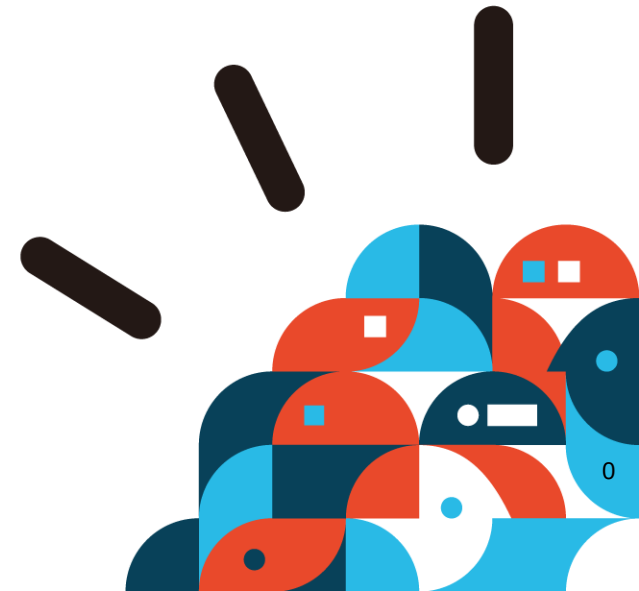




グローバル評判分析から有効なアクションを！ ～ IBM Content Analytics (ICA) ～

日本アイ・ビー・エム株式会社
クラウド・エバンジェリスト
米持 幸寿





- IBMにはテキストマイニング技術を駆使し、コールセンターログやSNS分析をしてきた経験と実績があります。
- 消費者の生の声を様々なテキストから分析しビジネスに活かした事例や実際の技術や方法についてご紹介します。



テキスト・データ

- 文章で表現されているコンピューター・データ
- コンピューターでこういったデータを処理するのは難しい
 - 通常（構造化データ）
 - 項目ごとにデータがばらばらになっている
 - 今増えているデータ（非構造化データ）
 - 説明や名前 と 値が、ひとつのデータにくっついている
 - あるいは、データとして取り出せない



構造化データと非構造化データ

名前	出身	年齢
佐藤	東京	35
斉藤	神奈川	28
高橋	埼玉	32
鈴木	埼玉	41

- 佐藤さんは東京出身で35才。斉藤くんは一番若くて28才で、神奈川県生まれ。高橋さんと私鈴木は揃って埼玉から来ました。彼は32才、私は41才です。

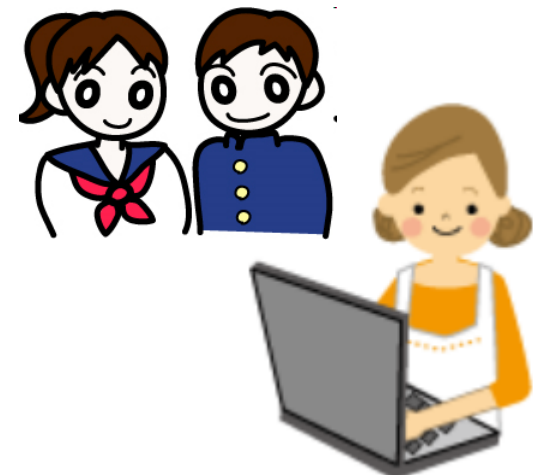
たとえばスプレッドシート

たとえばブログやツイッター

ソーシャル・ネットワーク分析では
右のタイプのデータを分析しなくてはならない

ソーシャル・ネットワークを例に

- 一般消費者にとって
 - ネットへの「不満のはけぐち」
 - 公共の場所での交信手段
- 企業ユーザーにとって（エンタープライズ・ソーシャル）
 - 社内への問題提起・議論
 - 社員同士の意見交換手段

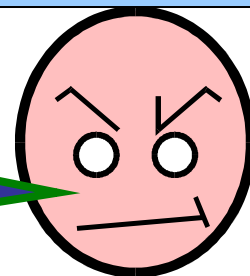


震災のとき、こんな書き込みがあったよ



サーバー倒れた
助けてーっ！

この人の住所、
実在しないよ！
ウソじゃないの！



ボトルウォーターで
赤ちゃんのミルク
作っちゃだめよ！

国産の水は、
軟水だから大丈夫

水、電池が
売ってないよー

ここで買えたよー

何千万メッセージもあって、
傾向を掴むのは手作業じゃ無理





で、どうしたら・・・？

- ソーシャル・ネットワークのテキストをプログラムで収集
 - mixi、ツイッター、facebook、その他・・・
- テキスト・マイニング技術を適用
 - 大量のテキストを、ソフトウェアで「読む」
 - 文章構成、係り受けなどを分析
 - 統計エンジンを使って可視化する

個別に読んだだけでは気が付かない
傾向や兆候を見つけ出すことができる



んなこと
できるわけ
ないじゃん！



いや・・・
できるらしい



米国人気クイズ番組『 Jeopardy! 』に挑戦！ **Watson** = IBMの自然言語処理の集大成



歴代出場者で
最も好成績をおさめた
チャンピオン、
ジェニングス氏(左)と
ラター氏(右)、中央は
Watsonのアバター



幅広いカテゴリーより自然言語で出題される質問



質問を瞬時に理解し解答するスピード

IBM Content Analytics 概観

The screenshot shows the IBM Cognos Content Analytics web application in a Mozilla Firefox browser. The interface includes a search bar at the top, a navigation bar with various tool icons, and a main content area displaying search results. Red boxes highlight specific features, and yellow callouts provide labels for these features.

検索フィールド (Search Field): Located at the top right, it contains the search query: `*:* /"keyword$.5aee2838-fc35-4345-983e-42bda5cb6b0a"/"東京"`. Buttons for "検索" (Search) and "クリア" (Clear) are visible.

操作ボタン (Operation Buttons): Located below the search bar, it includes buttons for "保存された検索 (0)" (Saved Searches), "拡張検索" (Advanced Search), and "照会ツリー" (Query Tree).

ファセットツリー (Facet Tree): Located on the left side, it lists various facets for filtering results, such as "受付日" (Received Date), "性別" (Gender), "住所" (Address), "申告方法" (Reporting Method), "会社名" (Company Name), "車名" (Car Name), "登録年月日" (Registration Date), "総走行距離" (Total Mileage), "型式" (Model), "原動機型式" (Engine Model), and "不具合装置" (Malfunction Device).

