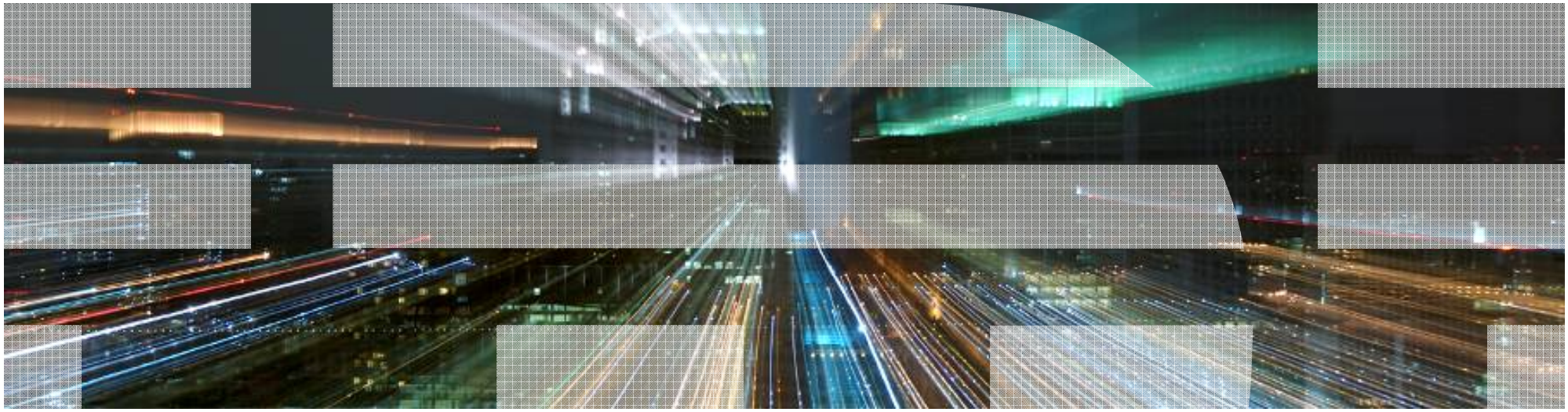
日本アイ・ビー・エム株式会社　東京基礎研究所
那須川哲哉
2012年3月2日

第６回音声ドキュメント処理ワークショップ

テキストマイニング
テキスト化された音声データによる価値創出の可能性



第６回音声ドキュメント処理ワークショップ講演論文集（２０１２年３月９日）
SDPWS2012-04

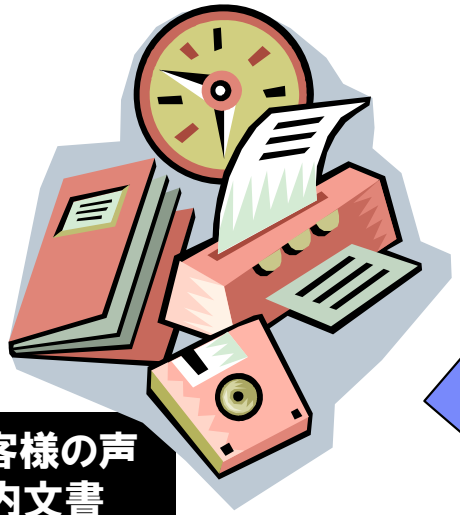
主なトピック

- テキストマイニングとは
  - 仕組みと特徴
  - 活用事例
- 会話データのテキストマイニング
  - 研究の背景
  - 試行事例
- テキスト化された音声データの活用可能性

テキストマイニングとは

テキストマイニングとは　－　テキストからの情報の抽出と分析

蓄積された文書を人が読んで分析・整理するには手間がかかるため死蔵されているケースが多い。
テキストマイニング技術に基づき、効率的なツール、作業のノウハウが提供される。



- お客様の声
- 社内文書
- 特許DBなど

大量の文書データ

人手による処理

非効率

- 精度は高いが時間・手間がかかる
- スケーラビリティがない

テキストマイニング

知識の抽出・発見

新商品、
新サービス、
お客様対応改善、
など

「お客様の声」から…

意向
- 特定商品で目立つ要望・クレームは?

変化
- 最近問題が増えている商品は?
- 直近の3ヶ月に特徴的な要望は?

関連性
- 商品と障害内容の相関関係は?

特性
- 特定の要望が多い年代や地域は?

…といった知見を効果的に獲得できる

テキストマイニングの本質とは

- **個々のテキストを読んだだけでは得られない知見**を獲得する技術
  - 例えば顧客からの問い合わせのテキスト
    - 個々のテキストを読むことで
      - ✓具体的にどのような問い合わせがあったか把握可能
    - 個々のテキストを読んだだけでは
      - ✓どのような問い合わせが多いのか
      - ✓どのような問い合わせが増えているのか

      把握できない

- 大量のテキストデータを全体として分析することにより可能となる

テキストマイニングの長所と短所

- 長所
  - 膨大な量のデータを対象にできる
  - 全体的な傾向や変化を捉えることができる
    - 個々のデータからは分からない
      データを全体を分析することで得られる知見

