

**NEC、大量の文書データを同じ意味で
自動的にグループ化できる技術を開発
～素早い概要把握で、マーケティングからリスク管理まで支援～**

**2014/11/18
日本電気株式会社**

ビッグデータ活用への期待

顧客の「期待」は4領域に集約
テキストデータを含めたビッグデータの活用が期待されている

オペレーション高度化/最適化

- プラントなど大規模施設異常の予兆監視や保全
- 自動車など輸送機器の品質の向上
- 社会インフラの安全管理の高度化

情報管理の強化、犯罪・不正の検知

- 企業内の要管理情報の検出による情報ガバナンス強化
- 金融取引の不正検知
- 治安維持・監視の強化

製品/サービス価値向上・改善

- 商品の需要予測による生産・発注の最適化
- 電力需要予測によるビルや都市のスマート化
- 適正価格予測による利益拡大

顧客獲得・維持、販売促進

- 商材・サービスのマッチングによる販売促進
- ポイント・クーポンの効果分析による顧客獲得
- カメラ画像を使用した動線分析によるマーケティング強化

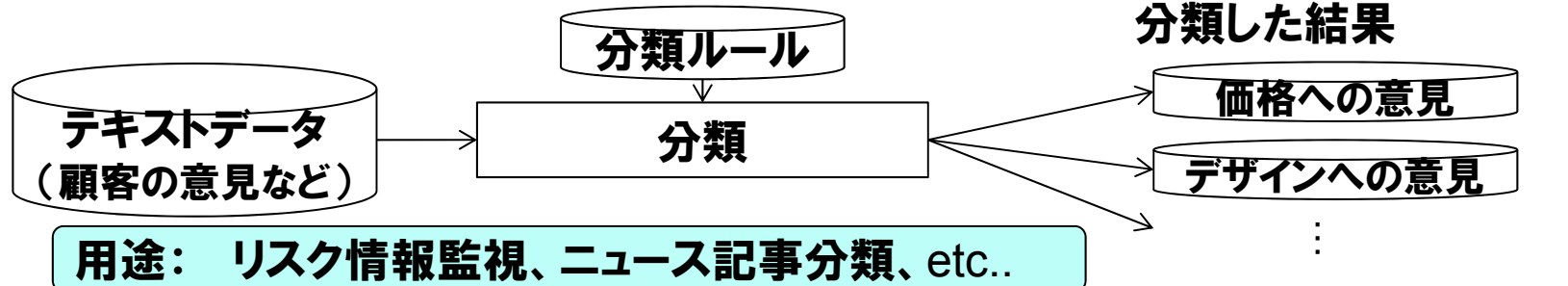
※当社、顧客商談内容などから類型化

大量のテキストデータの活用方法(現状)

大量テキストデータの活用に欠かせない「自動仕分け」の2つのアプローチ

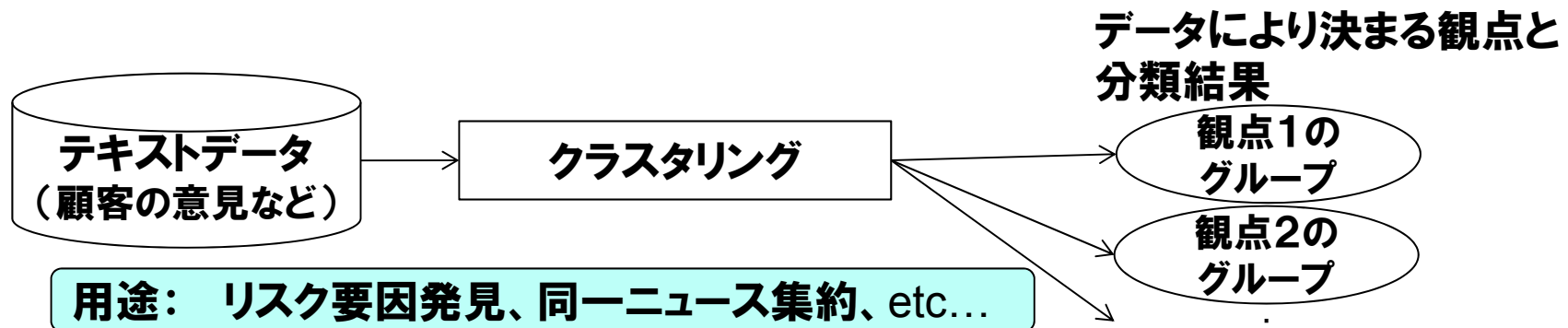
●分類

- ・ ユーザがあらかじめ決めた観点に自動仕分けする
- ・ データの見方が決まっている場合に用いる



●クラスタリング

- ・ ユーザがあらかじめ観点を決めず、機械が観点の決定と仕分けを同時に実施
- ・ データの見方が定まっていない、目的上定めるべきでない場合に用いる



テキストデータ活用の課題(1)

