4M1-2

# 日米スタートアップのキーワードによるクラスタリングを用いた事業トレンド予測

今井響 *1　大知正直 *2　松尾豊 *2
Imai Hibiki　Ochi Masanao　Matsuo Yutaka

*1東京大学工学部システム創成学科
Department of Systems Innovation, Faculty of Engineering, The University of Tokyo

*2東京大学工学系研究科技術経営戦略学専攻
Graduate School of Technology Management for Innovation, The University of Tokyo

The attention to the importance of startups has increased in Japan, but there is no confirmed method to choose which market. We focused on the globalization of business, and predict that which market will become hot in Japan by comparing startups' trends in Silicon Valley and in Japan. To analyze startups' trends, we used the data in AngelList and entrepedia, which are the biggest startup databases in Silicon Valley and Japan, and clustering startups by descriptions of startups' service.

## 1. はじめに

近年，国の政策にも現れているように，日本国内において起業の重要性が注目されてきている．一方で，有望な事業を選択する手法に関して確立された手法は存在せず，日本のスタートアップが事業を選択する方法にはまだ改善の余地がある．事業選択の方法の一つとして，ネット黎明期に注目された，米国の成功サービスを模倣して国内に持ち込むタイムマシン経営 [Arora 04] が挙げられるが，近年ではネットの普及による情報共有速度の向上 [Quelch 96]，国際間でのサービスの垣根の低下 [Lituchy 00] などにより，かつてのタイムマシン経営は難しくなってきているといわれている．本研究では，日米それぞれにおいて最大級のスタートアップデータベースであるentrepedia*1 及び AngelList*2 のデータを用いて，前半において米国で Exit したものと同様の事業が日本にて行われるまでの時差の縮小を示し，ネット黎明期におけるタイムマシン経営が現在においては難しくなっていることを確認した．また，米国で企業が Exit してから同様の事業が行われるまでの時差が短縮してきているだけでなく，米国での Exit を待たずして日本で同様の事業が行われるようになってきていることがわかった．そのため本研究の後半部では，有望な事業か否かを判断する指標を，米国でその事業を行う会社が Exit したというものでなく，その事業を行う会社の数が多い，とすることで，日本における有望な事業の予測を試みた．結果，機械学習による日米の企業群のクラスタリングと予測を用いることで，従来手法に比べて高い精度で，日本で今後有望な事業を予測することに成功した．

## 2. 関連研究

本章では，関連研究としてスタートアップの成功予測に関する研究と，スタートアップの分類に関する研究に触れる．

連絡先: 今井響，東京大学工学部，東京都板橋区成増 3-31-4-108，08030947955，0h3k2a4@gmail.com

∗1 http://entrepedia.jp/

∗2 https://angel.co/

### 2.1 スタートアップの成功予測に関する研究

ベンチャー企業の成功要因に関しては Key Success Factor[He 09]（以下 KSF として扱う）と呼ばれるものがある．ベンチャーの成功予測に関する既存研究の多くは，ベンチャー企業が成功に至るまでには多数の因子が複雑に関連している中にも，再現性のある KSF が存在するという前提 [Chang 04] に立ち，行われてきた．

ベンチャー企業を評価する指標としては，古典的なものに Resource Based View[He 09] がある．企業の成功要因は社内資源の最適配置にあるという立場である

社内要因だけでなく，社外の要因に関しても言及しているものとしては，Network Resource Combinations[Tolstoy 10] というアプローチがある．これは，対象とする企業の活動，内部要因だけに着目するのではなく，顧客や投資家，取引先といった外部の要因も視野に入れて KSF を解明するというものである．特に，人的な関係性という因子が成功に対する影響があることは，協業 [Yli-Renko 02] や投資家の持つネットワーク [He 09] の重要性が示されている．そのため，ベンチャー企業の人的な資産という意味で総じてソーシャルキャピタル [Yli-Renko 02] という言葉で語られている．関連研究では，ベンチャー企業の人材転職履歴情報と成功の因果関係 [上野山 14] が示されている．

### 2.2 スタートアップの分類に関する研究

スタートアップの成立要因や成功要因，分布等を議論する上で，事業の分類は以前から行われてきたテーマである．ベンチャーキャピタルがスタートアップを評価する際に用いている分類手法として，リスクを起点とした分類 [MacMilan 86] を用いているとの研究が 1985 年に行われた．