      ➔大きな成果に繋がる可能性

- 短所
  - 使いこなすには工夫が必要
    - 人手による分析との違いの把握

      ➔テキストマイニングで得られるのは気付き
    - 機械処理のノイズに対する理解

      ➔苦労するだけで終わってしまう危険性

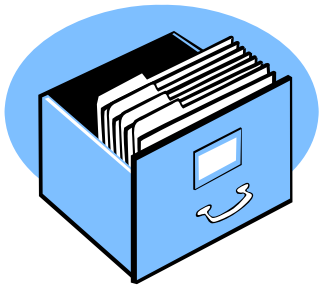
IBMが目指すテキストマイニング技術（TAKMI）

大量の文書データ

レベル1

検索

レベル2



分類整理

レベル3

●お客様の声
●社内文書
●特許DBなど

トピック、パターン、原因・結果などの抽出や、統計的な特徴・傾向の分析

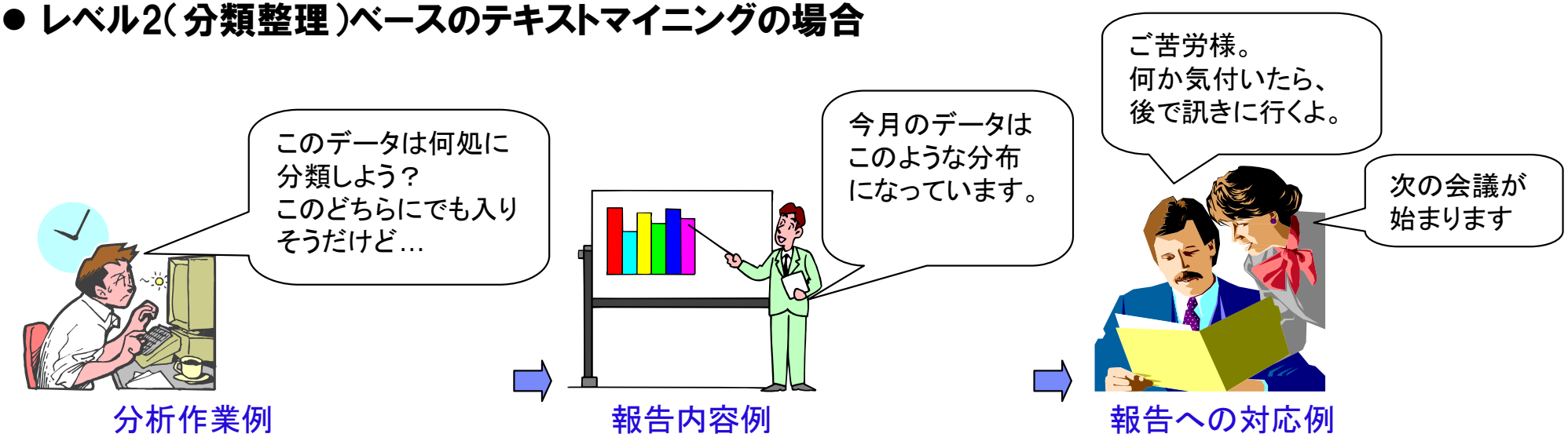
分析レポート作成やアクションに結びつく知識の生成

知識発見

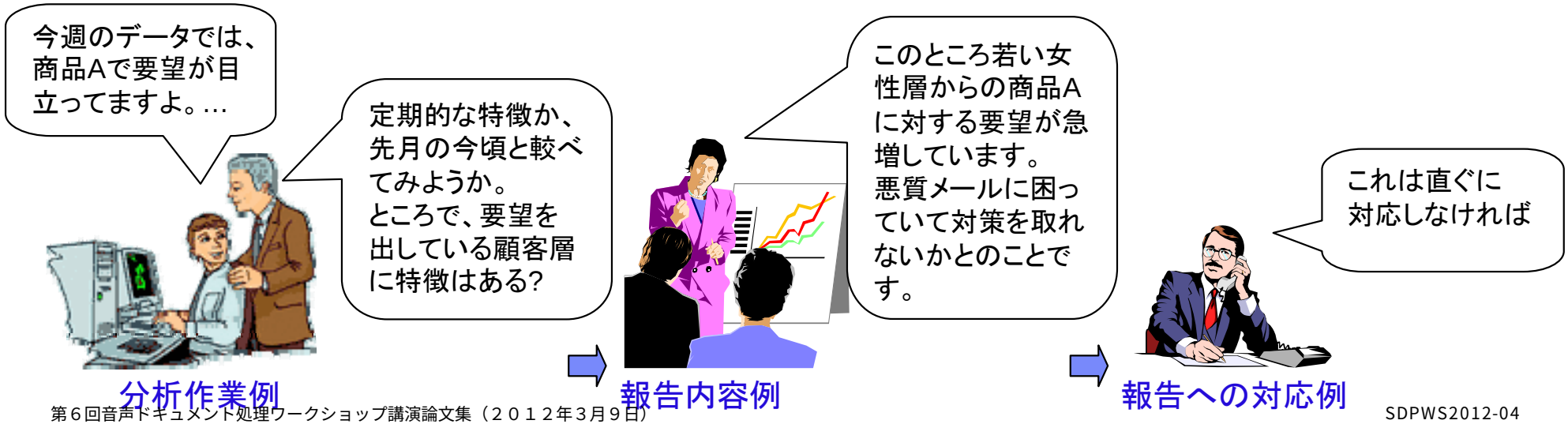
IBM TAKMI がカバーする範囲

アクションにつながるテキストマイニング

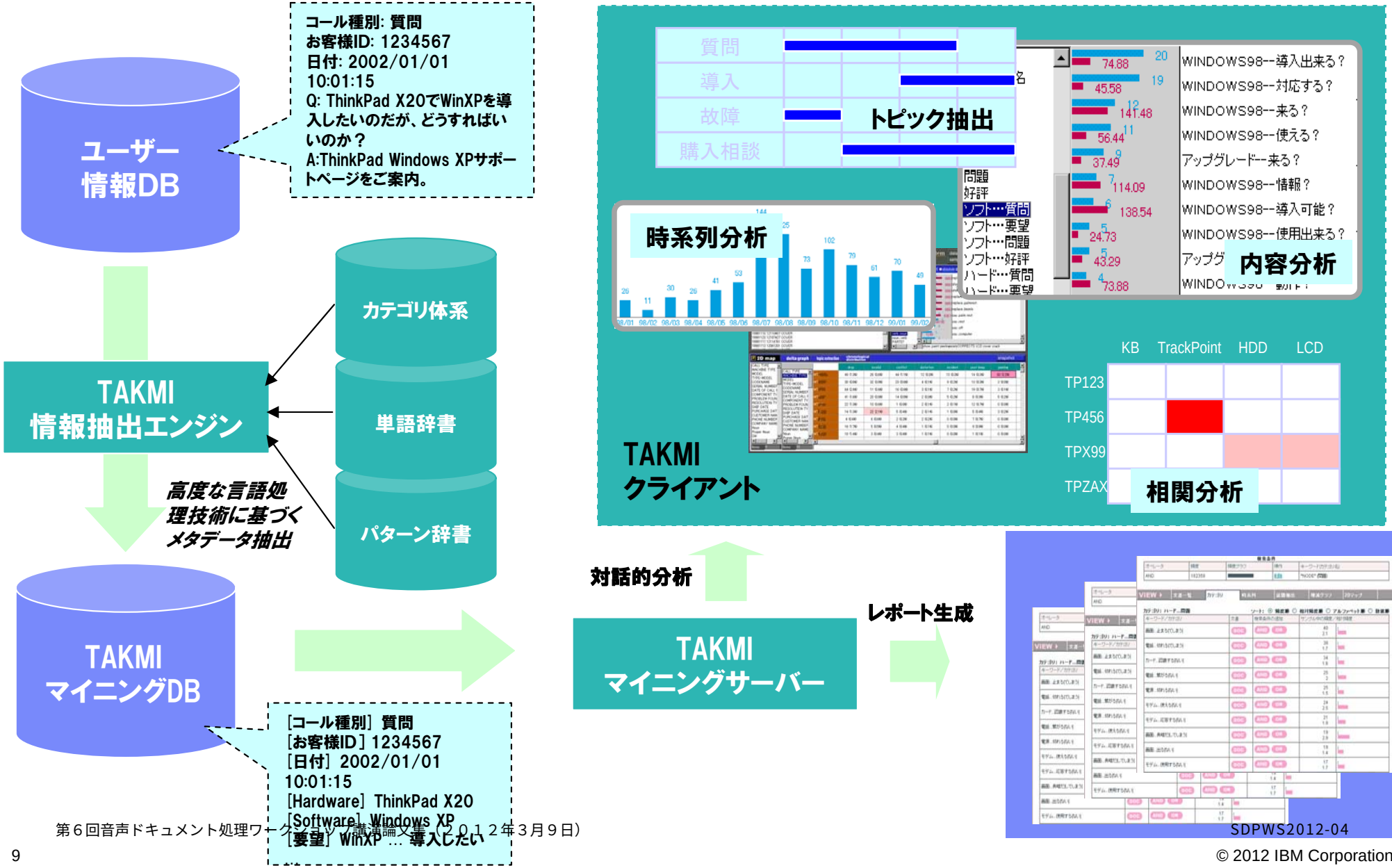
- **レベル2（分類整理）ベースのテキストマイニングの場合**