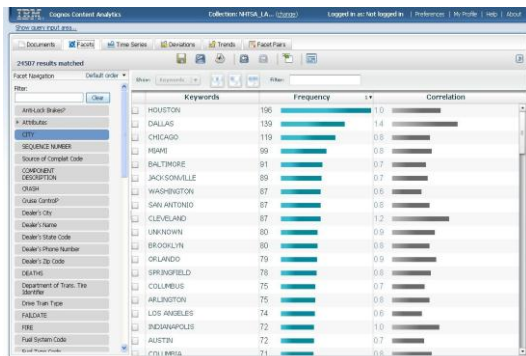
分析ビュー (Analysis View): Located on the right side, it displays a table of search results with columns for "キーワード" (Keyword), "頻度" (Frequency), and "相関" (Correlation). The table shows various keywords and their associated frequencies and correlation scores.

キーワード	頻度	相関
エンジン	612	0.9
動力伝達	259	0.8
制動装置	204	0.8
保安灯火	146	0.9
その他	120	0.8
かじ取り	101	0.7
車枠車体	92	0.7
乗車装置	91	0.7
燃料装置	82	0.7
電気装置	57	0.7

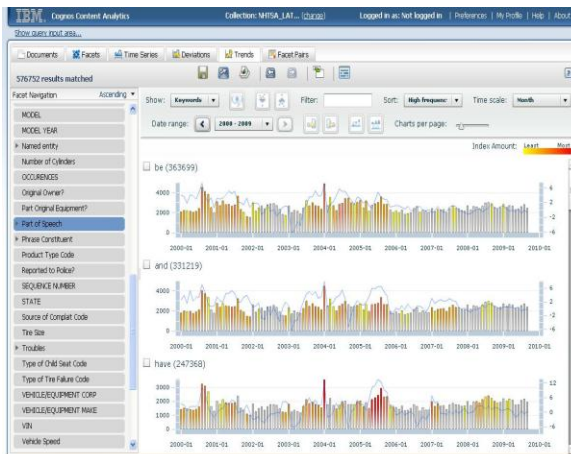
IBM Content Analytics の5つの分析機能

Content Analytics は、以下の5つの分析機能を提供します。

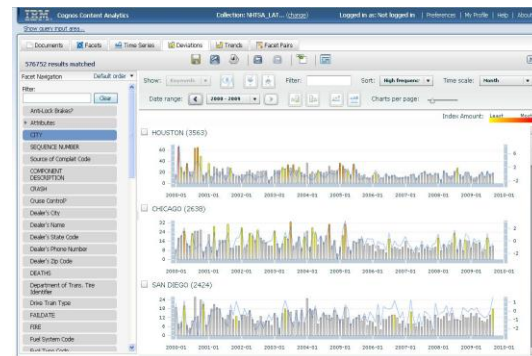
1. ファセット分析



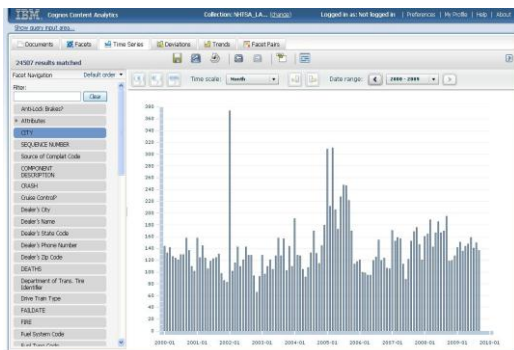
3. 偏差分析



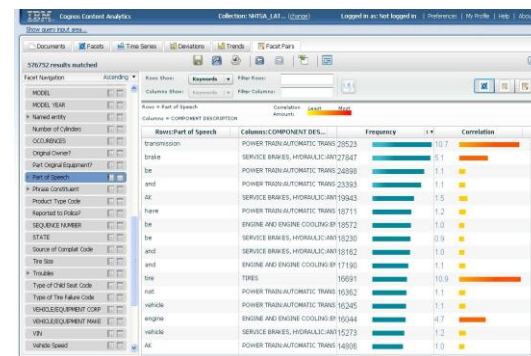
4. トレンド分析



2. 時系列分析



5. ファセット・ペア分析





カスタム辞書・同義語

CSVなどで用意してインポート

- マイニングに利用される単語の収集に使う単語を定義可能
- 同義語として複数語を定義可能

日本, にほん, ニッポン, Japan, ジャパン

新宿, しんじゅく, ジュク

パーソナルコンピューター, パソコン, PC



カスタム分析パターン

XMLのパターンファイルと、正規表現で記述

- 変化するものを収集
 - 行った、行っちゃった、行きました、行ってきた、
- 違う表現でも収集
 - XXXが売ってない、XXXが見つからない、XXXが買えない
- 対象物を収集
 - XXXXが欲しい、XXXはいらない、XXXXが出たら買う
 - XXX色が好き、YYYY色は嫌い、ZZZZ色は悪くない

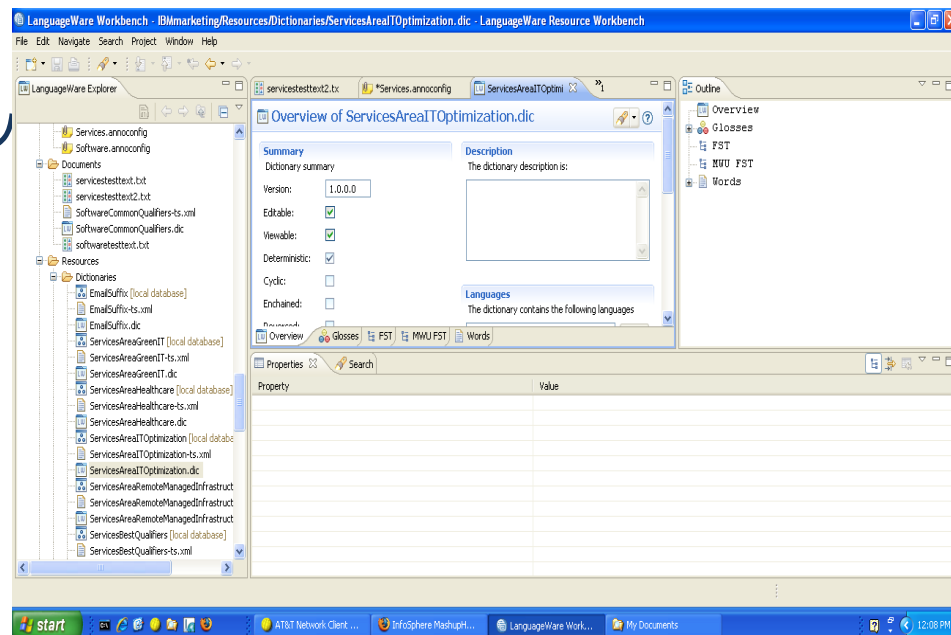
LanguageWare

- 増大するデジタルデータの80%近くが非構造化データと予測されている

- 自然言語テキストを処理するためのソフトウェア

- ノンプログラミングで索引付けを調整

- 人の名前
- 住所
- 製品名
- 電子メールアドレス



- IBM Content Analytics を購入すると、インターネットからダウンロードして利用可能



対応言語

- マイニング
 - 11言語
 - 固有名抽出
 - 5カ国語
 - 評判分析
 - 日本語／英語
 - Nグラム解析
 - 日本語／中国語／韓国語
- 単語レベル抽出
 - 19言語

