2E3-5

企業のウェブ情報を用いた取引マッチング支援システムに関する研究

Predicting Customer-Supplier Relationships using Web Information

伊藤　諒 *1　　森　純一郎 *2
Ryo Ito　　Junichiro Mori

*1東京大学工学部システム創成学科　*2東京大学大学院工学系研究科技術経営戦略学専攻
Faculty of Engineering, The University of Tokyo　Graduate School of Engineering, The University of Tokyo

Analytical lens of the concept of business ecosystem has clarified strategies and behaviors of actors embedded in business ecosystem. It helped researchers to understand dynamics of systems emerged from intended and unintended results of actors 's behaviors. This paper illustrates recent achievement of research on business ecosystem especially focusing on computational approach. It not only allows analysis of business ecosystem but also supports design of it. We demonstrate and discuss the effectiveness of machine learning approach to model customer-supplier relationships, which can be utilized to design business continuous plan and to find plausible business partners.

1. はじめに

学術および産業において，ビジネスエコシステムが近年注目を集めている [1] [2]．エコシステムは，そこに内在するアクター間の競争や協調と言った相互作用を理解するための重要な視座を提供するものである．現代において，企業の競争優位性は，その企業が保有する技術やビジネスモデルのみならず，協業の戦略やビジネスエコシステムにおけるその企業の位置づけに大きく依存していると広く認識されている．そのため，企業の経営陣の意思決定において，ビジネスエコシステムの構造的な分析ならびに協業戦略の立案は，必要不可欠な要素となっている．そして，そのような意思決定を支援するための方法論やツールを構築することは，企業の技術経営において現在，重要な研究課題となっている．

ビジネスの戦略立案のために，企業は，エコシステムにおける自身の位置づけを適切に把握し，ビジネスチャンスを広げるための機会の発見を行う必要がある．その手段となるのが，大規模なデータを分析することで現状を客観的に俯瞰し，戦略立案に資する知識を抽出する計算知能のアプローチである．我々はこれまでに，国内の大規模なサプライチェーンにおける企業間の取引関係をモデル化することで，取引先の候補を推薦するシステムに関する研究を行ってきた [3]．企業間の取引関係で構成されるサプライチェーンは，ビジネスエコシステムの典型的な例である．企業が，そのサプチェーンにおいて適切な取引先の候補を発見できることは，ビジネスエコシステムにおける協業戦略の立案において非常に重要である．

特にリソースの乏しい中小企業などの従来の取引先発見や販路開拓の支援は，中小企業支援を行う専門家や専門機関が提供する人知に大きく依存しており，非常に手間のかかるものであった．一方，我々の手法は，大規模な既知の企業の属性情報や取引関係の情報を元に，機械的に潜在的な取引先候補の推薦を行うものであり，ビジネスエコシステムにおける企業の戦略立案の支援ツールとしての活用が期待できる．しかしながら，従来手法 [3] においては，特に企業の既知の取引情報に依存しており，それらの情報が入手できない企業については適用が困難であるという問題が存在する．企業の取引情報は，企業の信用調査の一環として一部の企業データベースの提供者が保有しているが，それらの情報を網羅的に収集することは困難である．一方で，企業の保有技術や製品と言った情報については各企業のウェブページにテキスト情報として網羅されており，かつそれらは容易に収集可能である．

そこで本研究では，個々の企業が最適な取引先候補を発見することを支援することを目的とし，ウェブから収集した企業のテキスト情報を用いて企業間の取引関係を予測するモデルの構築，ならびにそのモデルの精度評価について考察を行う．さらに構築されたモデルを用いて，企業の取引先を推薦する Web システムの実装を行う．

2. 関連研究

産業構造が細分化ならび複雑化している現在，産業の川上から川下へという従来の直線的なサプライチェーンに対して，ネットワーク構造としてのサプライチェーンに注目が集まっており，サプラチェーンの構造と企業活動に関する定量的な分析が行われている [4][5][6]．特に近年においては，エコシステムとしてのサプライチェーンにおいて，データ分析に基いた，ビジネスの戦略立案を支援するための研究が行われいている [7][8]．