- **レベル3（知識発見）ベースのテキストマイニングの場合**

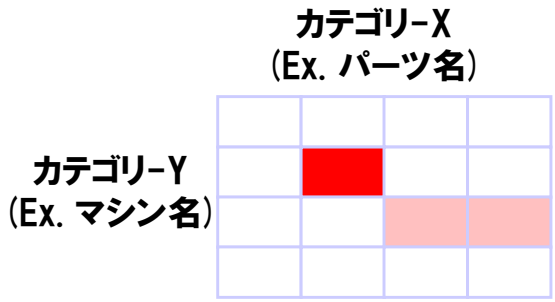


IBM TAKMI (Text Analysis and Knowledge Mining) における処理の流れ
製品名：IBM Content Analytics (ICA)



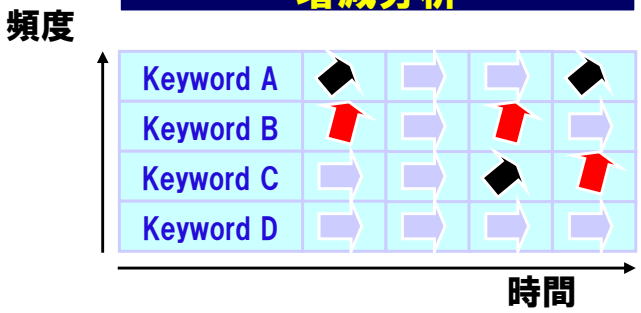
IBM TAKMIのマイニング機能の例

二次元マップ



二つのカテゴリ内の項目間の相関を把握する機能。例えば、製品名と故障部品の相関などを分析。

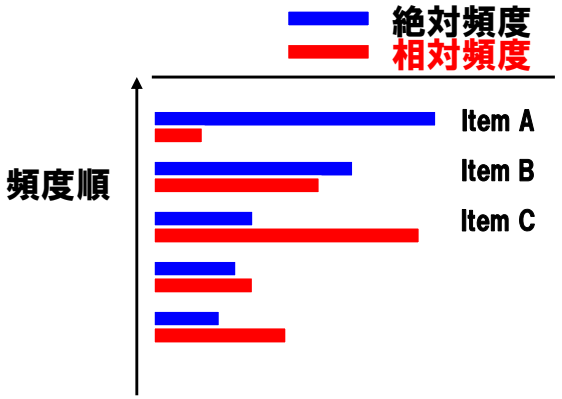
増減分析



ある項目が急に増えたり減ったりする様子を分析。

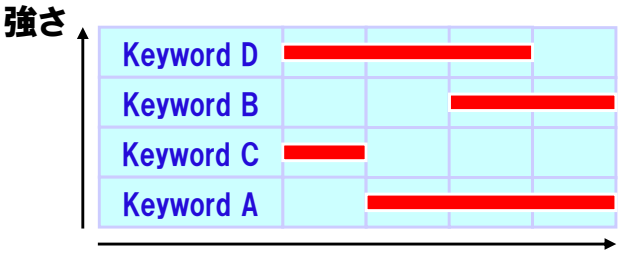
内容分析

（頻度：個々の項目を含む文書数）



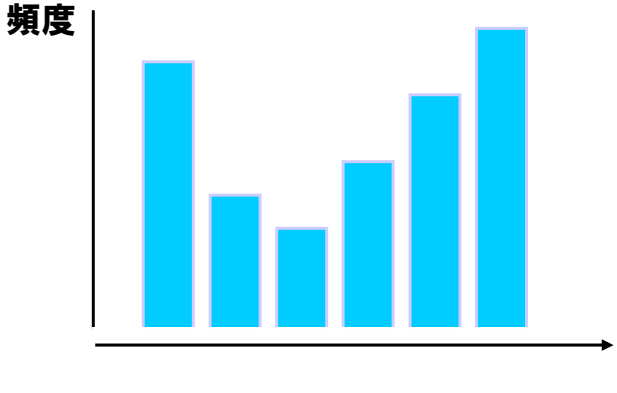
注目する文書集合の内容を概観。相対頻度をみることにより、その文書集合を特徴付ける項目を発見。

