第９回産業日本語研究会・シンポジウム
【第三部】ＡＩ活用時代の日本語データの高度活用事例

大量日本語データから得られる知見の産業応用
－内容分析から筆者の性格推定まで－

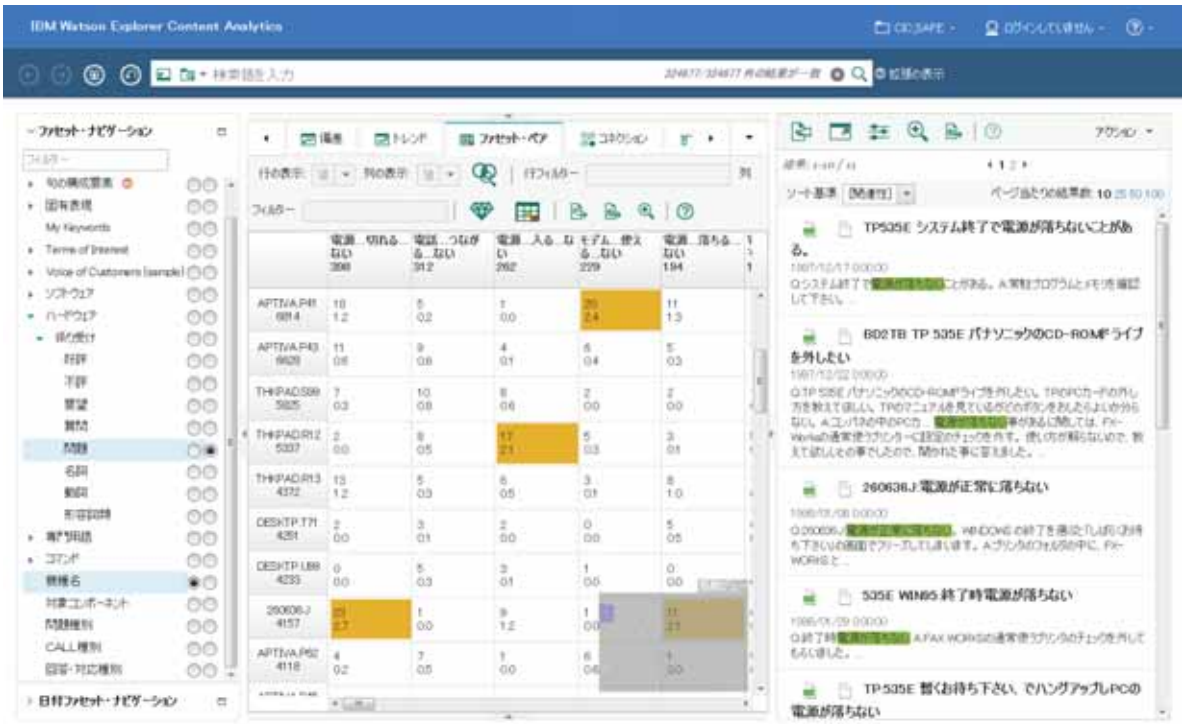
日本アイ・ビー・エム株式会社
東京基礎研究所
那須川哲哉

大量の日本語データの分析例

- 自動車不具合情報の分析デモ
- PCヘルプセンターへの問い合わせ記録の分析デモ

2

データ全体を分析すると、
特定の機種に特定の問題が集中している状況を見出すことができます。



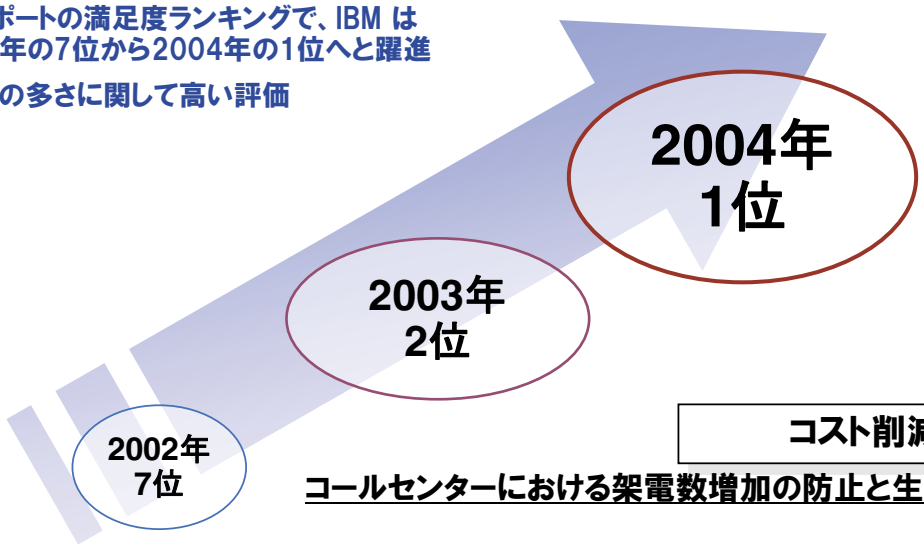
3

テキストマイニングの活用効果

お客様満足度の向上

日経パソコンお客様満足度調査 – PC サポートランキングでの順位向上

- Webサポートの満足度ランキングで、IBM は 2002年の7位から2004年の1位へと躍進
- 情報量の多さに関して高い評価



コスト削減

コールセンターにおける架電数増加の防止と生産性向上

4

TAKMI
Bringing Order to Unstr

The Optimization of Global Railways
A Computer Called Watson
The Rise of the Internet
RAMAC
Excimer Laser Surgery
Magnetic Stripe Technology
The First Salaried Workforce
Optimizing the Food Supply
The Floppy Disk
SAGE
IBM 1401: The Mainframe
UPC
Patents and Innovation

A Global Volunteer Network
Smarter Planet
The Globally Integrated Ent
The Networked Business Pl
The Automation of Personal
The Social Security System
The DNA Transistor
Corporate Leadership in En
The Origins of Computer Sc
The Apollo Missions
Fractal Geometry
Silicon Germanium Chips
Magnetic Tape Storage

5

IBMの100年の軌跡 - Icons of Progress