見るべきデータの抽出が困難

1. 設定したキーワードに関連する観点を拾うだけでは、新しい発見が得られない、漏れる
例: 新規トラブル、新発表情報、など
2. ひとつの文でも、いろいろな観点が混在
例: 「価格は高いけど、デザインはいいね」
それぞれの要素を、漏れ無く見たい
3. 分析対象としたくない、情報も記載されている
例: 挨拶文など



クラスタリングが必要 かつ、関係の薄いデータの除外が必須

テキストデータ活用の課題(2)

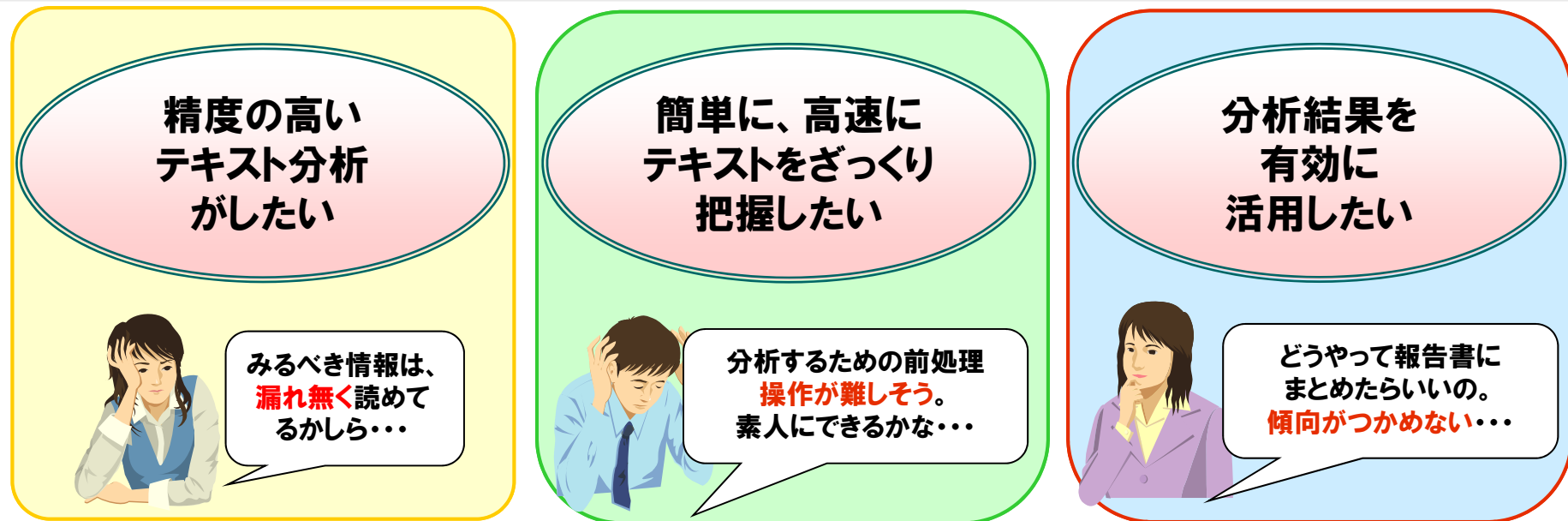
分析内容のタイムリーな活用が難しい

1. 使いこなすための専門知識・工数が必要
例： 分類するためのルール作りに、10人がかり・・・
2. 結果を概観しにくい
例： 結局、全文を読んで、人間がラベル付
3. 分析結果の根拠を調べるのが困難
例： 単語の羅列では、何を意味するのか、わからない



ざっくりと 高速に、傾向を把握したい

意味でテキストを自動グルーピングする「含意クラスタリング技術」



自動でグルーピング、分類観点(代表文)を抽出

「含意クラスタリング技術」を開発

NEW

1. どんな話題があるか、もれなく観点を抽出するので安心
2. 分析結果に、コンパクトな代表文がつくので、業務活用しやすい

含意クラスタリング技術のポイント

■ 全ての文を対象に、同じ意味を持つテキストを同じグループに分類しつつ、グループが共通して含む「意味」を代表文として抽出

1. どんな話題があるか、もれなく観点を抽出するので安心
2. 分析結果に、コンパクトな代表文がつくので、業務活用しやすい

分析結果イメージ:

1. 全テキストを意味毎のグループに分類することで、漏れなく観点(意味)を考慮

2. グループが共通で含む意味の文がコンパクトな代表文となる

価格が高い:62件

...
価格が高い
値段が高い

...
高価格だけどデザインが良い
値段が高いが、価値はある

...