トピック抽出



ある時期に目立って現れる項目を分析。

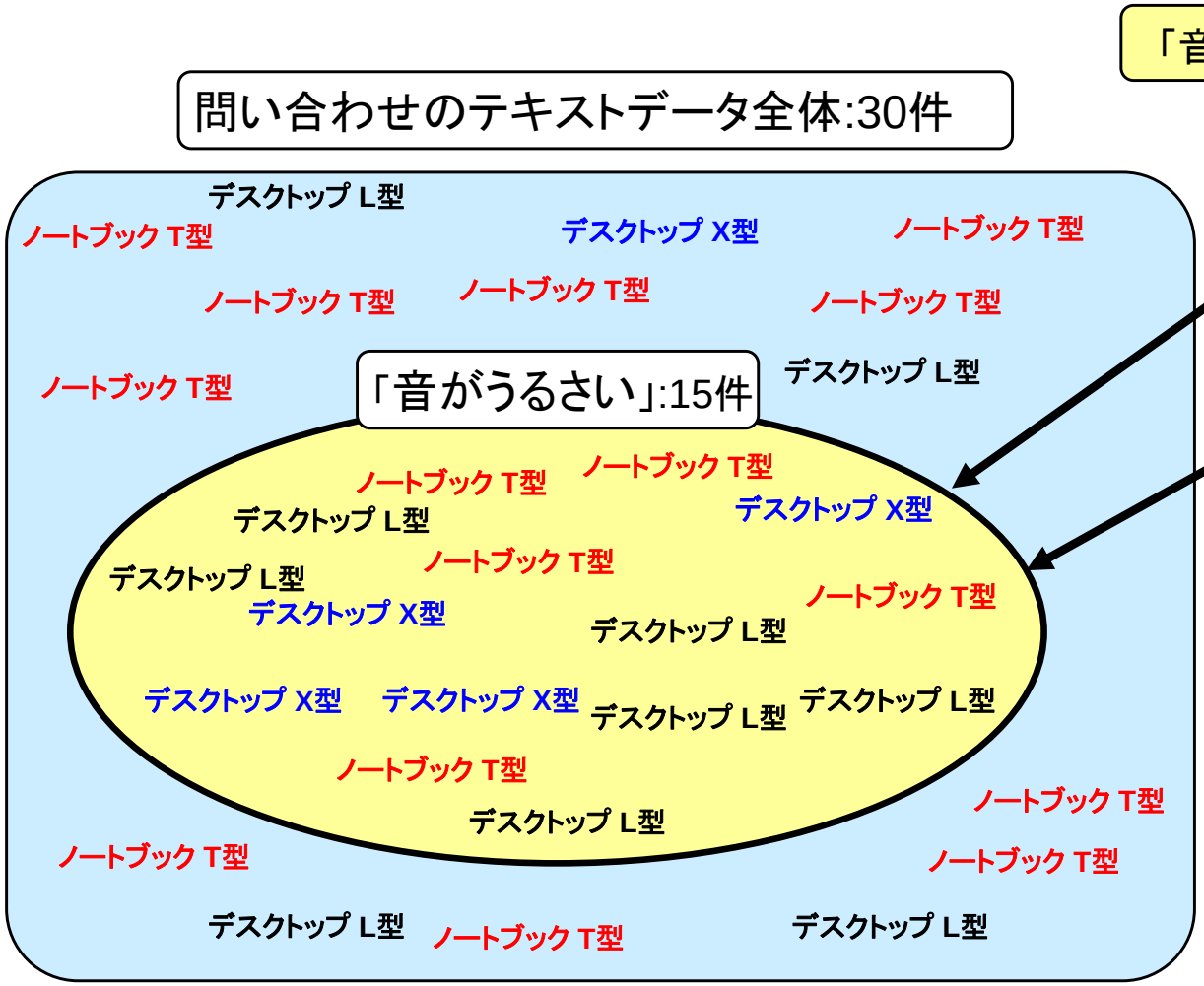
時系列分析



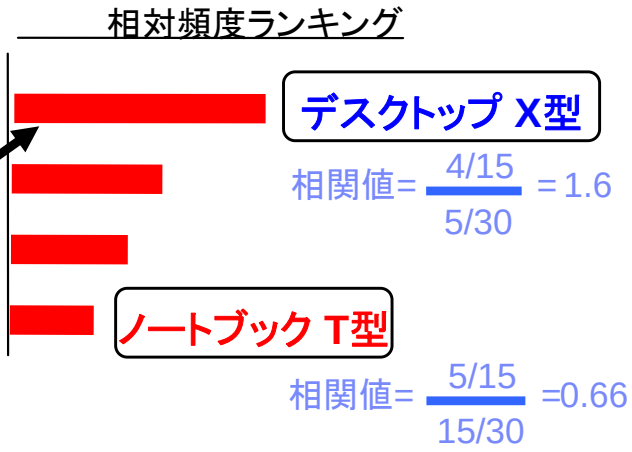
ある項目の時系列上でのトレンドを把握。

IBM TAKMIの相関分析

絶対的な頻度（件数）でなく相対的な頻度に着目



「音がうるさい」声の集合を機種別に見る



「音がうるさい」の集合を機種別で見ると、**デスクトップ X型**の方が**デスクトップ L型**や**ノートブック T型**よりも、相対的密度が高い

デスクトップ X型に関する声は全体的に少ないが、「音がうるさい」という評判では**デスクトップ X型**には特異的な関係がある

テキストマイニングの活用で重要なこと

- 件数分布でなく分布の偏りや変化に着目
- 得られた知見をアクションにつなげる
- 対象データと目的に応じた分析設定の試行錯誤

TAKMIの活用事例

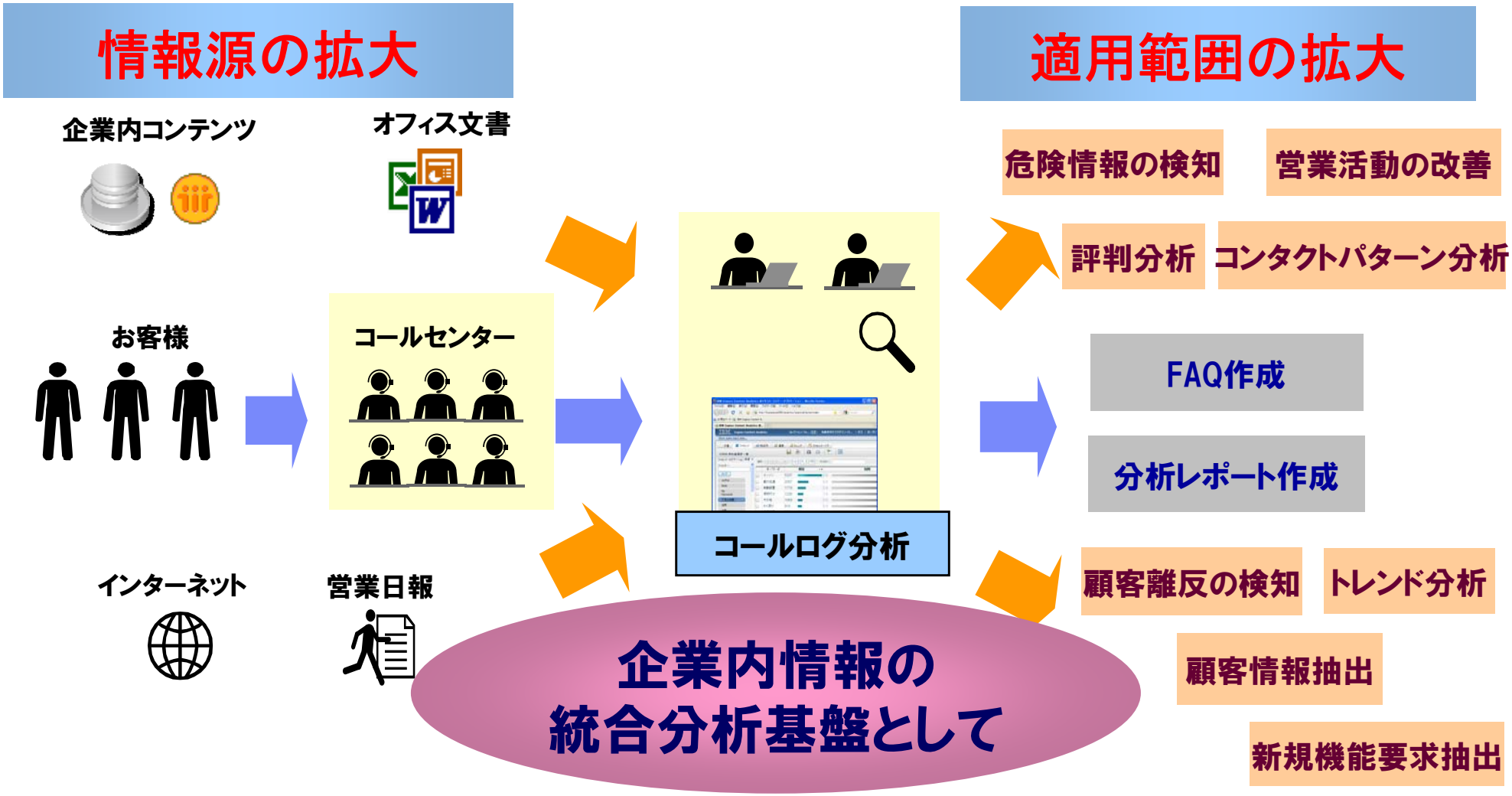
主なお客様導入事例

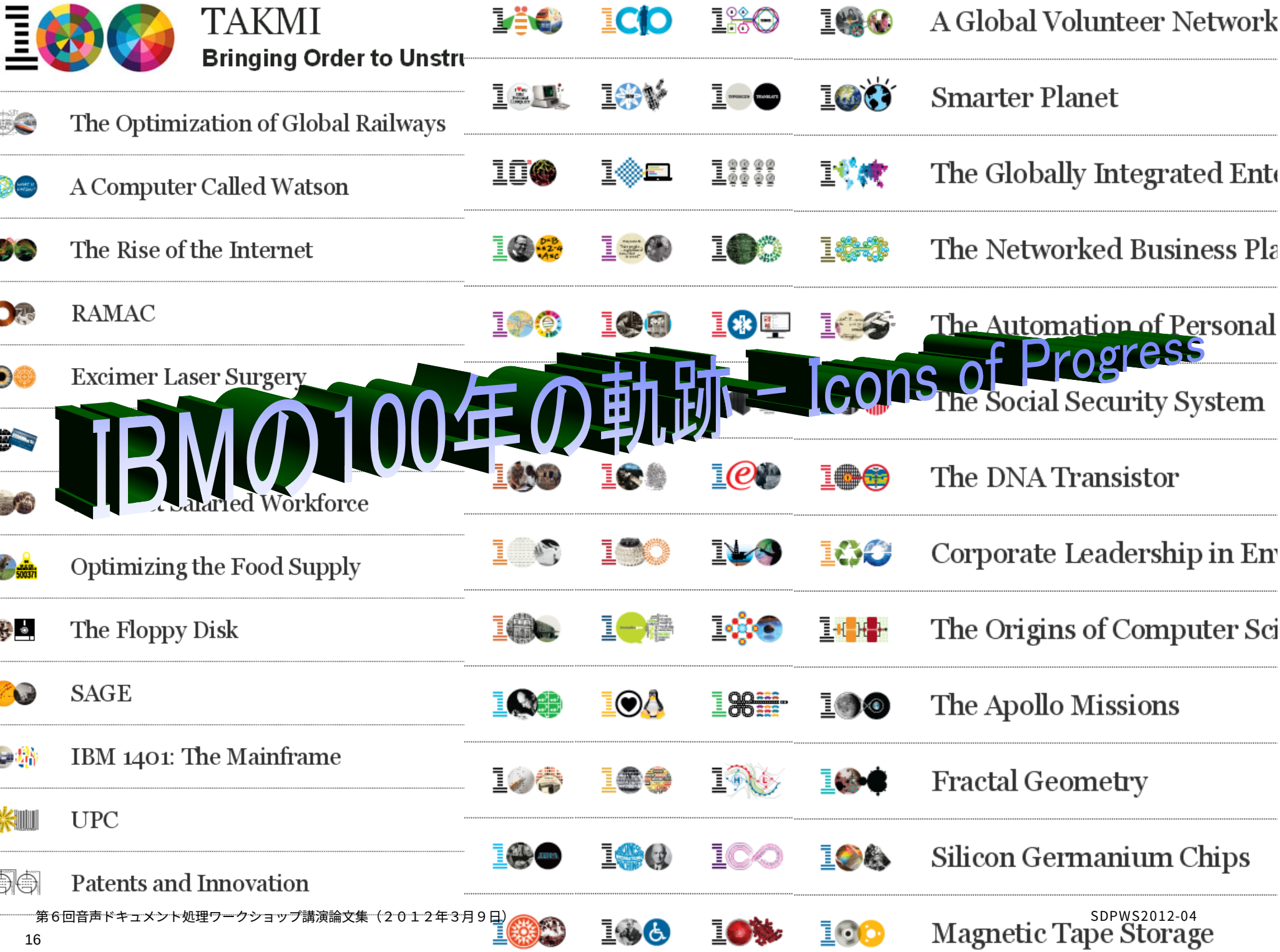
セクター	主なソリューション
放送	コールセンターに集まる年間数百万件のお客様の声の分析によるサービス改善。
製造	製品の不具合の早期発見。
製薬	1千万件以上の学術文献からのプロテオミクスに関する知識・情報抽出の支援。
金融	コールセンターに集まるお客様の声の分析による業務、製品・サービス改善とマーケティングへの活用、コンプライアンスのチェック、など。コールログから顧客の最新の属性情報を抽出し、キャンペーンに活用。
通信	店頭に持ち込まれる商品の障害に関するお客様の声の分析による業務、サービス改善。
流通	コンタクトセンターの問題点の早期発見、商品の問題点の分析、新商品開発のための顧客ニーズの発見、など。

IBM社内の導入事例

IBM社内	ソリューション