http://www-03.ibm.com/ibm/history/ibm100/jp/ja/stories/

TAKMI
Bringing Order to Unstructured Data

http://www-03.ibm.com/ibm/history/ibm100/jp/ja/icons/takmi/

TAKMI
構造化されていないデータに秩序をもたらす

1997年、IBM東京基礎研究所の研究員たちが新しい強力なテキスト分析ツールのプロトタイプを開発しました。膨大なテキストのデータベースの中にある大量の埋もれた知識を効率良く獲得し、利用するための新たな罪を、TAKMI (Text Analysis and Knowledge Mining) と名付けたこのシステムが開いたのです。

Mining the Talk: Unlocking the Business Value in Unstructured Information – Part 1 (audio)

この「進歩の象徴」に貢献したえり抜きのチーム・メンバー

- 服部川哲弘
  IBM主席研究員
- 武田浩一
  技術理事、IBM東京基礎研究所、アナリティクス&インテリジェンス マネージャー、自然言語処理
- 渡辺日出雄
  IBM東京基礎研究所、ナレッジ・インフラストラクチャー・グループ マネージャー、自然言語処理
- 荻野紫穂
  研究員、自然言語処理
- 村上明子
  研究員、ソーシャル・アナリティクス
- 金山博
  研究員、自然言語処理(構文解析・意味解析)
- 竹内広宜
  研究員、自然言語処理・知能ソフトウェア工学
- 吉田一星
  研究員、データベース・検索・大規模データ処理
- 坪井祐太
  研究員、統計的自然言語処理
- 宅間大介
  研究員、検索
- 伊川洋平
  研究員、データ工学、テキストマイニング
- 西山莉紗
  研究員、自然言語処理

6

日本語データの分析の深化

- 「何を表現しているか」から

  「どう表現しているか」へ

→ 表現方法から筆者の性格が見えてくる

7

AI時代におけるPersonalityの重要性: 人を理解し個別対応を可能にしたい

Personalityの測り方

- **人相学**
  - **人相から気質や性格を推測**
- **骨相学**
  - **骨格や頭蓋骨の形から知能や行動を推測**
- **血液型性格分類**
  - **血液型によって人の性格を分類**
- **ビックファイブ理論：特性5因子論**
  - **人間の多様な性格は5つの要素の組み合わせで構成される**
  - **膨大な統計調査や脳科学による科学的裏付けが進んできており、世界標準に**

