世界 166 カ国の Tweet データを用いた予防概念の類型化
－Hofstede による不確実性回避指標の深化に向けた探索的研究－

古川裕康（日本大学　経済学部）
小林良喜（日本大学　松戸歯学部）

キーワード：予防，不確実性回避指標，トピックモデル

1　はじめに

Hofstede, Hofstede, & Minkov (2010)の不確実性回避指標によれば，アジア圏において日本が最も高い値となっている。不確実性回避の強い国や地域において，人々は曖昧さや不安を恐れ，それらを積極的に予防しようとする傾向があるとされている。しかし予防の対象は多岐に渡る。一概に不確実性回避といえども対象を細分化すると予防意識の傾向は国によって異なる。たとえば日本においては健康予防に関して積極的に取り組もうとする人々は諸外国に比べて多いとはいえないことが分かっている。日本においては国民皆保険制度が充実しているという背景も健康予防に対する意識へ影響を与えていることが考えられるものの，不確実性回避の指標のみでは説明できない人々の予防に対する価値観の違いも存在していることが想定される。

国際ビジネスでは，各国の組織マネジメントやステークホルダー管理において文化的背景や人々の価値観を深く理解することが不可欠となっている。そして国際ビジネス研究領域においては，頻繁に Hofstede の国民文化概念を用いながら多国籍企業の行動と文化・価値観の関係性が論じられてきた。しかし上述の通り Hofstede の国民文化概念には概念の更なる精緻化が求められている（太田・佐藤, 2013）。本報告は不確実性回避の指標に焦点を当てながら，世界各国から収集したビックデータ分析により探索的に本概念を再検討し深化を図るものである。以上を通して本報告は，今後の国際ビジネスに関する実務的，理論的発展に寄与するものである。なお本報告の射程は予防に対する世界的な動向を類型化することである。各国の傾向に関する分析については今後の研究で明らかにしてきたい。

2　予防の対象

不確実性回避の概念は予防に対する人々の考え方と深く関連するものである。不確実性回避とは人々の抱える不安や曖昧さを回避する行動を意味しているが，この回避行動は人々の予防意識・行動にほかならない。予防に関する研究は防護動機理論，健康信念モデル，消費リスク，そして利他的行動といった観点から研究が進められてきた。これらの研究を整理すると，人々が予防を考慮する対象は，物理的脅威，心理的脅威，健康上の脅威，情報リスク，持続的社会に関するものに分類することができる。物理的脅威とは，災害や事故，戦争等といった脅威を意味し，人々は身体的な安全を確保しようとする。心理的脅威とは，内面的，精神的にネガティブな影響を与える脅威を意味する。健康上の脅威については，文字通り健康を損なう危険性に対する脅威を，そして情報リスクは人々が消費活動を実施する

際に必要な情報が不十分になることで被るリスクを意味している。最後に持続的社会に関するものとは，自然・社会環境との共存・共栄のために，自己ではなく他者にネガティブな影響を与えるリスクを予防しようとするものを意味している。

本報告では上記を踏まえ，可能な限り広範な国や地域における人々の行動データを用いて予防行動の傾向を探索的に検証することにした。

3　検証方法

本報告では行動データの中でも Twitter の中に出現するテキスト情報を用いている。Twitter は１日当たり約 1 億 8,600 万人のアクティブユーザーが日々の些細な出来事をつぶやく世界的なソーシャルメディアである。ここで発話されたテキスト情報を分析することによって，人々の思考や行動の傾向を明らかにすることができる。Twitter での発言を意味する Tweet データは近年，人々の景況感や性格・行動予測にも用いられており，社会科学分野における研究対象データとしての実績も存在している。