さらに，SVM やニューラルネットワークなどの機械学習を応用した取引先候補の推薦に関する研究が盛んに行われている [9][10]．森らは，企業を表す特徴量として資本金，従業員数，所在地などの企業の基本属性情報と，個々の取引関係からなるサプライチェーンネットワーク構造情報を用いて，企業間の取引関係を予測する手法を提案している [3]．その研究において評価指標となる F 値は 0.77 程度であり，企業属性情報を用いた取引関係の予測は有用であるという事が示された．

従来の研究において，企業を表す特徴量として企業の基本属性情報やサプライチェーンネットワーク構造情報を用いており，個々の企業の具体的な活動，製品，技術などが反映されていなかった．しかしながら，取引先候補の推薦を行う場合は，企業の具体的な活動に関連した根拠の提示が重要であり，個々の企業の具体的な活動，製品，技術などが反映されていない従来手法では，この根拠の提示が困難であるという問題点があった．一方，企業のウェブページには個々の企業の具体的な活動，製品，技術などを表すテキストが存在する．そこで本研究では，企業を表す情報としてテキスト情報を用いて，取引関係を予測する手法の提案を行う．

3. テキスト情報を用いた企業間取引関係のモデル化

まず，その取引関係が既知の企業群について，それらの企業の公式ウェブページを収集する．収集にあたっては，当該ウェブページのトップドメインについて，検索エンジンの ”site: ” オプションにより検索を行い，検索エンジンによるインデキシングが確認されたウェブページを対象として，ウェブクローラにより収集を行う．次に，収集された各企業のウェブページ群から，その内容を特徴付けるキーワードを抽出する．キーワードの抽出にあたっては，まずウェブページから html タグを除去した上で，形態素解析を行い，名詞ならび未知語を抽出をする．抽出された名詞ならび未知語を対象にコロケーション抽出の手法である C-value 法によりさらに名詞句を抽出し，これらを各企業のキーワード群とする．また出現頻度が相対的に少ない単語や多い単語についてはそれぞれ，単なる記号列や誤表記された単語などのノイズであったり，産業や技術に直接関連のない一般語のため，企業を表す特徴とするには不適切である．よって単語出現頻度に下限と上限の閾値を設定し, 下限閾値未満もしく上限閾値を超える単語を除去する．そして各企業のウェブページから抽出したキーワード群を用いて，各企業を 1 つの文書と見なした以下の式で表される単語文書行列 M を作成する．

$$M = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix}$$

行列 M の各成分 x_{ij} は単語 i の文書 j における出現重みを表しており，出現重み x_{ij} は，以下の 3 つの異なる手法により重み付けを行う．

- 単語が出現するか否かの 1 か 0 の Binary 重み
- 単語の出現頻度（TF）重み
- 単語の TF-IDF 重み

まず，2 つの企業 X_s と X_c のペアを表すベクトルを $X_{sc} := (X_s, X_c)$ とする．ここで，s はサプライヤー企業を，c はカスタマー企業を表しており，X_s と X_c は行列 M における，企業 s と c に対応するそれぞれの行ベクトルである．この企業ペア X_{sc} について，取引関係が存在するか否かを予測するモデルの構築を行う．$Y := y$ を取引関係を表す変数とし，y は，取引があれば $+1$ を，取引がなければ -1 をとるものとする．ここで目的は，企業ペア X_{sc} から取引関係 Y を予測する，関数 $Y = F(X_{sc})$ で表されるようなモデルを推定することである．この時，関数 F は，サプライヤーとカスタマーの企業ペアを表す重み付きのキーワードベクトルを入力とし，それらの企業間の取引関係の有無を予測するような関数である．そのような関数 F について，企業間の取引関係はサプライヤーとカスタマーの各企業を表すキーワードの特徴的な組み合わせによって表されることを仮定する．