8

性格関連研究の特徴

- **ビッグファイブという標準的なモデルが存在し、**Personality**が数値化可能に**
  - **ビッグファイブ**

    人間の性格を5つの要素の組み合わせで記述
    - Openness to experience
      独創的・好奇心が強い・開放的・知的 vs. **着実・警戒心が強い**
    - Conscientiousness
      手際が良い・勤勉・注意深い・まめ vs. **楽天的・不注意**
    - Extraversion
      外向的・社交的・エネルギッシュ vs. **内向的・孤独志向・控えめ**
    - Agreeableness
      人当たり良い・温情あり・協調的 vs. **冷たい・不親切**
    - Neuroticism
      繊細・神経質 vs. **情緒安定・自信家**

9

ビックファイブの測り方：心理テスト

	全く当てはまらない	当てはまらない	どちらでもない	当てはまる	とても当てはまる
1.人生を楽しんでいる	○	○	○	○	○
2.他人にあまり興味がない	○	○	○	○	○
3.準備を怠らない	○	○	○	○	○
4.ストレスに弱い	○	○	○	○	○
5.語彙が多い	○	○	○	○	○
6.たくさんしゃべるタイプではない	○	○	○	○	○
7.人と接することに興味がある	○	○	○	○	○
8.自分の持っているものを近くに置く	○	○	○	○	○

10

【性格関連研究から得られた知見】

子は親の性格を50％程度引き継ぐ

- Bouchard, Thomas J., and John C. Loehlin. "Genes, evolution, and personality." Behavior genetics 31, no. 3 (2001): 243-273.

11

【性格関連研究から得られた知見】

言葉に性格が反映される

- **文章の特徴に筆者の性格との関連性が見出せる**
  - Mairesse, F., Walker, M.A., Mehl, M.R., and Moore, R.K., Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. In Journal of Artificial intelligence Research, 30: 457-500, 2007.

12

テキストからの性格推定

https://personality-insights-livedemo.mybluemix.net/

13

Personality Insightsとは

- ソーシャルメディアへの書き込みや、フリーテキストから、筆者の性格特性を推定するシステム
  - https://personality-insights-livedemo.mybluemix.net/
    - ① Personality （性格・個性）
    - ② Needs （欲求）
    - ③ Values （価値観）
- 解析に必要なテキスト
  - 少なくとも3,500単語、理想的には6,000単語以上
- 現在対応している言語
  - 日本語、英語、スペイン語、アラビア語、韓国語

14

Personality Insightsが推定すること

- ビックファイブ/OCEAN の5軸でPersonalityを推定
  - Openness to experience:　好奇心が強い・独創的　vs.　着実・警戒心が強い
  - Conscientiousness：　勤勉・まめな人　vs.　楽天的・不注意
  - Extraversion：　外向的・エネルギッシュ　vs.　孤独を好む・控えめ
  - Agreeableness：　人当たりの良い・温情のある　vs.　冷たい・不親切
  - Neuroticism：　繊細・神経質　vs.　情緒安定な・自信家の
- 各軸のさらに細かい推定も可能
  - Openness to experience: 活発度、自己主張、明朗性、刺激希求性、友情、社交性
  - Conscientiousness：　大胆性、芸術的関心度、情動性、想像力、思考力、現状打破
  - Extraversion：　達成努力、注意深さ、忠実さ、秩序性、自制力、自己効力感
  - Agreeableness：　利他主義、協調性、謙虚さ、強硬さ、共感度、信用度
  - Neuroticism：　悲観的、自意識過剰、低ストレス耐性、激情的、心配性、利己的

15

Personality Insightsが推定すること：Needs(欲求)

- Kevin FordのUniversal Needs Map に沿った分析 (欲求と社会的価値の関係)
  - 個人の様々な習慣に関係　：　ブランドの選択、商品の選択、職業の選択
    - Challenge:挑戦
    - Closeness:親密
    - Curiosity:好奇心
    - Excitement:興奮
    - Harmony:調和
    - Ideal:理想
    - Liberty:自由主義
    - Love:社会性
    - Practicality:実用主義
    - Self-(expression):自己表現
    - Stability:安定性
    - Structure:仕組