3.1　サンプリング

筆者は米国の Twitter 本社から同社データの利用許諾を得ており，同社のデータベースから情報を収集するために必要なアクセス権利を譲渡されている。そこで本報告では 2019 年 1 月～12 月における全世界の「Prevention」に関わる tweet データを収集した。サーバーに過大な負荷を与えないため，サンプリングは 1～12 月の毎月末の数日とし，合計で 166 カ国に渡る 176 万 3821 件の tweet をサンプリングした。なお各 tweet の地域特定には，各ユーザーが申告した地名を利用している。

Sommer, et al. (2012)は tweet データを分析する前にクリーニング作業の必要性を提案している。そこで本報告においてもサンプルに含まれる絵文字や記号，URL 等を取り除くクリーニング作業を実施した。またサンプルの中に含まれる言語の活用形の違いを適切に処理するためにステミングと呼ばれる処理を実施しデータセットを作成した。

3.2　分析手法

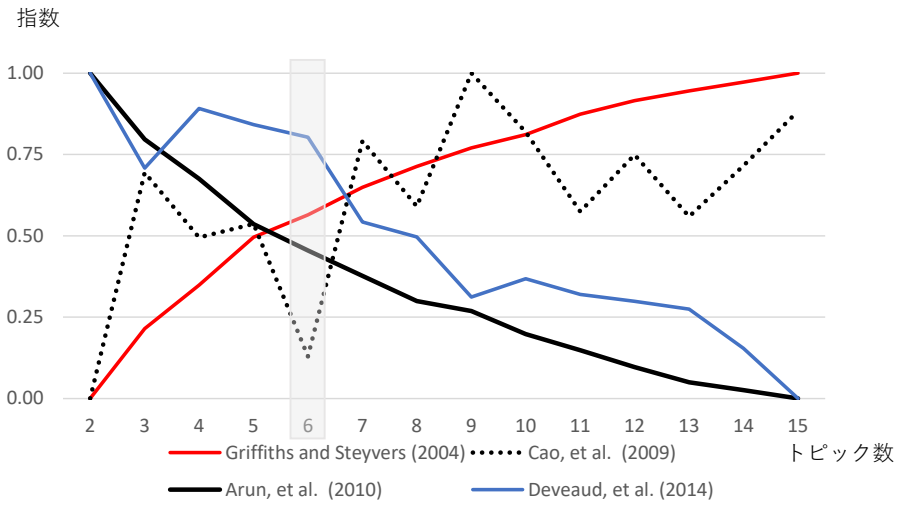
本報告では収集した「Prevention」に関する tweet データに対して，トピックモデルにより機械的に内容を分類した。トピックモデルはテキストデータの内容を特徴的なキーワードから機械的に分類する手法の一つであり，テキスト分析領域における多くの検証に用いられてきた。本報告ではトピック推定にあたり，Blei, Ng and Jordan (2003)の提唱した潜在的ディリクレ配分法（Latent Dirichlet allocation）を用いて検証を進めた。なお分析には R（version 3.6.1）を用いた。

4　結果

データに含まれる要素（トピック）の数は Arun, et al. (2010), Cao, et al. (2009), Deveaud, et al. (2014)，そして Griffiths and Steyvers (2004)によって提唱された各指数によって推定できる。Arun, et al. (2010)ならびに Cao, et al. (2009)による指数は最少と

なるように，そして Griffiths and Steyvers (2004)ならびに Deveaud, et al. (2014)による指数は最大となるトピック数が推奨されている。図表１に各指標を算出したものを示した。ここからデータに含まれるトピック数は６が最適であることが示唆された。

図表１：トピック数の推定



図表１の結果を基にデータセットに含まれるトピック数を６に指定し，トピックモデルによりサンプルの内容を機械的に分類した。図表２にその結果を示している。トピックモデルにおいては，各トピックを良く表すキーワード群を用いてトピックを推定する。そこでトピック分類に用いられたキーワードの内，影響力の強い上位 10 キーワードを抽出し図表２に併せて示してある。これらのキーワードや，それぞれが用いられている各種 tweet の内容を筆者らで解釈したところ，最終的に tweet データには生活，病気，利他的，物理的，心理的，犯罪に関係する内容が存在することが明らかになった。