IBM-J PC Services and Support （ブリスベン/オーストラリア） （旧IBM川崎PCヘルプセンター）	・コールセンターに集まるお客様の声の分析による業務、製品・サービス改善 ・PCサポートセンターに集まるお客様の声の分析による外部向けFAQ作成支援、問い合わせの各種統計作成支援
IBMラーレイPCヘルプセンター	・コールセンターに集まるお客様の声（英語）の分析による業務、製品・サービス改善
IBM-J アウトバウンド コールセンター	・顧客情報を抽出し、マーケティングに活用 ・エージェントの成功パターンの分析及び教育への活用

アクションにつながるテキストマイニングの広がり





IBMの100年の軌跡 – Icons of Progress

TAKMI
Bringing Order to Unstructured Data

IBM100 ▸ IBM 100年の軌跡 日本 [変更]

← 100年の軌跡のインデックスに戻る

TAKMI
構造化されていないデータに秩序をもたらす

IBMの専門家

那須川 哲哉

IBM東京基礎研究所 主席研究員

1989年IBMに入社。1997年からテキストマイニング・プロジェクトをリードし、開発した技術をTAKMI（タクミ）と命名。テキストマイニング以前には、機械翻訳や電子図書館などのプロジェクトに関与しており、その後も、評判分析、会話マイニング、言語横断テキストマイニングなど一貫して自然言語処理関係の研究に従事。著書に『テキストマイニングを使う技術／作る技術』。

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扉を、TAKMI（Text Analysis and Knowledge Mining）と名付けたこのシステムが開いたのです。

この「進歩の象徴」に貢献したえり抜きのチーム・メンバー

- **那須川哲哉**
  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）
- **竹内広宣**
  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