そこで，企業ペア X_{sc} を構成するベクトル X_s と X_c の要素積を取り，以下の式で表されるサプライヤー企業とカスタマー企業を表す任意の単語同士の組み合わせ特徴量 x を作成する．そして，以下のように関数 F をそれらの組み合わせ特徴量の重み付き線形モデル（1），もしくは重み付き非線形モデル（2）で表されるものとする．

$$Y = sign(wx + b) \tag{1}$$

$$Y = sign(w\Phi(x) + b) \tag{2}$$

ここで，$x = (x_{s1}x_{c1}, ..., x_{si}x_{cj}, ..., x_{sn}x_{cn})$ は，サプライヤー企業を表すベクトル X_s の i 番目の要素とカスタマー企業を表すベクトル X_c の j 番目の要素の積 $x_{si}x_{cj}$ からなる $n \times n$ 次元の特徴量ベクトルであり，$w = (w_{11}, ..., w_{ij}, ..., w_{nn})$ はモデルを表すためのパラメータベクトルである．また sign はシグモイド関数であり，$\Phi(\cdot)$ は任意の関数である．

n は，行列 M の単語数であり，一般に数万から数十万である．そのため，特徴量ベクトル x の次元数は膨大な数となり，モデルの表現力が増加し過ぎることでモデルの汎化性が低下することや，スパースな行列が得られるという問題点がある．さらには計算の上で多くメモリを必要とする問題が生じる．ここで企業間取引が形成される要因を考えた際,「ねじとプレス」といった単語レベルの関係性よりも，より抽象的な「部品製造と部品組立」といった産業や技術などの概念の関係性が取引形成に影響を及ぼしていると考えられ，潜在的な変数の導入が必要であるとする．

以上の事から，潜在的な変数を得るために高次元のデータを低次元に変換する次元削減手法である LSI を用い行列 M の次元数の削減を行った後，得られたベクトル X_s^*，X_c^* に基づいて要素積を取り，特徴量ベクトル x^* を生成する．

関数 F のようなモデルの最適なパラメータベクトル w の推定にあたり，線形 SVM とニューラルネットワークを用いる．

4. 実験

4.1 データ

帝国データバンクが提供している企業概要データベース [*1] のデータを用いて，当該企業の取引関係が既知でかつ公式ウェブページが確認された関東甲信越地域の製造業企業を対象とした．なお，帝国データバンクのデータには URL 情報が含まれていないため，クラウドソーシング [*2] を用いて企業の公式ウェブページの URL 情報の収集を行った．クラウドソーシングのワーカーには企業名と所在地から企業の公式ウェブページの URL を探すタスクを依頼し，各企業に割り当てられた 2 名のワーカーが発見した URL が共通していれば当該企業の公式ウェブページの URL 採用することとした．その結果，7,112 企業の公式ウェブページを取得した．各企業のウェブページから合わせて 657,567 語のキーワードを抽出した．

取引関係データには，帝国データバンクの信用調査において各企業が自己申告した取引先のサプライヤーならびカスタマーが最大 5 社まで含まれている．これらのデータをもとにして，取引関係がある 5,039 の企業ペアを抽出し，提案手法を用いた取引関係予測モデルを構築するための学習データの正例とした．一方，取引関係のない同数の企業ペアを負例サンプリングとしてランダムに 3 セット抽出し同学習データの負例とした．このようにして正例と負例のペアを 3 セット作成し，提案手法による取引予測モデルを作成し，5-folds のクロスバリデーションにより予測精度の評価を行った．予測精度の評価指標は F 値とし，各正例と負例のペアの予測精度を平均した値を，その学習モデルの予測精度とした．

4.2 単語重みの評価

まず，テキスト情報を用いた企業間の取引関係予測のベースラインを示す為に，次元削減および組み合わせ特徴量を用いずに，学習データの正例，負例の企業ペアの重み付きキーワードベクトルの要素をそのまま特徴量として用いて取引関係予測モデルを構築した際の予測精度の評価を行った．単語の重みづ

∗1 http://www.tdb.co.jp/lineup/cosmos2/
∗2 http://www.lancers.jp/

けには，単語が出現するか否かを 1 または 0 で表した Binary 重み，TF 重み，TF-IDF 重みの三種類を用いた．学習モデルには L2 正則化の線形 SVM を用いた．