16

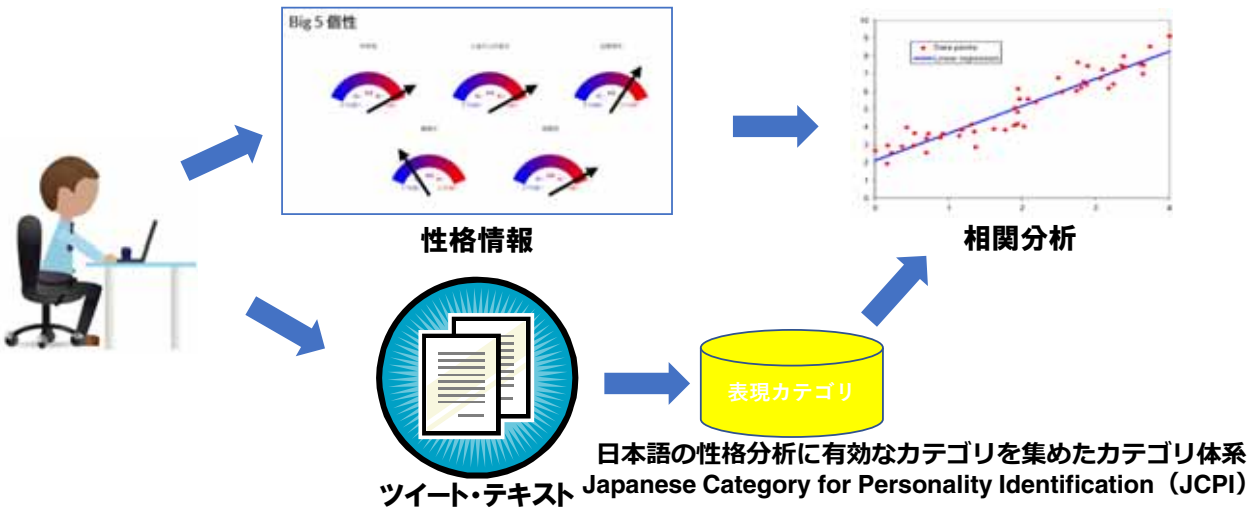
Personality Insightsが推定すること：Value(価値観)

- Schwartzの価値概説 (Schwartz Value Survey) に沿った分析
  - 4つの上位価値と10個の価値によって構成される
    - ４つの上位価値
      - Self-transcendence ：自己超越
      - Conservation ：現状維持
      - Self-enhancement ：自己増進
      - Open to change ：変化許容性
    - 10の価値
      - 博識、善行、調和、伝統、秩序、権勢、達成、快楽、刺激、自決

17

テキストから性格を推定する仕組み

1. ネット上でTwitterユーザに性格診断アンケートを依頼
2. 性格診断アンケートへの回答から推定される性格と回答者のツイートを蓄積
3. ツイート・テキストにおける各表現カテゴリの表現の割合と性格との相関を分析



