows. Assume that the number of sentences predicted by phase 2 as $B \rightarrow A$ be $w$ and the number of sentences predicted by phase 2 as $A \rightarrow B$ is $z$. We define a link strength threshold $\alpha$. If $w - z \geq \alpha$, we predict that $B \rightarrow A$; if $z - w \geq \alpha$, we predict that $A \rightarrow B$; otherwise, we consider that there is no supplier relationship between $A$ and $B$. Figure 3 shows the tuning result of different thresholds for $\alpha$ for the focal company (i.e., Apple Inc). As threshold $\alpha$ increases, precision improves at the cost of recall. The highest F1 score is achieved when $\alpha$ is set to 4 for our benchmark method.

Figure 3. Tuning result of the benchmark method (on Apple Inc).

4.3 Comparative Evaluation Results

Using the supplier list of our testing company (i.e., Apple), we first evaluate the proposed SRM system by varying values for $\beta$ (i.e., the ratio between false instances and true instances) in the link assessment phase (i.e., phase 3). Specifically, we evaluate the range of values for $\beta$ from 1 to 15 in increments of 1. As Figure 4 illustrates, $\beta$ is equal to or larger than 5, the resultant precision reaches to 1.00. The best effectiveness attained by our proposed SRM system is precision = 1.00 and recall = 0.27 ($F1 = 0.43$) at $\beta = 5$ to 13 and 15. As Table 1 show, the best effectiveness of our proposed SRM system is better than that of the benchmark method (with precision = 0.19, recall = 0.18, and $F1 = 0.1849$). This comparative evaluation result suggests that our proposed SRM system outperforms the benchmark method in all evaluation metrics.
5 CONCLUSION

In this study, we proposed a novel supplier mining system, namely Supplier Relationship Mining (SRM) system, to identify supplier relationships extracted from financial news documents on Web, and construct the supplier/customer graph for a concerning focal company. The SRM system consists of three phases, including sentence classification, direction classification, and link assessment. We develop 11 variables for link assessment purposes. In the evaluation analysis, since supplier mining has not been investigated by prior research, a benchmark method is developed when evaluating our SRM method. The company Apple Inc. is used as the focal company in our evaluation. The results in Section 4 show that the proposed SRM system, achieving a precision of 100% and a recall of 27%, noticeably outperforms the benchmark method.

ACKNOWLEDGMENT

This work was supported by the National Science Council of the Republic of China under the grant NSC 100-2410-H-002 -021-MY3.

References