4.3 次元削減と組み合わせ特徴量の評価

次に，予測精度改善の為に，単語文書行列を次元削減した後に，各トピック同士の組み合わせ特徴量を作成し，予測精度の評価を行った．単語文書行列の重みの種類は Binary 重みと TF-IDF の二種類を用い，LSI による次元削減後の次元数は 100 次元とした．学習モデルは L2 正則化の線形 SVM を用い，各々の単語重みの場合の予測精度の比較を行った．

4.4 取引関係予測モデルの評価

次に，学習モデルとして L2 正則化の線形 SVM を用いて，各種パラメーターによる予測精度の違いを検証する為に，Binary 重みで重み付けした単語文書行列を次元削減した後に，各トピック同士の組み合わせ特徴量を作成し，予測精度の評価を行った．パラメーターは正則化項，単語出現頻度の下限と上限，トピック数である．正則化項のコストパラメーターは 1,10,100,1000,10000 で変化させた．単語出現頻度の下限は 10 から 10 ずつ増やした 100 までの 10 通り，上限は 300 から 100 ずつ増やした 1200 までの 10 通りとした．また次元削減後の次元数は 100 と 200 の 2 種類とし，以上合計 200 通りのデータセットを作成し学習モデルの入力とした．

4.5 取引関係予測モデルの拡張

非線形モデルであるニューラルネットワークを学習モデルとして分析を行った．実験の条件は線形 SVM におけるパラメーター評価時の条件と同じく，単語文書行列の重みの種類は Binary 重み，次元削減後の次元数は 100 と 200 の 2 種類，単語出現頻度の下限は 10 から 10 ずつ増やした 100 までの 10 通り，上限は 300 から 100 ずつ増やした 1200 までの 10 通り，合計 200 通りのデータセットを用いた．また中間層が 1 層のニューラルネットワークを用い，中間層・出力層の次元はそれぞれ 1024，2 とし，活性化関数は順に ReLU，Softmax を用いた．また，正則化の強さを表す weight decay は 0.001 とし，確率的勾配降下法（SGD）によりパラメーターの更新を行った．さらに学習エポック数は 30 として実験を行った．

5. 結果と考察

5.1 単語重みの評価

表 1 は単語重みとして，Binay 重み，TF 重み，TF-IDF 重みの 3 種類の重みと，それぞれの推定精度を示したものである．予測精度が良かった順に，TF-IDF 重み，Binary 重み，TF 重みであったが，ここから単語の出現頻度を表す TF 値よりも，単語の文書頻度を表す DF 値の方が，取引予測精度に寄与していることが分かった．これは一般語と比べて企業の産業・技術・製品を表す産業用語は多くの文書には出現せず，DF 値が小さいためだと考えられる．また同様に TF 重みよりも Binary 重みの方が，取引予測精度に寄与していることが分かった．これは TF 重みの場合，出現頻度の多い一般語が産業用語よりも大きな重みを持っているが，Binary 重みを用いる事で一般語と産業用語が同一の重みで扱われ，TF 重みの場合よりも相対的に産業用語の重要性が増したためと考えられる．

表 1: 単語文書行列の重みと取引予測精度

重み	Binary	TF	TF-IDF
F 値	0.688	0.674	0.700

5.2 次元削減と組み合わせ特徴量の評価

表 2 は，Binary 重みと TF-IDF 重みの 2 種類の重みからなる単語文書行列を次元削減し，組み合わせ特徴量を生成した際の予測精度の比較をした表である．単語重みの評価での実験では TF-IDF を用いた時に F 値 0.700 で最大であったが，今回の実験では Binary 重みを用いた際に F 値 0.748 とさらに予測精度が改善していることから，次元削減をした後に組み合わせ特徴量を用いるという提案手法の有効性が確認された．このことはサプライヤー企業とカスタマー企業を表す産業や技術などの概念の潜在的な関連性が，取引関係予測に寄与していることを示唆している．Binary 重みを用いた場合，次元削減により得られたトピックは表 3 のように纏まりのあるものとなった．一方，TF-IDF 重みを用いた際は，記号文字や限定的な単語から成るトピックが抽出された．このことを踏まえ，これらの単語によるノイズを減らしより有効な特徴量を得るために，単語の出現頻度に対して，上限と下限の閾値を設け，条件を満たさない単語を除くフィルタリングを行う必要があると考えられる．よって次の実験では低頻度語と高頻度語に閾値を設けて実験を行った．