18

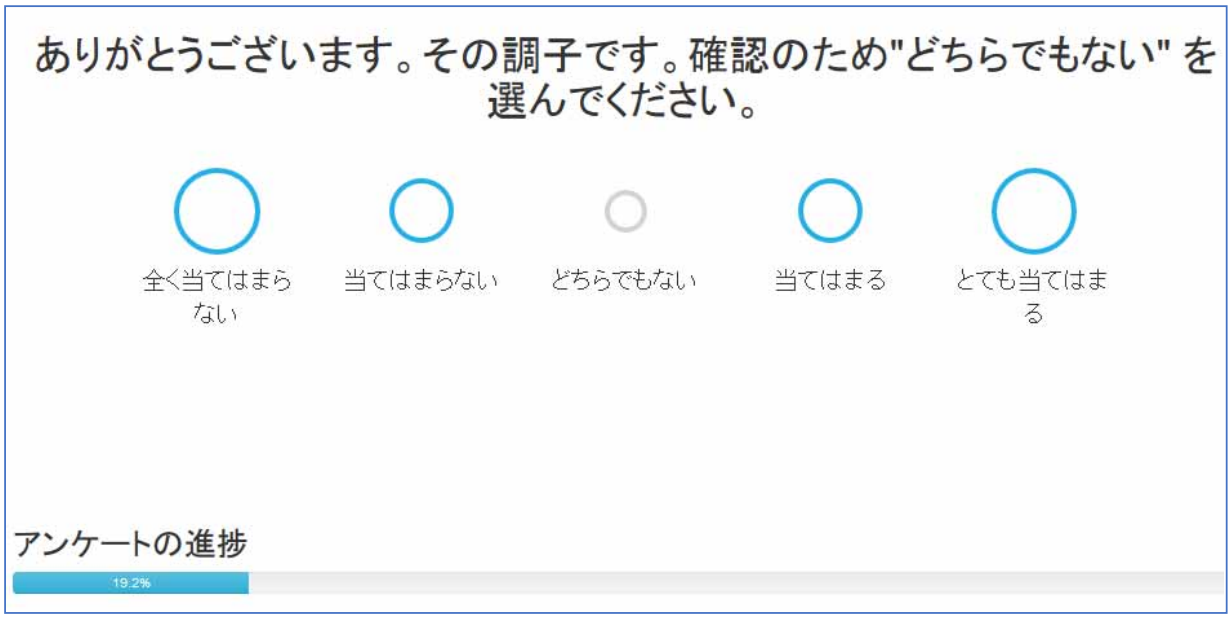
性格診断アンケート

	全く当てはまらない	当てはまらない	どちらでもない	当てはまる	とても当てはまる
1.人生を楽しんでいる	○	○	○	○	○
2.他人にあまり興味がない	○	○	○	○	○
3.準備を怠らない	○	○	○	○	○
4.ストレスに弱い	○	○	○	○	○
5.語彙が多い	○	○	○	○	○
6.たくさんしゃべるタイプではない	○	○	○	○	○
7.人と接することに興味がある	○	○	○	○	○
8.自分の持っているものを近くに置く	○	○	○	○	○