また，成功するか否かを発見するための分類軸を模索した結果，競合といかに断絶されているか，及び事前のデモンストレーションでどれほど市場に受け入れられたかが重要であるとする研究 [MacMilan 87] がある．2000 年以降，スタートアップの成長段階と資金調達のラウンドによる分類に触れている研究 [Gompers 01] も行われるようになった．この研究では，ベンチャーキャピタルが早期のベンチャーに対して集中的に投資を行っている，と述べている．

本研究では，特定のスタートアップに対しての成功するか否

かの予測や，成功要因の特定を行うのではなく，日米という二国間の事業の流行に関する関係性を元に，有望な事業の予測を行う．また，近年増加してきた Web 上のスタートアップに関する情報を用いて，日米それぞれに関して事業内容にもとづいたスタートアップの分類をし，比較分析に用いる点も新規性のある点となる．

## 3. 日米スタートアップの比較分析

本章では，日米のスタートアップの比較分析を行い，米国で Exit した事業と同様のものが日本において行われるまでの時差が短くなってきていることを確認する．そのために，日米のスタートアップをそれぞれ，事業内容に基づいてクラスタリングする．分析に使うデータは，米国及び日本においてそれぞれ最大級のスタートアップデータベースである AngelList 及び entrepedia のものを用いる．対象とするデータ数は，米国はシリコンバレー発のスタートアップ 15,876 社，日本は 2000 年以降に設立されたスタートアップ 5,949 社分とする．なお，AngelList と entrepedia には，スタートアップごとに会社名，サービス説明概要文，過去の資金調達時期と額，その際の投資家の情報や，創業者に関する情報などがある．クラスタリングは，会社説明文からキーワードを抽出して会社間の類似度を計算し，ネットワーク図を作成したうえで、Newman 法によりモジュラリティを計算して行う．キーワード抽出には tf-idf 法を，会社間の類似度計算にはコサイン類似度を用いる．抽出されたクラスタには，クラスタ内の会社説明文における出現頻度の高い語を用いてラベル付けを行う．2012〜2014 年における米国スタートアップをクラスタリングした結果をネットワーク図として描画して，ノード数の多いクラスタのラベルを注釈としてつけたものを次の図 1 に示す．

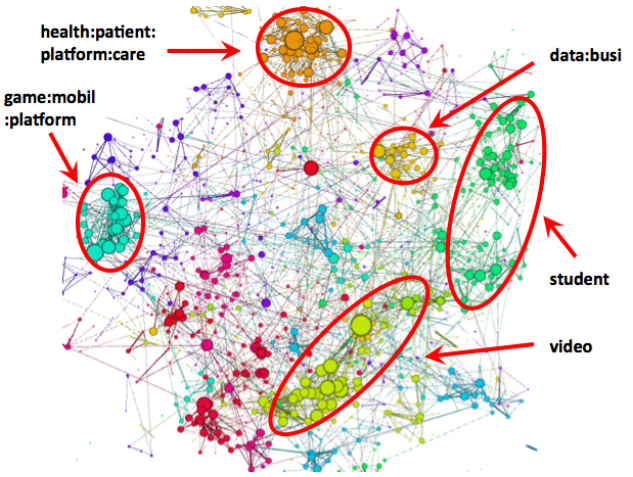


図 1: 2012〜2014 年における米国スタートアップの事業内容に基づくネットワーク図

ノード数が多いクラスタには，「video」「health:patient:platform:care」「student」「data:busi」「game:mobil:platform」などのラベルがついたものがあった．日米スタートアップの比較分析は，米国において Exit した会社を含むクラスタのうち，日本に対応する事業を行うクラスタが存在するものと，対応する日本のクラスタの比較により行う．米国の Exit した会社を含むクラスタと，対応する日本のクラスタを以下の表 1 及び表 2 に示す．

表 1: 日本に同様クラスタが存在する，米国の Exit 企業を含むクラスタ

| No. | Date | Cos | Label | Node Num |
|---|---|---|---|---|
| 1 | 03〜05 | 0.1 | data:custom:big | 6 |
| 2 | 03〜05 | 0.1 | softwar:cloud:solution:busi:applic:center:integr | 5 |
| 3 | 06〜08 | 0.1 | data:cloud | 31 |
| 4 | 06〜08 | 0.1 | onlin:game | 20 |
| 5 | 06〜08 | 0.1 | cisco:network:cloud:leader:security:secur | 3 |
| 6 | 06〜08 | 0.2 | media:platform:social:optim:manag:engag:dashboard | 7 |
| 7 | 09〜11 | 0.25 | big:data:analyt:intellig:transform:analytics:busi | 4 |
| 8 | 12〜14 | 0.2 | game:mobil:platform:develop:social:play:player | 25 |
| 9 | 12〜14 | 0.3 | payment:card:credit:servic:api:pay:let | 4 |