図表２：Prevention に関する tweet のトピック分類

	Topic label	上位10キーワード
Topic 1	生活	help, need, know, call, talk, someone, holiday, crisis, remind, life
Topic 2	病気	drug, look, poster, hotline, cancer, disease, new, control, use, treatment
Topic 3	利他的	suicide, nation, year, time, lifeline, hard, folk, simple, past, home
Topic 4	物理的	violence, gun, day, today, week, support, work, live, aware, learn
Topic 5	心理的	health, provide, mental, veteran, pour, always, service, plan, billion, congress
Topic 6	犯罪	people, stop, die, freak, line, crime, alone, free, hiv, act

5　解釈とまとめ

世界的に予防に関する人々の傾向には生活，病気，利他的，物理的，心理的，犯罪に関するものが存在していることが分かった。予防に関する既存研究と関連させてみても，生活

（情報リスク），病気（健康上の脅威），利他的（持続的社会），物理的（物理的脅威），心理的（心理的脅威），犯罪（物理的脅威＆心理的脅威）が該当していることが考えられる。心理的トピックに分類されている health というキーワードについては健康上の脅威とも捉えることができる。しかし health というキーワードは精神的な内容（メンタルヘルス）として多く用いられていたため，その他の病気に関するトピックとは別に分類されたことが考えられる。また犯罪トピックのように物理的と心理的脅威に跨る内容も検出された。

不確実性回避の指標に関わらず，予防に対する人々の考え方についてはこれまでも研究が蓄積されてきた。しかし膨大な行動データを用いて世界の人々の予防に関する傾向を国家横断的に測定した研究は筆者らの知る限り存在していない。本研究では今回の検証結果を用いて各国における 6 つのインデックス値を作成し，予防に対する考え方の国別傾向を定量的に示す取り組みを進捗させているところである。報告当日までにその成果の一部を共有できれば幸いである。本報告を足掛かりにして，人々の予防に対する考え方の傾向を国別・地域別に明らかにすることで，本研究は国際ビジネス研究領域で用いられてきた Hofstede の国民文化研究，とりわけ不確実性回避指標の深化に取り組むものである。

※本研究は 2018 年度，日本大学学部連携研究スタートアップ研究費の助成を受けたものです。

参考文献

Arun, R., Suresh, V., Veni Madhavan, C.E. and Narasimha Murthy, M.N. (2010), "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations", In Zaki, M.J., Yu, J.X., Ravindran, B. and Pudi, V. ed. *Advances in Knowledge Discovery and Data Mining, PAKDD 2010, Advances in Knowledge Discovery and Data Mining*, pp.391-402.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, Vol.3, No.4-5, pp.993-1022.

Cao, J., Xia, T., Li, J., Zhang, Y. and Tang, S. (2009), "A density-based method for adaptive LDA model selection", *Neurocomputing*, Vol.72, Issues 7–9, pp.1775-1781.

Deveaud, R., Sanjuan, E. and Bellot, P. (2014), "Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval", *Document Numérique*, Vol.17, No.1, pp.61-84.

Griffiths, T.L. and Steyvers, M. (2004), "Finding scientific topics", *PNAS*, Vol.101, Suppl.1, pp.5228-5235.

Hofstede, G., Hofstede, G.J., Minkov, M. (2010), *Cultures and Organizations –Software of the Mind–*, 3rd edition, McGraw Hill.

Sommer, S., Schieber, A., Heinrich, K. and Hilbert, A. (2012), "What is the Conversation About?: A Topic-Model-Based Approach for Analyzing Customer Sentiments in Twitter", *International Journal of Intelligent Information Technologies*, Vol.8, Issue 1, pp.10-25.

太田正孝・佐藤敦子 (2013),「異文化マネジメント研究の新展開と CDE スキーマ」,『国際ビジネス研究』, 国際ビジネス研究学会，第 5 巻，第 2 号，107-120 頁。