表 2: 単語文書行列の重みと次元削減後組み合わせ特徴量による取引予測精度

重み	Binary	TF-IDF
F 値	0.748	0.662

表 3: 次元削減後のトピックの例（単語重み Binary，次元数 100）

加工	工程	樹脂	プラスチック	素材	ロット	金属
金型	勤務	CAD	台数	履歴書	CAM	勤務地
島根	徳島	和歌山	北海道	山形	奈良県	
建物	工事	社	建築	施工	昭和	道路	トップページ
施工	本体	建物	空間	幅	設置	構造	タイプ	壁
環境マネジメントシステム	表紙	廃棄物	環境方針			
金型	検査	医療	銅	量産	金属	材料	試作	損害

5.3 取引関係予測モデルの評価

表 4 は最適パラメーター探索の有無による取引予測精度の違いをまとめたものである．パラメーター探索の結果予測精度が向上し，最大予測精度はトピック数 200，低頻度語の閾値 10，高頻度語の閾値 1000，正則化項の強さ 10000 の時に 0.769 となった．この値は森らの先行研究 [3] の 0.777 とほぼ同程度の精度であり，従業員数や企業ランキングなどの企業属性情報では無くテキスト情報を用いた場合でも，同程度の予測精度を得られる事が実験により示された．

表 4: 最適パラメーター探索の有無と取引予測精度

重み	探索無し	探索有り
F 値	0.748	0.769

5.4 取引関係予測モデルの拡張

ここでは取引推薦モデルを拡張して，非線形モデルであるニューラルネットワークを学習モデルとして分析を行った．図 5 は学習モデルとしてニューラルネットワークを用いた場合と線形 SVM を用いた場合のそれぞれの最大予測精度を比較したものである．学習モデルとしてニューラルネットワークを用い，次元削減後のトピック数 200，単語出現頻度の下限閾値 30，上限閾値 1200 とした時に，予測精度が F 値 0.790 となり本研究において最大の予測精度となった．線形 SVM よりも

ニューラルネットワークの方が予測精度が良かった理由としては，企業ペアの正例と負例空間上に複雑な識別境界面を形成しており，非線形分離が可能なニューラルネットワークがその識別境界面をよりよく学習出来た為と考えられる．なお SVM はカーネル関数の種類によって非線形分離が可能であるため，非線形 SVM を用いることでも精度が向上する可能性があると考えられる．この点の検証は今後の課題である．

表 5: 各学習モデルにおける最大予測精度

学習モデル	線形 SVM	ニューラルネットワーク
F 値	0.769	0.790

6. 取引マッチング支援システムの実装

まず，ユーザーは対象となる企業のホームページ URL をホーム画面の URL 入力ボックスに入力をする．学習済みモデルを用いた結果，潜在的に取引があると予測された企業を対象企業の取引先候補としてグラフの形で可視化する（図 1）．

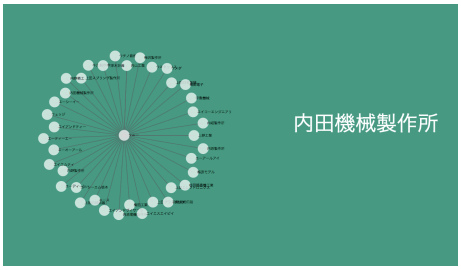

図 1: 企業間取引推薦システムの結果画面

取引マッチングシステムを用いた結果，どのような企業が推薦されるかを具体的な企業 URL を入力し検証した．表 6 は，山本精機 *3 の URL を作成したシステムの入力とし，その結果推薦された企業の産業種別を割合で多く含まれた上位 5 件を表示したものである．

山本精機は精密金型製作技術をベースとした企業であり，金属プレス加工，両面テープ・光学フィルムの打ち抜き加工，車載・医療機器部品の組立を手がける，長野県に北安曇郡に所在するものづくりの企業である．取引先としては，カスタマー企業として車両部品メーカー，各種機器製造メーカーが，サプライヤー企業として金属プレス加工機器を製造するメーカーが考えられるが，推薦の結果は仮説の通り，自動車部品製造，金属プレス製品製造などの企業が推薦結果に含まれていた．このように推薦の内容にある程度の合理性が見られた．