テキストマイニングに十年以上従事して感じていること

世の中には活用されていないテキストデータが溢れている
- 企業内に限っても
  - 多大な労力をかけて入力された多種多様な報告書が蓄積され
  - 多大なコストをかけて保守されている
  - データの存在は認識されていても内容は誰も知らないことが多い
  - 活用できるにも関わらず活用されていないことも多い
    - ✓活用するのは自分の仕事でないと考えている人が多い

大きなチャンスが存在
- 誰も使っていないデータを活用すれば誰よりも大きな成果を出せる可能性
- 情報を賢く使い、無駄を排して、より良い世の中へ

IBM東京基礎研究所におけるテキストマイニング研究

	トピック・技術名	概要	対外的評価
1997	TAKMI (Text Analysis and Knowledge MIning)	文書中の表現の出現分布の偏りを有益な知見につなげる技術	TAKMIで1999年情報処理学会第59回全国大会大会優秀賞受賞
	MedTAKMI	医療文献をマイニングする技術 ・千数百万件のMedline文書 ・巨大なオントロジー ・カルテのように連続性を持つデータ	
2002	**評判分析・嗜好分析・要望分析** Extraction of Sentiment and Preference ExpRessions (ESPER)	感情や意見を分析する技術 ・好評・不評の判断 ・好不評を示す表現の学習 ・要望の抽出	評判分析ESPERで2005年度情報処理学会山下記念研究賞受賞 要望分析で2005年言語処理学会第11回年次大会優秀発表賞受賞
	ディスカッション分析 (JASMIN) Jam AnalysiS and MINing	意見の出し合いを分析する技術 ・良いアイディアの選択	会話マイニングで2007年言語処理学会第13回年次大会優秀発表賞受賞
2007	**会話マイニング**	生の会話を分析する技術 ・音声認識エラーへの耐性の検討 ・会話の流れの着目点の認識	会話マイニングで2007年度人工知能学会研究会優秀賞受賞
	将来技術予測のための技術的可能性表現抽出 (CAPHMIT) CApability PHrase MIning Tool	特長表現を認識する技術 ・何が特長となるか ・意外性の評価	会話マイニングで2008年度人工知能学会論文賞受賞
	TExt and Network Analysis (TENA)	ネットワークで表現される多様な言語外文脈を考慮してテキストマイニングを行う技術	テキストマイニングの研究開発および実用化の取り組みで2009年度人工知能学会現場イノベーション賞金賞受賞
	言語横断テキストマイニング	多言語データを一元的に母国語で分析する技術	

会話データのテキストマイニング

〔会話データのテキストマイニング〕
研究の背景

- テキストマイニング活用の進展に伴う、より高度な分析への期待の高まり
  - コールセンターでの活用において
    人手により要約したテキストデータではなく、
    生の会話内容を直接分析したい
    - 人手による要約・入力の手間が省ける
      担当者が顧客対応に専念することで、
      生産性・顧客満足度向上の可能性
    - より詳細な内容が分析対象になる
      担当者が従来入力しなかった情報
      担当者にとって不都合な情報
      担当者が重要性に気付いていない情報
- 音声認識技術レベルの向上
  - 電話のやり取りを高い精度で自動的にテキスト化できる可能性の高まり