19

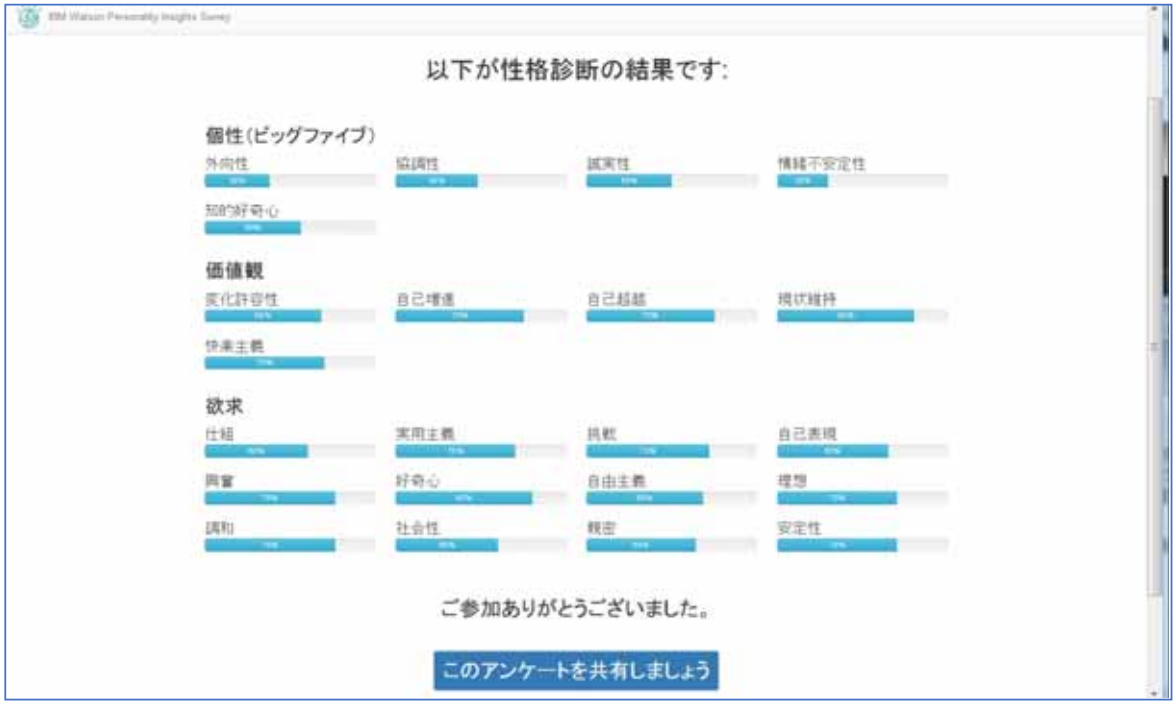
アンケート例 - 確認用質問

- 途中に以下のような質問を何回か入れ込み、まじめに答えていない人（例えば、全て同じところにマークしている人）のデータを排除



20

アンケート結果例



21

テキスト中の表現の分析

- 特定カテゴリの表現がどの程度含まれているかを分析
- 日本語の性格分析に有効なカテゴリを集めたカテゴリ体系を構築
  Japanese Category for Personality Identification（JCPI）
- 下記構成要素からなる約90種類のカテゴリを設定

  【構成要素】

  - Linguistic Inquiry and Word Count (LIWC)を参考にしたカテゴリ
  - 助詞（〔格助詞〕〔係助詞〕など）
    字種（〔漢字〕〔カタカナ〕など）といった日本語独特カテゴリ
  - 〔読書〕〔遊び〕〔イベント〕などの独自カテゴリ

22

JCPIの構成要素： LIWCを参考にしたカテゴリ群

- **LIWC：Linguistic Inquiry and Word Count**
  言語表現の特徴を心理学的観点から整理・体系化し分析する枠組み

I.STANDARD LINGUISTIC DIMENSIONS　（言語的特徴）
- 代名詞（1人称か, 2人称か, 3人称か）・否定形・同意表現・他

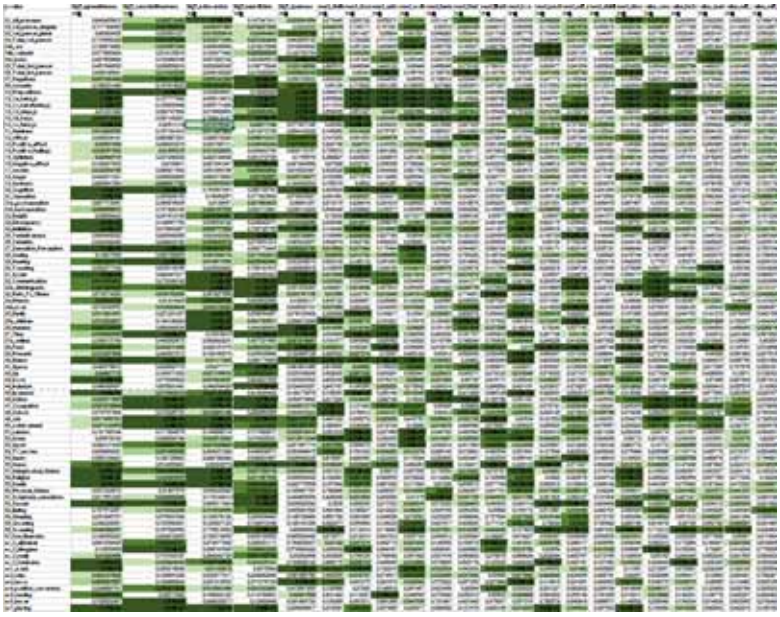
- II. PSYCHOLOGICAL PROCESSES（心理作用、精神的・知覚的プロセス）
  - 好評表現・不評表現・視覚表現・聴覚表現・他
- III. RELATIVITY　（相対的概念）
  - 未来への言及・過去への言及・空間表現・他
- IV. PERSONAL CONCERNS　（関心の対象）
  - 職業関連表現・学業関連表現・趣味関連表現・宗教関連表現・他