	自然言語分析(*) (テキストマイニング機能)	単語レベルの抽出 (検索機能)
形態素解析	✓	✓
対応言語	• Arabic • Chinese (Simplified 簡体字) • Chinese (Traditional 繁体字) • Danish • Dutch • English(*) • French(*) • German(*) • Italian • Japanese(*) • Portuguese • Spanish(*)	• Arabic • Chinese (Simplified 簡体字) • Chinese (Traditional 繁体字) • Czech • Danish • Dutch • English • Finnish • French (National) • French (Canadian) • German (National) • German (Swiss) • Greek • Italian • Japanese • Korean • Norwegian (Bokmal) • Polish • Portuguese (National) • Portuguese (Brazilian) • Russian • Spanish • Swedish

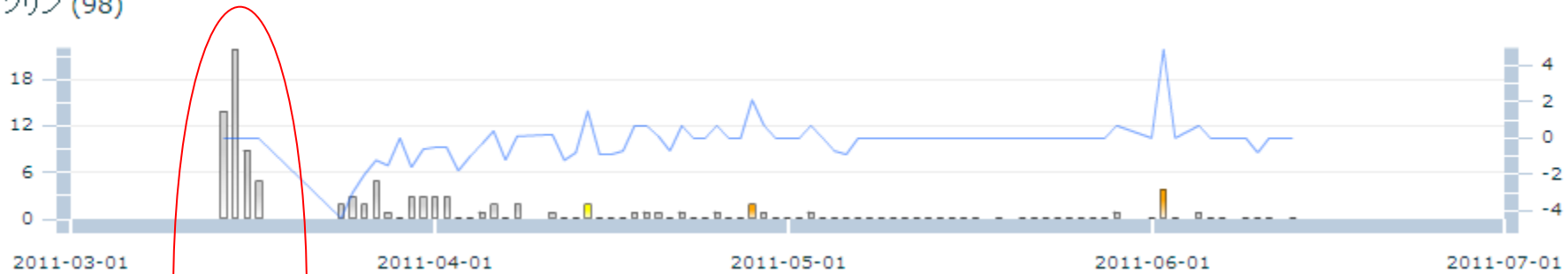
- (*)固有名抽出 (Named Entity Recognition)に対応
(例:「川崎」が人名か地名かを自動認識)
- Japanese, Chinese(Simplified&Traditional), Koreanは、Nグラム解析も可能

<http://www.ibm.com/software/jp/data/search/language.html>

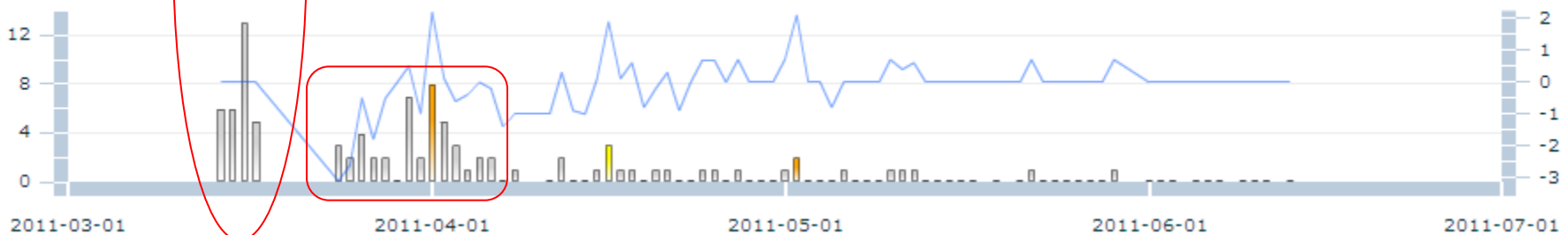


分析例：買い占めが起きたもの

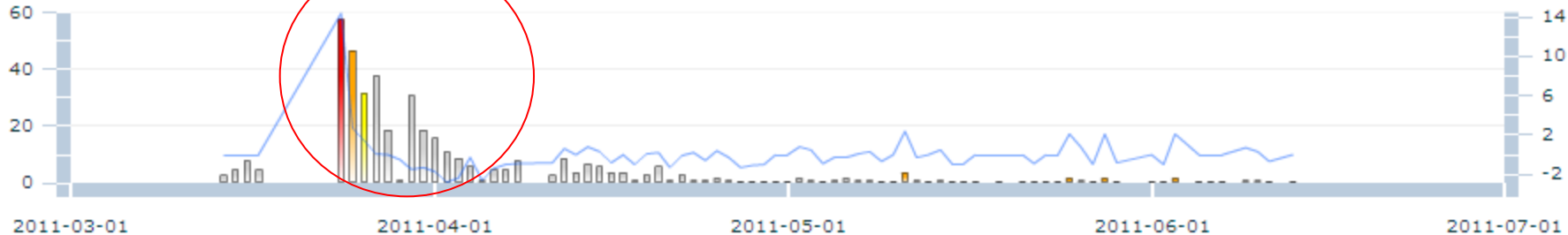
□ ガソリン (98)



□ 電池 (96)



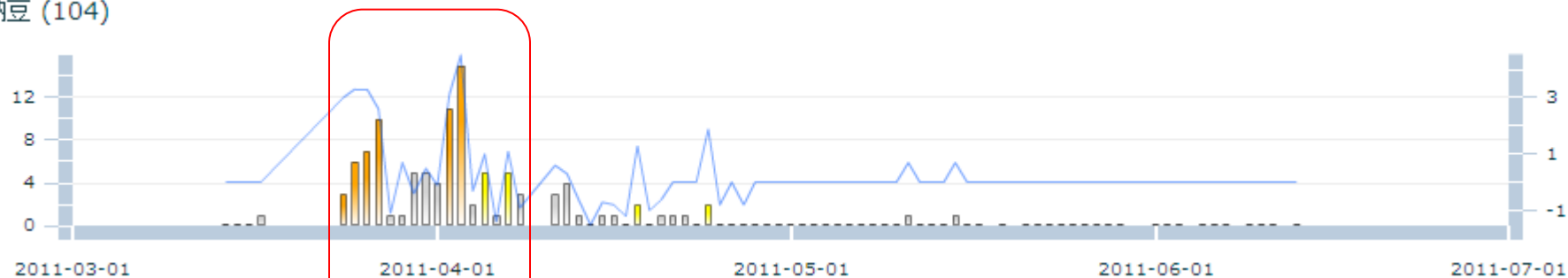
□ 水 (406)