〔会話データのテキストマイニング〕
研究の概要

- 会話マイニングの有益性の検討

  【疑問】生の会話全てをテキスト化したデータの分析は、人手による要約の分析より有益か

  - 生の会話は冗長なことが多い

    本筋と必ずしも関係の無い雑多な内容がノイズとなる可能性

  【実験】実際に生の会話全てをテキスト化したデータをテキストマイニング

  - 実際のコールセンターにおける約千件近い応対を全て人手でテキスト化

    テキストマイニングシステム IBM TAKMIにより有益な知見が得られないか分析

- 会話マイニングの実現可能性の検討

  【疑問】自動音声認識システムのエラーによるノイズがどの程度までなら有益な分析が可能か

  - 自動音声認識システムが100％の認識精度を実現することは不可能

    人間が聞き取っても完全な書き起こしは不可能

  【実験】人手で書き起こしたテキストにノイズを入れてTAKMIで分析

  - 多段階のノイズを人為的に入れたテキストと自動音声認識によりテキスト化したデータを用意

    TAKMIによる分析で、元データと同じ特徴が検出できるか確認

〔会話データのテキストマイニング〕 実データを用いた有益性の検討

デモ

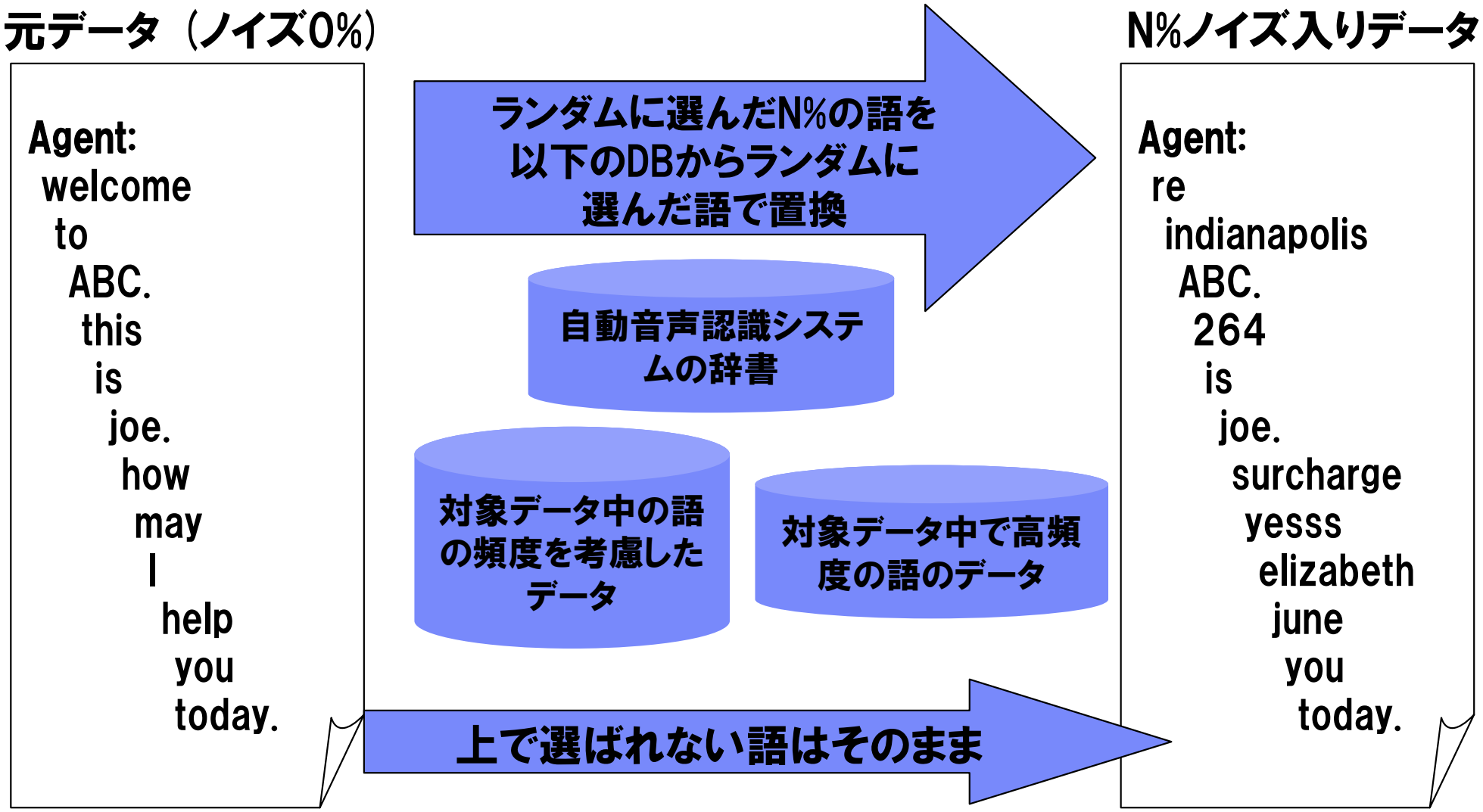
〔会話データのテキストマイニング〕 現実的な活用可能性の検討

- 多段階のノイズを人為的に挿入したデータの作成

〔会話データのテキストマイニング〕
元データ
《応対の全会話を人手により書き起こしたテキストデータ》

Agent: welcome to ABC. this is joe. how may i help you today.
Customer: ya. I want to see how much is the reservation is.
Agent: sure from which location.
Customer: from Houston George Bush Intercontinental.
Agent: what date and time.
Customer: umhh what date is on April 20th is it a Thursday or Friday.
Agent: that's thursday.
Customer: ok. 20th till may the second.
Agent: and at what time that would be.

〔会話のマイニング〕
多段階のノイズを人為的に挿入したデータの作成



50% ノイズ入りデータ

Agent: re indianapolis ABC. 264 is joe. surcharge yesss elizabeth june you today.
Customer: birmingham ehhh I fare to see assistance talk is venus dialled code
Agent: sure from members talked
Customer: send Houston 03:22 Bush Intercontinental.
Agent: thousanc r downtown
Customer: virgir
Agent: that's hil
Customer: gues
Agent: 55 at wh

元データ

Agent: welcome to ABC. this is joe. how may i help you today.
Customer: ya. I want to see how much is the reservation is.
Agent: sure from which location.
Customer: from Houston George Bush Intercontinental.
Agent: what date and time.
Customer: umhh what date is on April 20th is it a Thursday or Friday.
Agent: that's thursday.
Customer: ok. 20th till may the second.
Agent: and at what time that would be.

〔会話データのテキストマイニング〕
現実的な活用可能性の検討

- 人手で書き起こしたテキストにノイズを入れたデータをTAKMIで分析

デモ

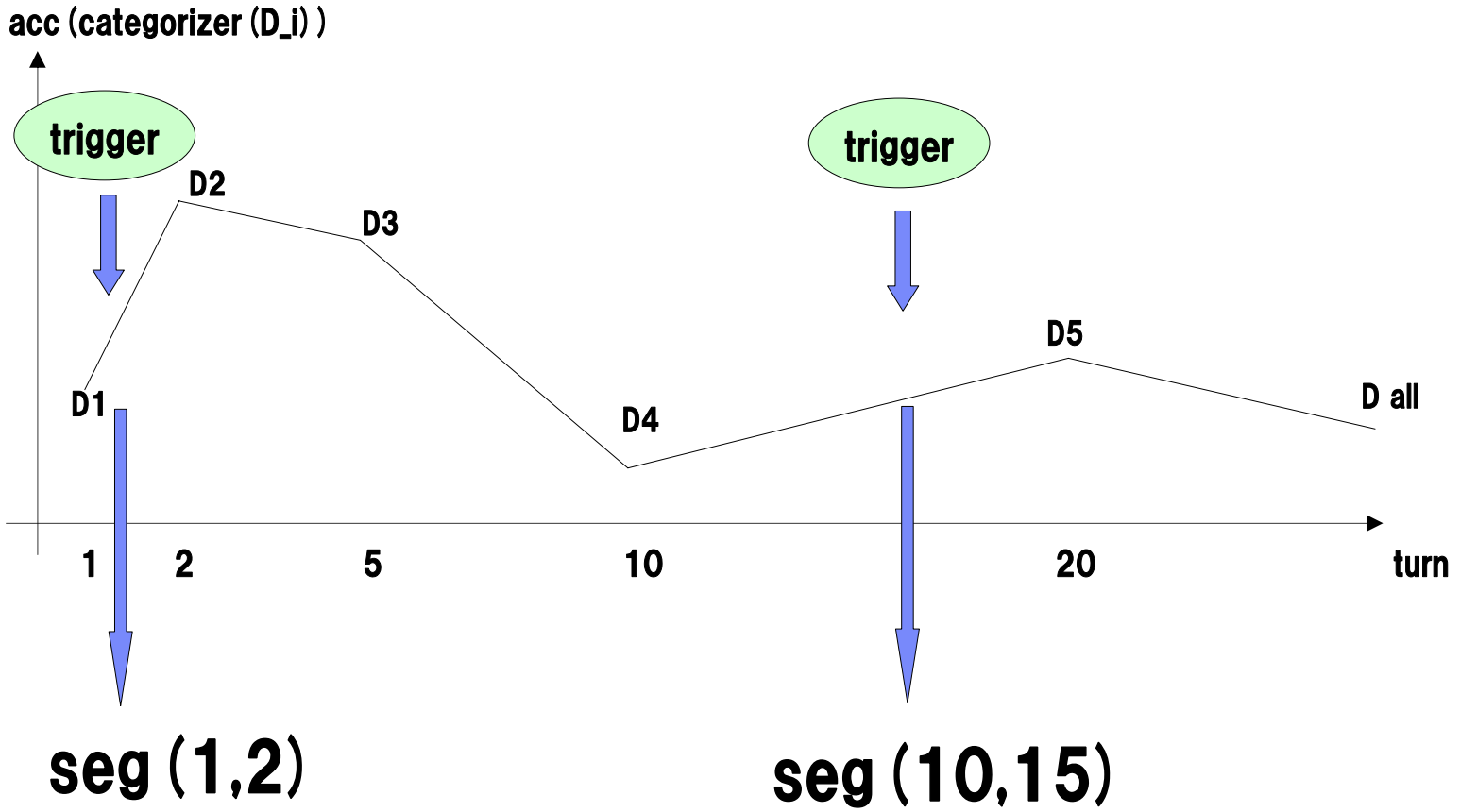
〔会話データのテキストマイニング〕
ノイズの割合と元データの特徴の検出度合いの関係