表 2: 米国 Exit クラスタに対応する日本のクラスタ

| 対応 No. | Date | Cos | Label | Node Num |
|---|---|---|---|---|
| 1, 7 | 12〜14 | 0.5 | データ:ビッグ:プラットフォーム | 5 |
| 2 | 09〜11 | 0.2 | クラ:ウド:システム | 10 |
| 3 | 12〜14 | 0.2 | クラ:ウド:データ | 19 |
| 4 | 06〜08 | 0.5 | 運用:インターネット:ソフトウェア:ゲーム:オンライン | 4 |
| 5 | 12〜14 | 0.3 | クラ:ウド:管理:保守:システム | 13 |
| 6 | 09〜11 | 0.4 | ソーシャルメディア:コンサルティング:Twitter:Facebook | 6 |
| 6 | 09〜11 | 0.4 | ソーシャルメディアマーケティング:コンサルティング | 3 |
| 8 | 12〜14 | 0.5 | フォン:スマート:ゲーム | 7 |
| 9 | 12〜14 | 0.2 | 決済 | 7 |

上記の日米の対応するクラスタごとに，米国においては初回資金調達を行った年度，Exit した年度，日本においては会社が設立された年度を指標とし，クラスタ内で最も早期の年度を用いて時差に関する比較を行う．なお，米国で初回資金調達を行った年度は，クラスタの表す事業が Exit ほど明確な成功を収めずとも注目され始めた時期を表す指標として用いる．比較を行った結果は下記の図 2 となった．この結果，米国の Exit から同様の事業が日本において行われるまでの時差が短くなっていることがわかった．それだけでなく，近年では Exit 前から日本においても同様の事業が行われ始めていることがわかった．

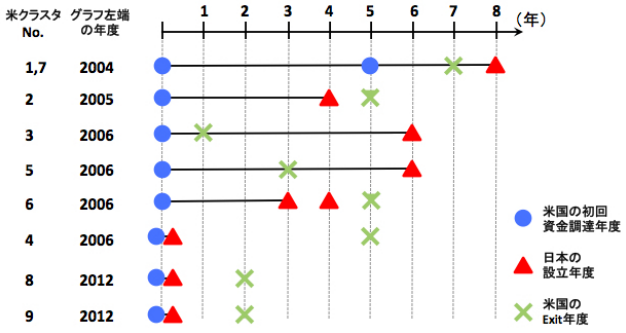


図 2: 日米の対応クラスタに関する時差比較

なお，従来のタイムマシン経営は，ここで挙げられてる米国の Exit した企業を含むクラスタが示す事業を日本に持ち込む手法であると考えることができる．従来の手法を現在に適用しようとしても成功することは難しいことがわかる．この結果を受けて本研究では次章以降，日本において有望な事業か否かの指標を，米国で Exit した企業がある，というものから，米国で同様の事業を行っている企業が多い，つまり事業が流行している，注目されている，というものに変えることで日本においてい有望な事業の予測を試みる．

# 4. 提案手法

本章では，日本において有望な事業の予測手法を提案する．今回提案する手法は，大きく以下の２つにわかれる．

1. 米国の企業群を，事業に関する説明文を元にクラスタリングする
2. 各クラスタを１つの事業分野と捉え，それが今後日本で成功するかを予測する

## 4.1 米国の企業群のクラスタリング

米国の企業群を，事業に関する説明文を元にクラスタリングする流れを示す．

始めに，tf - idf 法により各企業の説明文からキーワードを抽出する．次に，各企業間の類似度を，キーワードを用いたコサイン類似度により算出する．続いて，コサイン類似度に下限を設け，下限をこえるコサイン類似度を持つ企業間にエッジを持たせたネットワーク図を作成する．最後に，ネットワーク図を元に，モジュラリティを用いたクラスタリング手法である Newman 法を用いてクラスタリングを行う．モジュラリティとは，クラスタ内部のエッジ数が，クラスタ間のエッジ数と比べて多いほど高い値となる．Newman 法による計算式は以下の式（1）となる．

$$\varrho = \sum_i \left( e_{ii} - \left( \sum_j e_{ij} \right)^2 \right) = \sum_i (e_{ii} - a_i^2) \qquad (1)$$