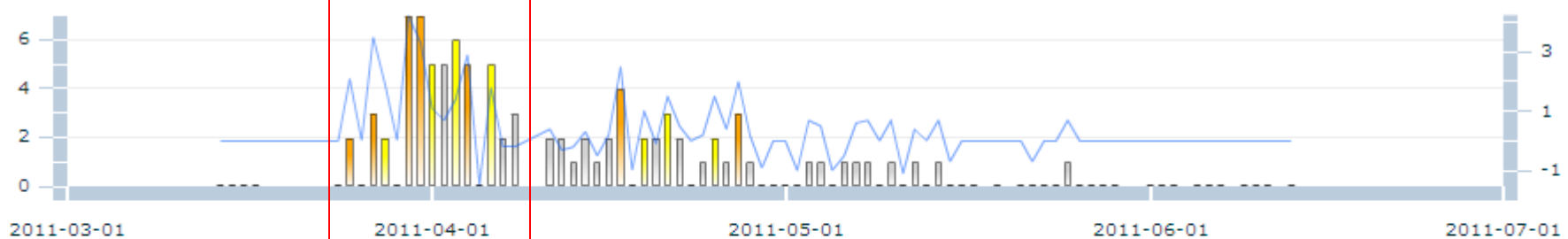


分析例：生産がとまってしまったもの

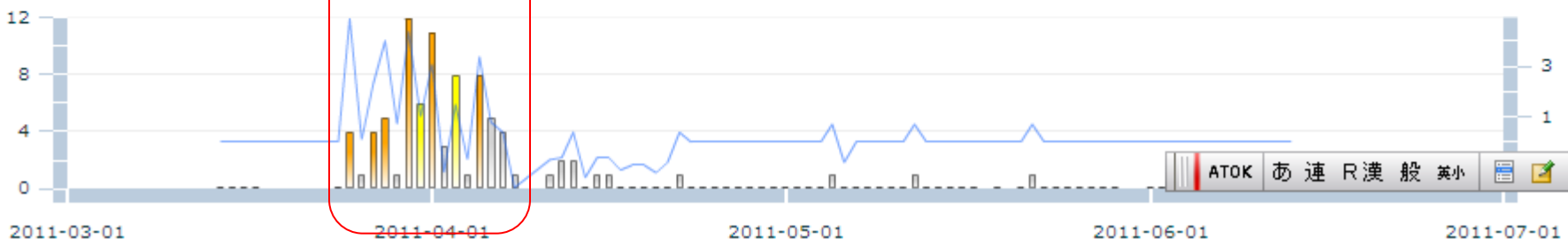
納豆 (104)



タバコ (92)



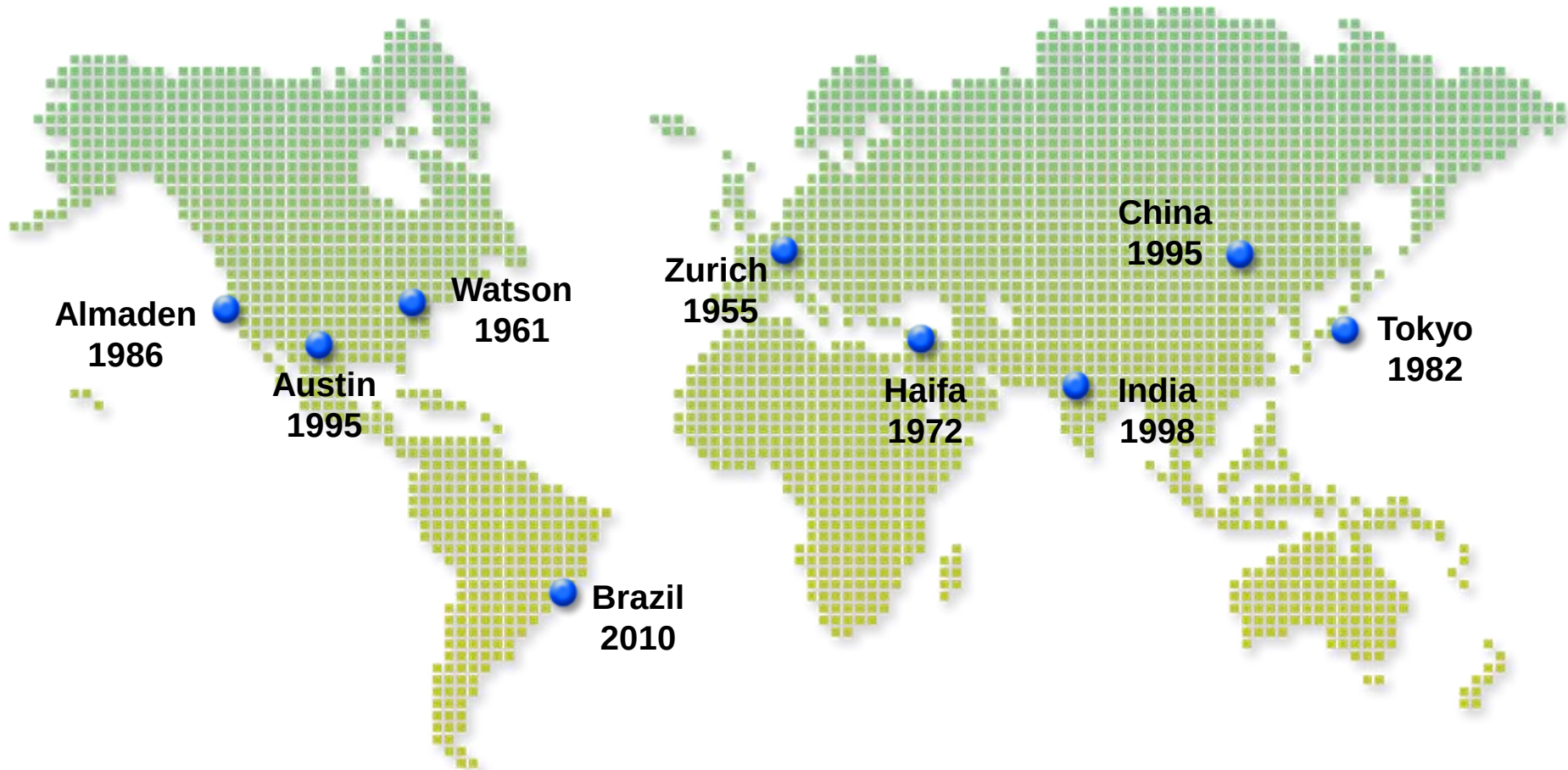
ヨーグルト (85)