	自動音声認識システムの辞書から選択した語との置換によるノイズ		書き起こしデータ中の出現分布を考慮して置換した語によるノイズ		書き起こしデータ中で出現頻度3以上の語と置換した語によるノイズ	
ノイズの割合	*debit card* の相関強度（頻度）	*debit* の相関強度（頻度）	*debit card* の相関強度（頻度）	*debit* の相関強度（頻度）	*debit card* の相関強度（頻度）	*debit* の相関強度（頻度）
0 %	5.3（13）	5.0（14）	5.3（13）	5.0（14）	5.3（13）	5.0（14）
10 %	5.8（13）	5.3（14）	5.1（12）	4.2（14）	4.7（11）	4.1（13）
20 %	4.5（10）	4.9（13）	4.3（10）	4.1（13）	4.2（9）	4.0（14）
30 %	4.4（8）	4.6（12）	4.4（9）	3.5（13）	3.5（8）	3.8（13）
40 %	2.8（6）	5.3（13）	2.8（6）	2.5（12）	1.4（4）	2.3（8）
50 %	1.2（3）	2.7（7）	3.3（6）	2.5（11）	4.3（7）	3.6（12）
60 %	0.0（1）	3.0（7）	0.1（1）	1.3（8）	2.4（3）	2.2（8）
70 %	-	0.7（3）	1.9（3）	1.8（10）	-	1.1（6）

〔会話データのテキストマイニング〕 結論

- 会話データのテキストマイニングの有益性の検討
  - 生の会話全てをテキスト化したデータの分析により有益な知見を獲得できる例を確認
- 現実的な活用可能性の検討
  - 50％程度のノイズを含んだデータからも有益な分析が可能な例を確認

会話データのテキストマイニングの高度化に関する研究成果（1/2）：分析に有効な領域（trigger segment）の自動認識



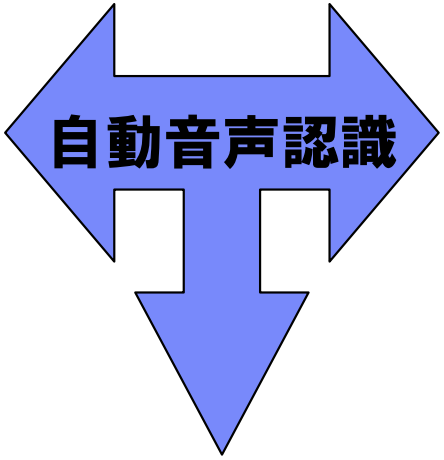
各trigger segmentごとに有効な表現を抽出

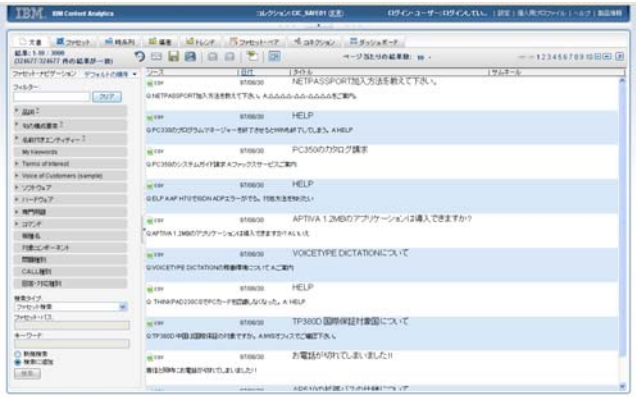
会話データのテキストマイニングの高度化に関する研究成果（2/2）：有効表現の抽出結果

Trigger	Selected Expressions	
	pick up	not picked up
seg (1,2)	make, return, tomorrow, day, airport, look, assist, reservation, tonight	rate, check, see, want, week
seg (10.15)	number, corporate program, contract, card, have, tax surcharge, just <NUM> dollars, discount, *customer* club, good rate, econom	

- **最初のcustomerの発言が重要**
- **discountや提供料金の良さへの言及が重要**

〔会話データのテキストマイニング〕
コールセンターにおけるテキストマイニングの将来像





自動音声認識によりテキスト化された会話のテキストマイニング

テキスト化された音声データの活用可能性

テキスト化された音声データによる価値創出の可能性

- 生の声を生かせる魅力
  - –企業は顧客の声を把握したがっている
  - –企業だけでなく自治体等でも住民の声の把握は重要課題
  - ➔顧客対応記録やソーシャルメディアのテキストマイニングが伸びている背景

- より幅広く、より自然な声を把握できる魅力
  - –ネットへの書き込みをする人は限られている
  - –バイアスのかかるアンケートなどでは真の声の把握が困難
  - –老若男女の自然な声が把握できれば大きな成果の可能性

➔人々の声を適切に活かすことでより良い世の中へ
  - –需要と供給のミスマッチを減らすことで無駄を排除
  - –より良いサービスを実現
  - –他

テキスト化された音声データとテキストマイニングの親和性

- 量の多さ
  - 膨大な音声データからは、
    人手ではとても扱えない量のテキストデータが生成される
  - 膨大な量のテキストデータを活用するための技術が
    テキストマイニング
- ノイズの許容性
  - データ及び言語処理が完全でないことを前提に
    分布の偏りや変化に着目するのがテキストマイニングのポイント
  - 認識エラーが含まれていても、有用な気付きにつながる可能性

➔テキストマイニングは、テキスト化された音声データから価値を創出するための有望な技術の一つ

まとめ

- テキストマイニングとは
  - 個々のテキストを読んだだけでは得られない知見を獲得する技術
  - 件数分布でなく分布の偏りや変化に着目し、得られた知見をアクションにつなげることが重要
- 会話データのテキストマイニング
  - 生の会話全てをテキスト化したデータの分析の有効性を確認
  - 50％程度のノイズを含んだデータからも有益な分析が可能な例を確認
- テキスト化された音声データの活用可能性
  - テキスト化された音声データに対する大きな期待
  - テキストマイニングは、テキスト化された音声データから価値を創出するための有望な技術の一つ

参考文献

- 那須川哲哉. テキストマイニングを使う技術/作る技術―基礎技術と適用事例から導く本質と活用法. 東京電機大学出版局, 2006
- Tetsuya Nasukawa and Tohru Nagano, "Text analysis and knowledge mining system", IBM Systems Journal, Vol.40, No.4, pp.967 -984, 2001.
- 那須川哲哉、宅間大介、竹内広宜、荻野紫穂, "コールセンターにおける会話マイニング", 言語処理学会 第13回年次大会,2007.
- Hironori Takeuchi, L Venkata Subramaniam, Tetsuya Nasukawa, Shourya Roy, "Getting insights from the voices of customers: Conversation mining at a contact center", Information Sciences, 179 (11) , 1584-1591, 2009.
- 竹内広宜, 那須川哲哉, 渡辺日出雄, "コールセンターにおける目的をもったビジネス会話のマイニング",人工知能学会論文誌, 23 (6) , 384--391, 2008.