JCPIの各カテゴリの表現の出現頻度と性格特性の分析

5%水準以上の有意性	1%水準以上の有意性	0.5%水準以上の有意性	0.1%水準以上の有意性

		big5_agreeableness	big5_conscientiousness	big5_extraversion	big5_neuroticism	big5_openness
JCPI	表現の例	P値	P値	P値	P値	P値
代名詞全体	これ、自分、それ	0.842455815	0.006572465	6.79E-04	0.187047911	0.020041654
一人称代名詞：単数	自分、私、俺	0.717784062	0.001689219	0.028267285	0.016327302	0.5236259
一人称代名詞：複数	我々、私たち、僕ら	0.89593425	0.920691227	0.012050414	0.03464401	0.001365726
一人称代名詞全体	自分、私、俺	0.715662228	0.001857832	0.022377054	0.021823403	0.442395129
俺	俺、おれ、オレ	0.128874083	0.034687343	0.406046669	0.318915648	0.518908003
私	私、わたし、あたし	0.077865434	0.514125816	0.190477489	0.005898906	0.133750754
僕	僕、ぼく、ボク	0.437609884	0.162483168	0.097032723	0.693896514	3.42E-05
二人称代名詞全体	あなた、お前、そちら	0.451549602	0.308659319	0.603990242	0.169776949	0.022198693
三人称代名詞全体	彼、彼女、彼ら	0.946196641	0.010095858	0.346100903	0.215280805	0.005902104
否定形	わからない、知らない、来ない	0.007181493	0.001979067	0.006141266	6.55E-05	0.068792822
肯定表現	OK、認める、了解	0.158221446	0.147016032	0.001622674	9.39E-04	0.002657395
助詞	の、て、が	5.32E-08	8.14E-04	0.226364851	0.15985436	1.32E-10
格助詞	が、に、を、で	1.59E-07	0.137777083	0.458119012	1.51E-07	1.31E-10
格・係・副・終助詞	が、は、に、を	1.72E-07	0.135000448	0.955171438	6.09E-05	1.31E-10
終助詞	か、な、よ、ね	0.91221098	0.082747699	0.273663854	0.006962603	5.57E-05
係助詞	は、も、しか、こそ	0.004214694	0.281142921	0.764013197	3.01E-05	1.31E-10
副助詞	とか、だけ、まで、くらい	8.20E-04	0.23511118	0.016150243	0.03677798	0.577927186
数値	2、1、3	0.013395053	0.187181413	0.011437838	0.312372736	0.854142293

JCPIの各カテゴリの表現の出現頻度と性格特性の分析



- 大半のカテゴリが何らかの性格特性と統計的に有意な相関を示している

5%水準以上の有意性
1%水準以上の有意性
0.5%水準以上の有意性
0.1%水準以上の有意性

25

JCPIの各カテゴリの表現の出現頻度と性格特性の分析

JCPI	表現の例	agreeableness 人当たりの良さ		neuroticism 繊細さ		openness 好奇心の強さ	
		R値	P値	R値	P値	R値	P値
僕	僕、ぼく、ボク	−0.019229991	0.437609884	0.009752	0.6938965	0.101884196	3.42E−05
格助詞	が、に、を、で	−0.128299128	1.59E−07	0.128534	1.51E−07	0.257380587	1.31E−10
係助詞	は、も、しか、こそ	−0.07064671	1.72E−07	0.102578	6.09E−05	0.178145598	1.31E−10
副助詞	だけ、まで、とか、くらい	−0.082488225	0.91221098	−0.05164	0.0069626	0.013786148	5.57E−05
終助詞	か、ね、な、よ	0.002732013	0.004214694	−0.06665	3.01E−05	−0.099139128	1.31E−10

- P値と相関係数（R値）が示唆すること
  - 〔僕〕を多用しがちな人は、好奇心が高い傾向
  - 〔格助詞〕を多用しがちな人は、冷静で、繊細で、好奇心が高い傾向
  - 〔格助詞〕を省略しがちな人は、人当たりが良く、自信家で、着実性が高い傾向
  - 〔係助詞〕を多用する人は、繊細で、好奇心が高い傾向

26

JCPIを用いた性格推定システムの実装

- 性格診断アンケート回答者のツイートを利用し、アンケートから得られた性格プロファイルと各カテゴリ表現の出現頻度とを回帰分析
- GloVeの学習済みWord Vector(200次元)のデータの日本語表現部分を利用し、カテゴリ体系に依存しない文字Nグラム(2≦N≦10)の表現の情報を併用
- 1630人のデータを用いて、0から1の値をとるBig Fiveの各性格プロファイルに対して、10-分割交差検証により、実測値と推定値とのMean Absolute Error (MAE)を計算し、精度確認