表 6: 山本精機の取引先候補の産業分類

産業種別	占有率
自動車部品製造	7%
荷役運搬設備製造	5%
機械同部品製造修理	5%
プリント回路製造	4%
金型・同部品製造	4%

7. まとめ

近年地域創成が行政における重要な課題となっているが，地域創成を実現する上では，地域産業の活性化が必要であり，個々の企業が最適な取引先を選定する事が，経営リスクの分散や新規技術，市場の獲得という観点からも重要である．本研究では，個々の企業が最適な取引先候補を発見することを支援することを目的とし，ウェブから収集した企業のテキスト情報を用いて企業間の取引関係を予測するモデルの構築，ならびにそのモデルの精度評価や企業間取引の形成要因について考察を行った．

企業を表す特徴量としてテキスト情報を用い，学習モデルに線形 SVM を用いた場合において，F 値 0.769 となり先行研究と同程度の予測精度が実現可能であることを，実験により示した．特に，次元削減によって作成された組み合わせ特徴量や単語の出現頻度による一般語のフィルタリングが予測精度向上に寄与することを示した．さらに本研究では学習モデルとしてニューラルネットワークを用いることで，F 値 0.790 と先行研究を上回る予測精度を得られることを実験により示した．さらにこれらの知見を生かし，潜在的な取引先候補発見の支援を行う取引マッチングシステムの実装を行い，推薦結果に一定の合理性があることを確認した．

今後の展望としては，さらなる予測精度の向上の他に，推薦内容の多様性を考慮することで，多様な取引先候補を提案する手法を本研究に統合する事が考えられる．

*3 http://www.yamamotoseiki.co.jp/

参考文献

[1] R. Adner and R. Kapoor, “ Value creation in innovation ecosystems: How the structure of technological interdepen- dence affects company performance in new technology gen- erations, ”Strategic Management Journal, vol.31, pp.306-333, 2010.

[2] R. Kapoor and J.M. Lee, “ Coordinating and competing in ecosystems: How organizational forms shape new technol- ogy investments, ”Strategic Management Journal, vol.34, pp.274-296, 2013.

[3] J. Mori, Y. Kajikawa, H. Kashima, and I. Sakata, “Machine learning approach for finding business partners and building reciprocal relationships”, Expert Systems with Applications, vol.32, pp.10402-10407, 2012.

[4] A. Cox, J. Sanderson, and G. Watson, “ Supply chains and power regimes: Toward an analytic framework for manag- ing extended networks of buyer and supplier relationships, ”Journal of Supply Chain Management, vol.37, no.1, pp.28-35, 2001.

[5] S.P. Borgatti and X. Li, “ On social network analysis in a supply chain context, ” Journal of Supply Chain Management, vol.45, no.2, pp.5-22, 2009.

[6] M.A. Bellamy, S. Ghosh, and M. Hora, “ The influence of supply network structure on firm innovation performance, ” Journal of Operations Management, vol.32, no.6, pp.357-373, 2014.

[7] R. Carbonneau, K. Laframboise, and R. Vahidov, “ Application of machine learning techniques for supply chain demand forecasting, ” European Journal of Operational Re- search, vol.184, no.3, pp.1140-1154, 2007.

[8] S.Y. Chou and Y.H. Chang, “ A decision support system for supplier selection based on a strategy-aligned fuzzy smart approach, ” Expert Systems with Applications, vol.34, no.4, pp.2241-2253, 2008.

[9] G. Hu and G. Zhang, “ Comparison on neural networks and support vector machines in suppliers ’selection, ”Journal of Systems Engineering and Electronics, vol.19, no.2, pp.316-320, 2008.

[10] X. Guo, Z. Yuan, and B. Tian, “ Supplier selection based on hierarchical potential support vector machine, ” Expert Systems with Applications, vol.36, no.3, pp.6978-6985, 2009.