IBMリサーチ部門とテキストマイニングの研究体制

- 世界7ヶ国・9つの研究所で、約3,000名が基礎研究にあたる。
- テキストマイニングでは、東京基礎研究所が中心となって、グローバル製品のコア技術を研究。





東京基礎研究所における自然言語処理研究の歩み

研究プロジェクト

IBM製品・アセット

研究成果

1982

かな漢字変換

機械翻訳

5550 日本語ワードプロセッサ

英日機械翻訳システム SHALT
(社内のマニュアル翻訳向け)

コーパスベースの自然言語処理

知識ベースの機械翻訳
(joint project with CMU)

日本語形態素解析

1990

電子図書館と情報検索

インターネット翻訳の王様
“Internet King of Translation”
(Web向け翻訳ソフト)



制約依存文法

日本語構文解析

用例ベース機械翻訳手法

テキストマイニング

mySiteOutliner
(インターネットの情報の収集/分析ツール)

情報アウトライニング

テキストマイニングへの
自然言語処理的アプローチ

2000

WebSphere Translation Server

Enterprise Information Portal

評判分析

ビジネス的価値のある
情報の抽出にテキスト
分析技術を応用

TAKMI and MedTAKMI
(テキストマイニング ツール)

医療/生命科学分野での
情報マイニング

2007

日本及び英語圏向けに製品化・販売

技術的特徴表現の抽出

2009

11言語に対応し世界中で販売
製品名: IBM Content Analytics

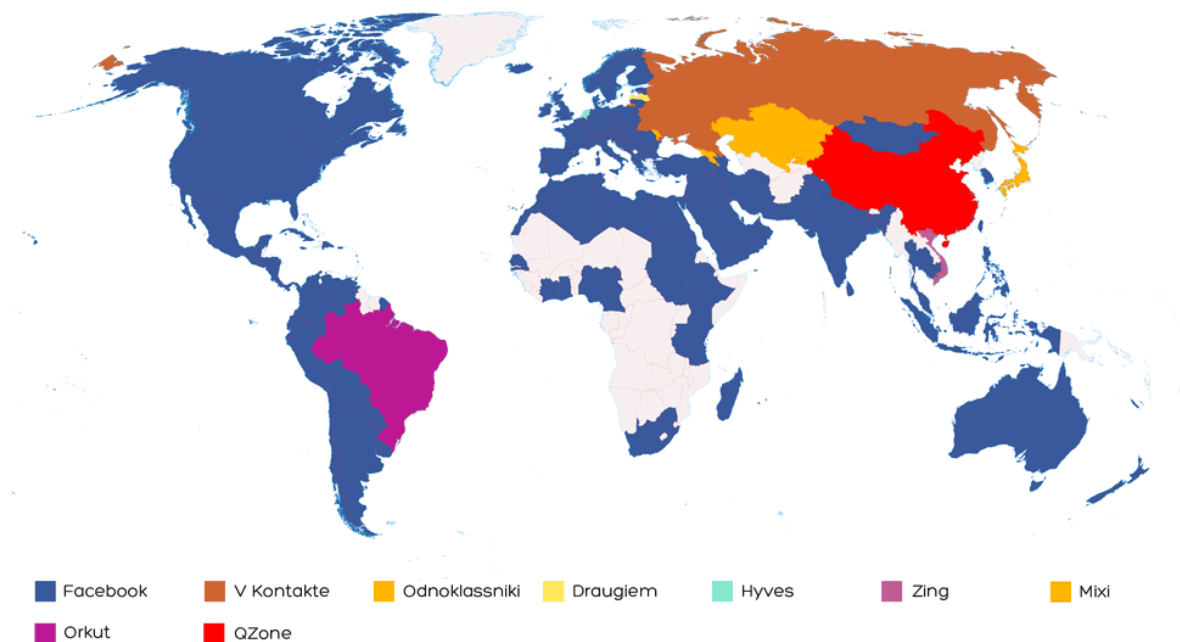
言語横断テキストマイニング



世界のSNSマップ

WORLD MAP OF SOCIAL NETWORKS

June 2011



credits: Vincenzo Cosenza www.vincos.it

license: CC-BY-NC

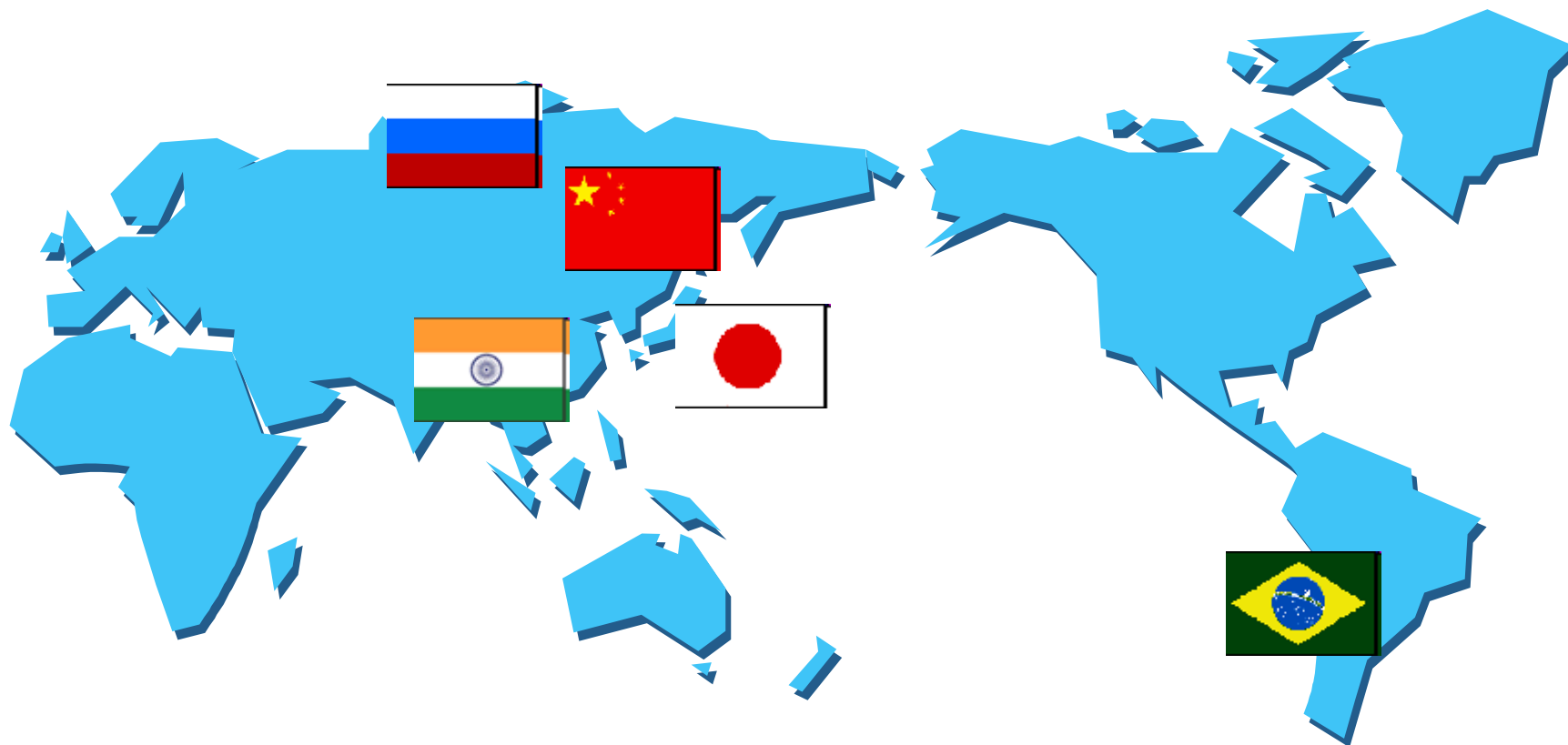
sources: Google Trends for Websites/Alexa