27

性格推定システムの性能（MAE）評価結果

	JCPI	GloVe	JCPI+GloVe
協調性	0.1057	0.1027	0.1001
誠実性	0.0962	0.0941	0.0939
外向性	0.1211	0.1158	0.1145
感情起伏	0.1186	0.1147	0.1122
知的好奇心	0.1099	0.1064	0.1063
平均	0.1103	0.1067	0.1054

- カテゴリ体系を用いるより、Word Vectorを用いた方が精度が高いが、カテゴリ体系とWord Vectorを併用した場合に、最も高い精度

28

Personality Insights のチューニング：
結果を正しく解釈するために

- 学習データの増加に応じて、不定期に再学習
  - 同じデータに対して結果が変わる可能性
  - 画面表示・用語も随時変更
- 10万ユーザのツイートデータで分布を正規化
  - 「外向性が90%」という結果の場合
    - 100人中
      - より外向性が高い人が10人程度
      - より内向的な人が90人程度
  - Twitter以外のデータでは百分率は意味を持たない
    - 同じタイプのデータとの比較が重要
      - 同じ業務報告書のテキストでAさんBさんCさんの結果を比較
        - 「誰が最も外交的か」「誰と誰が似ているか」

29

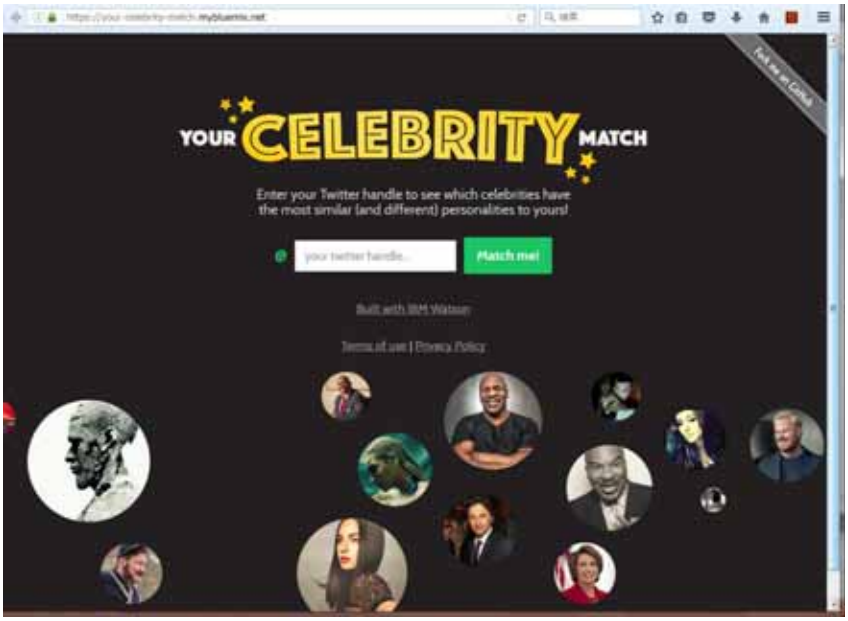
性格推定機能の活用可能性

- 相性が良さそうな人の検出・推薦
- 商品推薦

30

相性が良さそうな人の検出・推薦

https://your-celebrity-match.mybluemix.net/



31

国内導入事例

- **金沢工業大学**

  卒業生や上級生の履修した授業や成績、課外活動などのデータを基に、約7,000人の学生一人ひとりに的確でタイムリーなアドバイスを提供

- **フォーラムエンジニアリング**

  公平で正確かつスピーディーな人材マッチングシステムを構築

- **JAL**

  利用者の性格にあったおすすめ情報が提供するほか、性格診断の結果によって画面のデザインを変化

32

ご清聴ありがとうございました

本資料に記載されている内容は、全て、那須川哲哉個人の見解に基づいており、IBMの見解を示すものではありません。
IBM Watson は、International Business Machines Corporation の米国およびその他の国における商標です。

33