表 3: 米国 Exit クラスタに対応する日本のクラスタ

| 変数 | 意味 |
|---|---|
| $e_{ij}$ | 「総エッジ本数」に対する、コミュニティ i から j に張られているエッジ本数の割合 |
| $a_i$ | 「総エッジ本数」に対する、コミニティ i から張られているエッジ本数の割合 |

以上により，米国企業群をクラスタリングする．続いて，各クラスタを１つの事業分野と捉え，どの事業分野が日本において有望かの予測を行う．

## 4.2 今後日本で成功する事業分野の予測

米国企業群のうち，どのクラスタが日本において成功するかの予測を行う．今回，日本においての成功の定義は，特定の事業分野を扱う企業の数が多くなることとする．予測の方法は，まずクラスタ内の企業の数が多いものを抽出し，抽出したクラスタに対して機械学習を用いて予測を行う．

予測は，機械学習を用いた 2 値分類器の作成，適用により行う．分類器は，2 値分類の際によく用いられるサポートベクターマシン（以下 SVM）を使う．SVM に用いる素性は，1. 米国の対象とする期間における資金調達総額に占める，対象クラスタ内のスタートアップの資金調達総額，2. 対象クラスタ内のスタートアップの，対象とする期間における資金調達額の増加率，3. 対象クラスタ内のスタートアップの，対象とする期間における会社数の増加率を用いる．米国で流行しているクラスタのうち，同年代において他の事業に比べて相対的に注目されているほど，また過去にくらべて対象の期間に急激に成長しているほど日本において流行する可能性が高いと考え，1. の素性は前者を，2. 及び 3. の素性は後者を表すと考えての設定である．

従来手法は，米国企業群をクラスタリングした結果から，クラスタ内の企業の数や機械学習による予測を行わず，Exit した企業を含むクラスタ（事業分野）を日本において有望な事業とするものと捉えられる．

# 5. 実験・結果

本章では，実験方法及び結果について述べる．

## 5.1 実験方法

実験方法については，実験に用いるデータ，実験方法及び条件，評価方法の順に述べる．

実験に用いるデータは，今回 entrepedia 及び AngelList から取得したデータは，2000 年以降に設立された日本のスタートアップ日本のスタートアップ 5,949 社，及び米国のうちシリコンバレー発のスタートアップ 15,876 社である．

実験は，以下の流れで行う．

1. entrepedia 及び AngelList からのデータの取得
2. サービス説明文からの tf-idf 法によるキーワードの抽出
3. キーワードを用いた会社間のコサイン類似度の算出
4. コサイン類似度を用いたスタートアップのクラスタリング
5. 米国において流行しているクラスタの選択
6. SVM による予測分類器の作成と精度の検証
7. 従来手法との精度の比較

なお，コサイン類似度を用いたスタートアップのクラスタリングの際には，ネットワーク図を描画するツールである gephi [Bastian 09] を用いる．

今回の実験の条件としては，2011 年の段階で本提案手法を用いたことを想定し，2009〜2011 年に初の資金調達を行った米国のスタートアップのクラスタのうち流行しているものに対して，同様の事業がその後日本において流行するか否かを予測する．米国の流行しているクラスタは，クラスタ内の会社数が 15 以上のものとする．ノード数が 15 以上のクラスタに対して，SVM による予測を行う．学習用の正解データは，2012〜2014 年に設立された日本のスタートアップをクラスタリングした結果を用いて作成する．日米間の事業分野が同様のものか否かの判断は，ラベル及びクラスタ内の企業の説明文を参考に判断を行う．

精度の評価について，機械学習による予測の正解は，予測された事業分野と同様のものが日本においてもその後流行する，とする．正解データには 2012〜2014 年に設立された日本のスタートアップのクラスタリング結果を用いる．日本において流行しているクラスタの定義は，ノード数 5 以上とし，分類器の精度評価は K 分割交差検証法を利用し，K=4 とした．従来手法の評価は，2009〜2011 年に Exit した米国のスタートアップが所属するクラスタに対応するものが，日本において 2012〜2014 年に流行するか否か，により行う．分類器の精度評価指標について，正解率，適合率，再現率，F 値がある．分類器の予測と結果の組み合わせについて以下のように分類した際，それぞれ正解率は式（2），適合率は式（3），再現率は式（4），F 値は式（5）により算出することができる．

表 4: 分類器の予測と結果の組み合わせ

| | 実際が正 | 実際が負 |
|---|---|---|
| 予測が正 | True Positive （TP） | False Positive （FP） |
| 予測が負 | False Negative （FN） | True Negative （TN） |

$$Accuracy\text{（正答率）} = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

$$Precision\text{（適合率）} = \frac{TP}{TP + FP} \quad (3)$$

$$Recall\text{（再現率）} = \frac{TP}{TP + FN} \quad (4)$$

$$F - measure\ (F\text{ 値}) = \frac{2PrecisionRecall}{Precision + Recall} \quad (5)$$