出典: VINCOSBLOG

<http://blogs.itmedia.co.jp/saito/2010/06/snsfacebook13-1.html>



グローバル・分析



IBM Content Analytics では11言語に対応
海外市場の消費者の声を分析できます。

「今後5年間に生活を一変させる5つのイノベーション」第2弾

環境保全、都市づくりにも貢献する将来の交通システムとは

はじめに

交通のイノベーション

日本の専門家に聞く

環境、安全、品質の視点から変革が求められる交通システム

道路特定財源、整備新幹線、羽田空港の国際化交通システムの将来についてではさまざまな議論が活発化しています。急速に進む国内の少子高齢化問題、模で関心が高まっている環境問題に、どのように対応するかが喫緊の課題からです。

IBMが2006年に続き2007年に公表した将来予測レポート「Next 5 in 5」のイノベーションが期待される分野の一つに未来の交通システムがあります。東京など大都市への人口集中が続いており、自動車を中心に発達してきた環境、安全、品質の視点から変革が求められているためでしょう。



「日本の専門家に聞く」で交通のイノベーションについて語る安井さん、日高さん

Next 5 in 5のインタビュー・シリーズ第2弾として、交通システムのイノベーションにおいて重きが期待される自動車業界を取り上げました。IBM インダストリアル・サービス事業部、ビジネス・サービス・リサーチの日本代表、自動車業界の最新動向を聞きました。

うちの会社、
これから
どうして
いくのさ！？

情報ボックス - 当コンテンツに関連するさまざまな情報を、下記のリンクからもご覧になれます。

関連情報

- IBMプレスリリース: 「IBM Next 5 in 5」の第2弾を公表
- IBMが技術革新を予測 5年後の世界はどうなっている？

関連情報(続き)

- IBM Next 5 in 5 第2弾
- IBM Next 5 in 5 第1弾

交通のイノベーション

- 交通輸送イノベーション
- スtockホルムの道路交通イノベーション

サービス/ソリューション

- 製造業向け情報ポータル「匠」業界の部屋 - 自動車業界

イノベーションに関する情報庫

- Ideas from IBM ライブラリー
- IBM CIO イノベーション・チャンネル

当記事のPDF版(236KB)

→ HTML版はこちら

→ 当記事のオリジナル英語版(US)

ようこそ to InnovationJam*

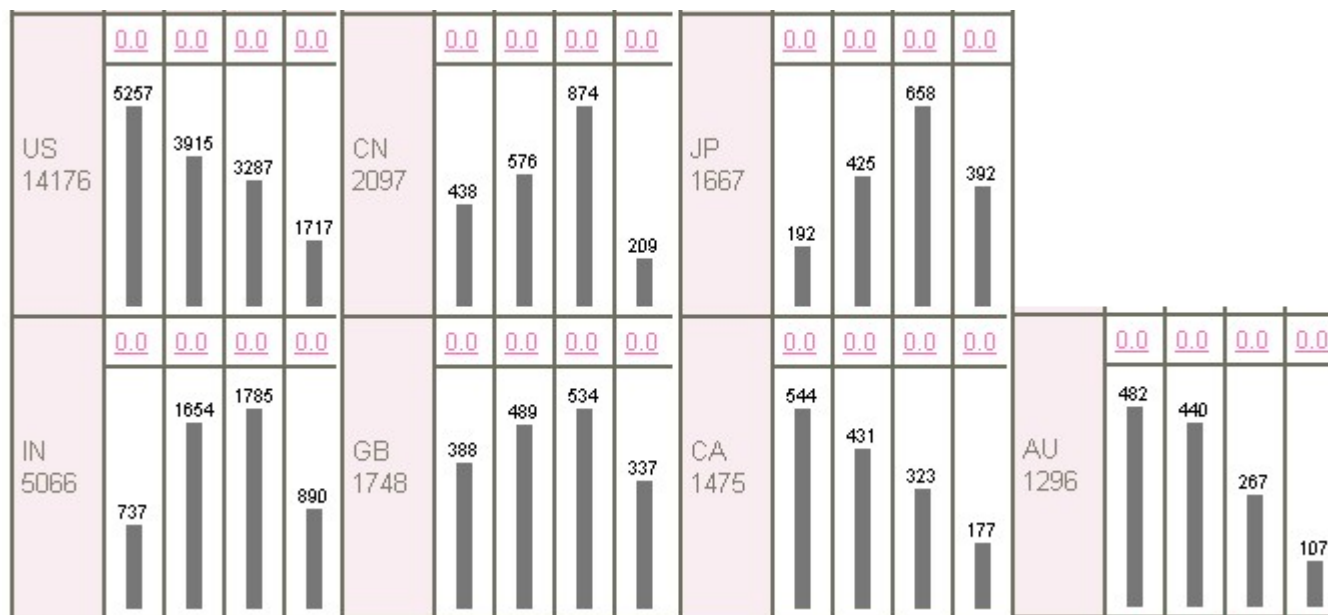
Tens of thousands of innovators from around the world recently collaborated to create new ways of transforming our businesses and society. We will bring those people back together on September 12 to refine those ideas. If you missed Phase One, you can still participate. First, get ready to jam by learning more about...