本研究では，予測と実際の意味はそれぞれ，以下の表 5.1 に示す通りである．

表 5: 予測と実際の意味

| | 正 | 負 |
|---|---|---|
| 予測 | SVM による出力が正 | SVM による出力が負 |
| 実際 | 入力した米事業が日本でも流行 | 入力した米事業が日本では流行しない |

### 5.2 結果

提案手法の精度評価の結果は以下の表 5.2 のとおりとなった．一方，従来のタイムマシン経営と呼ばれる，米国で明確な成功をおさめた事業を模倣する手法を今回のデータに対して適用した際は，正答率を最大化した際に正答率が 0.5 であった．よって，今回得られた結果は従来手法による精度に比べて高い結果となった．

表 6: 評価指標ごとの分類器の評価値

| 評価指標 | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| 評価値 | 0.55 | 0.59 | 0.88 | 0.70 |

## 6. 考察

本研究の限界について，米国で既に行われている事業を元にするという本提案手法の性質上，日本発で米国を始めとする世界に対し，今までなかったイノベーティブな事業の創出には用いることができない．また，情報源の性質や会社数を流行の基準にしているなどの理由から，対象となる事業分野が IT 関連のものに偏りがちになってしまうという限界がある．

長期的な視点に立つと，日米の時差がさらに縮小していき，本手法事態が成立しなくなる可能性も十分に考えられる．

有用性に関して，今回得られた予測精度は，本手法のみを用いて事業選択を行うのであれば低いといえる．しかし，本提案手法を用いて予測した結果を事業選択の意思決定における 1 つの判断材料として用いるのであれば，有用なものであるといえる．

## 7. 結論と今後の展望

本研究では，現代において従来のタイムマシン経営をそのまま行うのは難しいことを示したうえで，米国企業の情報を活用するうえでの新たな指標を用いて，日本において有望な事業の予測を行った結果，従来手法に比べて高い精度で予測を行うことが可能であり，また一定の範囲で有効な手法であるということもいえた．

今後の展望としては，日米間のみに用いるのではなく，日本とアジア諸国に用いる，など適用範囲を広げることが考えられる．

## 参考文献

[Arora 04] Arora, A., & Gambardella, A. (2004). The globalization of the software industry: perspectives and opportunities for developed and developing countries (No. w10538). National Bureau of Economic Research.

[Quelch 96] Quelch, J. A., & Klein, L. R. (1996). The Internet and international marketing. Sloan Management Review, 37(3).

[Lituchy 00] Lituchy, T. R., & Rail, A. (2000). Bed and breakfasts, small inns, and the Internet: The impact of technology on the globalization of small businesses. Journal of International Marketing, 8(2), 86-97.

[He 09] He, J., & Fallah, M. H. (2009). Is inventor network structure a predictor of cluster evolution?. Technological forecasting and social change, 76(1), 91-106.

[Chang 04] Chang, S. J. (2004). Venture capital financing, strategic alliances, and the initial public offerings of Internet startups. Journal of Business Venturing, 19(5), 721-741.

[Tolstoy 10] Tolstoy, D., & Agndal, H. (2010). Network resource combinations in the international venturing of small biotech firms. Technovation, 30(1), 24-36.

[Yli-Renko 02] Yli-Renko, H., Autio, E., & Tontti, V. (2002). Social capital, knowledge, and the international growth of technology-based new firms. International Business Review, 11(3), 279-304.

[上野山 14] 上野山勝也, 大澤昇平, & 松尾豊. (2014). 人材の転職履歴情報を素性としたベンチャー企業の Exit 予測. 情報処理学会論文誌, 55(10), 2309-2317.

[MacMilan 86] MacMillan, I. C., Siegel, R., & Narasimha, P. S. (1986). Criteria used by venture capitalists to evaluate new venture proposals. Journal of Business venturing, 1(1), 119-128.

[MacMilan 87] MacMillan, I. C., Zemann, L., & Subbanarasimha, P. N. (1987). Criteria distinguishing successful from unsuccessful ventures in the venture screening process. Journal of business venturing, 2(2), 123-137.

[Gompers 01] Gompers, P., & Lerner, J. (2001). The venture capital revolution. Journal of economic perspectives, 145-168.

[Bastian 09] Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. ICWSM, 8, 361-362.