Why we're jamming:	Where we'll focus:	What we'll use:	How we'll innovate:
An Introduction	Market Opportunities	Emerging Technologies	Beyond Products

ソーシャル・ネットワークで
15万人が参加して議論



国別に見たアイデア入力数の時間的変化



国別に見た各フォーラムのアイデア投稿数

Subcategories/ keywords	Open discussion 9820	Encouraging healthier living 4058	Vehicles of the future	Maximizing existing resources 2562	Digital entertainment 2486	Global travel 2199	Sustainable energy systems 1937	Managing my transactions 1567	Small business goes global 1522	Managing mass transit 1511	Meeting current & future demand 1354	Better diagnosis 1104	Identifying risk & opportunity	Managing costs 1007	Winning my business 999	Pandemic preparation & response 625
US 14176	<u>3324</u> 0.8	<u>1406</u> 0.8	<u>1106</u> 0.8	<u>920</u> 0.9	<u>1048</u> 1.0	<u>894</u> 1.0	<u>882</u> 1.1	<u>706</u> 1.1	<u>671</u> 1.1	<u>494</u> 0.8	<u>436</u> 0.7	<u>533</u> 1.1	<u>434</u> 1.0	<u>632</u> 1.5	<u>382</u> 0.9	<u>308</u> 1.1
IN 5066	<u>1670</u> 1.2	<u>630</u> 1.0	<u>364</u> 0.7	<u>376</u> 0.9	<u>213</u> 0.5	<u>178</u> 0.5	<u>216</u> 0.7	<u>272</u> 1.1	<u>120</u> 0.5	<u>208</u> 0.8	<u>268</u> 1.2	<u>141</u> 0.8	<u>129</u> 0.7	<u>67</u> 0.4	<u>123</u> 0.7	<u>91</u> 0.8
CN 2097	<u>499</u> 0.8	<u>117</u> 0.4	<u>274</u> 1.3	<u>40</u> 0.2	<u>244</u> 1.5	<u>265</u> 1.8	<u>20</u> 0.1	<u>32</u> 0.2	<u>136</u> 1.3	<u>103</u> 0.9	<u>38</u> 0.3	<u>55</u> 0.6	<u>174</u> 2.4	<u>29</u> 0.3	<u>50</u> 0.6	<u>21</u> 0.3
GB 1748	<u>584</u> 1.1	<u>160</u> 0.7	<u>165</u> 0.9	<u>167</u> 1.1	<u>163</u> 1.1	<u>67</u> 0.5	<u>90</u> 0.8	<u>47</u> 0.4	<u>46</u> 0.4	<u>83</u> 0.9	<u>58</u> 0.7	<u>25</u> 0.3	<u>19</u> 0.2	<u>28</u> 0.4	<u>32</u> 0.4	<u>14</u> 0.2
JP 1667	<u>398</u> 0.8	<u>58</u> 0.2	<u>393</u> 2.4	<u>150</u> 1.1	<u>69</u> 0.5	<u>83</u> 0.6	<u>113</u> 1.0	<u>23</u> 0.2	<u>50</u> 0.5	<u>98</u> 1.1	<u>8</u> 0.0	<u>43</u> 0.6	<u>17</u> 0.2	<u>26</u> 0.3	<u>117</u> 2.1	<u>21</u> 0.4
CA 1475	<u>371</u> 0.8	<u>145</u> 0.7	<u>97</u> 0.6	<u>183</u> 1.5	<u>140</u> 1.1	<u>76</u> 0.6	<u>66</u> 0.6	<u>64</u> 0.7	<u>39</u> 0.4	<u>61</u> 0.7	<u>41</u> 0.5	<u>33</u> 0.5	<u>59</u> 1.0	<u>34</u> 0.5	<u>44</u> 0.8	<u>22</u> 0.5

クロス・チャンネル分析



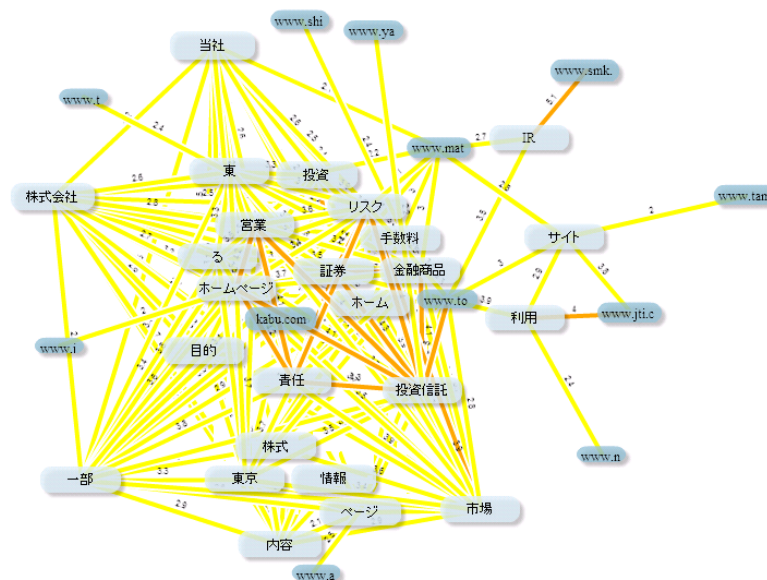
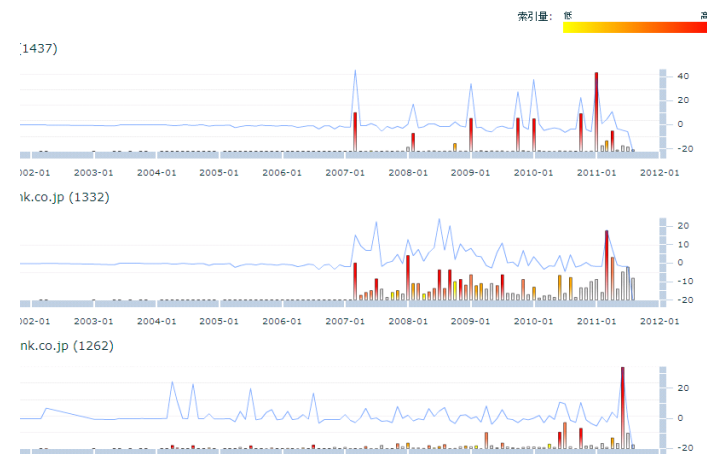
コールセンターログ

修理履歴

ネットニュース

ブログ

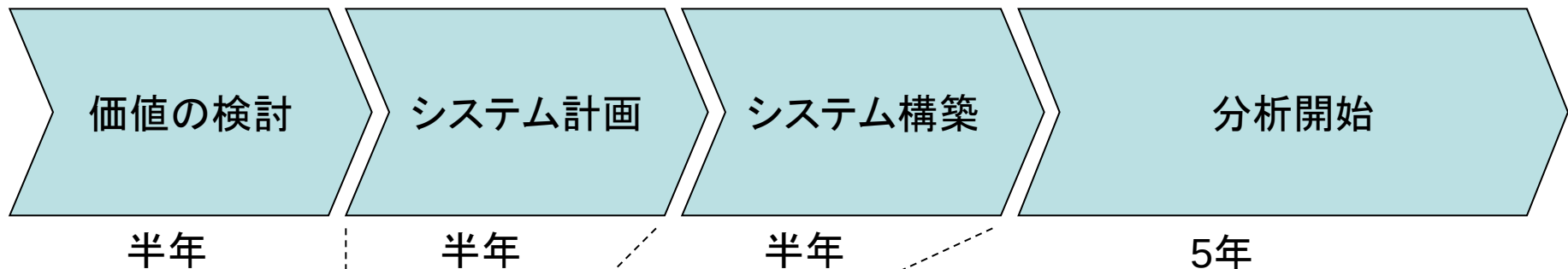
ツイッター



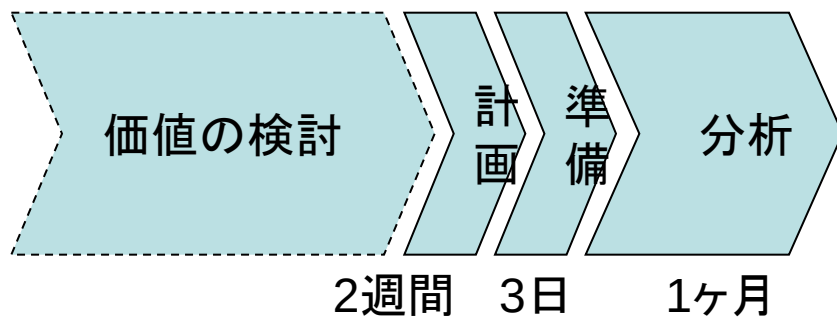


クラウドでの分析

これまでのテキスト分析環境



クラウドでの分析環境



クラウドなら
短期間で準備可能



まとめ

- ソーシャル・ネットワークはユーザー反響意見の「坩堝（るつぼ）」
- ITと言語解析技術で分析可能
- IBM Content Analytics による分析
 - 辞書・パターンのカスタマイズ
 - 多言語に対応
 - クロスチャネルで分析可能
- クラウドにより、手軽に、短期間で利用可能



ご清聴ありがとうございました

- http://ibm.com/software/jp/data/search/text_mining.html
 - CMoffice@jp.ibm.com

グローバル・風評分析ソリューション

http://ibm.com/software/jp/info/post_disaster/analytics.html



© IBM Corporation 2011. All Rights Reserved.

ワークショップ、セッション、および資料は、IBMまたはセッション発表者によって準備され、それぞれ独自の見解を反映したものです。それらは情報提供の目的のみで提供されており、いかなる参加者に対しても法律的またはその他の指導や助言を意図したのではなく、またそのような結果を生むものでもありません。本プレゼンテーションに含まれている情報については、完全性と正確性を帰するよう努力しましたが、「現状のまま」提供され、明示または暗示にかかわらずいかなる保証も伴わないものとします。本プレゼンテーションまたはその他の資料の使用によって、あるいはその他の関連によって、いかなる損害が生じた場合も、IBMは責任を負わないものとします。本プレゼンテーションに含まれている内容は、IBMまたはそのサプライヤーやライセンス交付者からいかなる保証または表明を引きだすことを意図したもので、IBMソフトウェアの使用を規定する適用ライセンス契約の条項を変更することを意図したものでなく、またそのような結果を生むものでもありません。

本プレゼンテーションでIBM製品、プログラム、またはサービスに言及していても、IBMが営業活動を行っているすべての国でそれらが使用可能であることを暗示するものではありません。本プレゼンテーションで言及している製品リリース日付や製品機能は、市場機会またはその他の要因に基づいてIBM独自の決定権をもっていつでも変更できるものとし、いかなる方法においても将来の製品または機能が使用可能になると確約することを意図したものではありません。本資料に含まれている内容は、参加者が開始する活動によって特定の販売、売上高の向上、またはその他の結果が生かると述べる、または暗示することを意図したもので、またそのような結果を生むものでもありません。

パフォーマンスは、管理された環境において標準的なIBMベンチマークを使用した測定と予測に基づいています。ユーザーが経験する実際のスループットやパフォーマンスは、ユーザーのジョブ・ストリームにおけるマルチプログラミングの量、入出力構成、ストレージ構成、および処理されるワークロードなどの考慮事項を含む、数多くの要因に応じて変化します。したがって、個々のユーザーがここで述べられているものと同様の結果を得られると確約するものではありません。

記述されているすべてのお客様事例は、それらのお客様がどのようにIBM製品を使用したか、またそれらのお客様が達成した結果の実例として示されたものです。実際の環境コストおよびパフォーマンス特性は、お客様ごとに異なる場合があります。

IBM、IBM ロゴ、InfoSphere は、世界の多くの国で登録されたInternational Business Machines Corporationの商標です。

他の製品名およびサービス名等は、それぞれIBMまたは各社の商標である場合があります。

現時点での IBM の商標リストについては、www.ibm.com/legal/copytrade.shtmlをご覧ください。

Adobe、Adobeロゴ、PostScript、PostScriptロゴは、Adobe Systems Incorporatedの米国およびその他の国における登録商標または商標です。

IT Infrastructure Libraryは英国Office of Government Commerceの一部であるthe Central Computer and Telecommunications Agencyの登録商標です。

Intel、Intelロゴ、Intel Inside、Intel Insideロゴ、Intel Centrino、Intel Centrinoロゴ、Celeron、Intel Xeon、Intel SpeedStep、Itanium、Pentium は Intel Corporationまたは子会社の米国およびその他の国における商標または登録商標です。

Linuxは、Linus Torvaldsの米国およびその他の国における登録商標です。

Microsoft、Windows、Windows NT および Windowsロゴは Microsoft Corporationの米国およびその他の国における商標です。

ITILは英国Office of Government Commerceの登録商標および共同体登録商標であって、米国特許商標庁にて登録されています。

UNIXはThe Open Groupの米国およびその他の国における登録商標です。

Cell Broadband Engineは、米国およびその他の国におけるSony Computer Entertainment, Inc.の商標であり、同社の許諾を受けて使用しています。

JavaおよびすべてのJava関連の商標およびロゴは Sun Microsystems, Inc.の米国およびその他の国における商標です。

他の会社名、製品名およびサービス名等はそれぞれ各社の商標。