デザインが良い:112件

...
デザインが良い
デザインが最高

...
高価格だけどデザインが良い
デザインも良いが使用感も良い

...

【参考】NECのテキスト含意認識技術 RTE: Recognizing Textual Entailment

テキスト含意認識とは二つの文が同じ意味を含むか判定するタスク

NECは文中の単語の重要性、主語・述語等の文構造を考慮する独自の技術で**グローバルトップレベルの精度**

- アメリカ国立標準技術研究所(NIST)主催の評価タスク、TAC2011-RTE7で**第一位**
- 日本の国立情報学研究所(NII)主催の評価タスク、NTCIR10-RITE2のEntrance Exam Taskの2つのタスクで**第一位、第二位**

元の文	比較対象	既存方式 (単語の一致/不一致)	NEC方式 (単語重要性、文構造)
私はリンゴが好きだ	単語: 同じ 構造: 同じ 私はリンゴが好きだ	○ 正しく判定	○ 正しく判定
	単語: 違う 構造: 同じ 私はアップルが好物だ	× 誤って判定	○ 正しく判定
	単語: 同じ 構造: 違う 彼はリンゴが好きだが 私は嫌いだ	× 誤って判定	○ 正しく判定

NIST : National Institute of Standards and Technology
TAC : Text Analysis Conference
RTE : Recognizing Text Entailment

含意
含意
非含意

従来のクラスタリング技術との違い

従来（キーワードの共有度合い）

- ・複数の観点が混じるため、観点が見落とされやすい
- ・グループの観点が不明確

グループ1:52件

値段は高いがデザインが良い
...
サイズはあるが値段は高い
...
サイズに合った服が良いが無い
...

グループN:302件

値段が高い
値段が高いとは思わない
値段が高いが、価値はある
...
値段が安いのが価値だ
...

結局どういうことだ？
原文で再整理しよう。



含意クラスタリング(文意)

- ・意味の重複なく、全観点に対するグループを生成
- ・代表文がグループの観点となるため明確

価格が高い:62件

価格が高い
値段が高い
...
高価格だけどデザインが良い
値段が高いが、価値はある
...

デザインが良い:112件

デザインが良い
デザインが最高
...
高価格だけどデザインが良い
デザインも良いが使用感も良い
...

デザインの反応は
思ったより良いな。
価格への不満は
予想ほどではないな。



適用例： 顧客の苦情に基づく品質管理

■ サービスの品質改善策を検討するために、1万件の苦情の概要を把握したい

■ 人手による分類作業は約2分/件とする

・既存技術では一つのグループに複数の観点が混ざり、グループ単位の再分類ができない
・そのため、半数(5000件)を、人手で既定された観点到分類することで、概要を把握する
⇒ 約5日
・既定観点到分類困難なものは「その他」



・グループのメンバは代表文の意味を含むため、件数の上位500グループの代表文に基づき、グループ毎に既定された観点到分類
⇒ 約4時間（約10倍の効率化）
・既定観点到分類困難ならば新規観点候補
⇒ **新規観点的半自動発見**

他の利用シーンの例

苦情からのリコール
製品の早期発見

ソーシャルメディア
からのトレンド把握

消防・自治体に
寄せられる災害情報
の自動整理

デモ： 自動車の不具合情報の分析

「国土交通省 自動車のリコール・不具合情報」のテキストを分析した、件数トップ5グループ

含意クラスタリング

端的でわかりやすく、どのような不具合が多いかわかる

グルーピングされる文は代表文の意味を含んでいる

異音が出る。	563
エンジンが停止した。	542
走行不能！	417
ブレーキが利かなくなった。	412
警告灯が点灯した。	400

CVT(無段変速機)の不具合により、寒い時のエンジン始動直後に変速機から異音が出る。
高速道路を走行中にガタンという異音がし、そのまま走行してみると、ガタガタと異音が出るため、最寄りのタイヤ販売店で確認したところ、右前輪のドライブシャフトが外れていた。
ウォーターポンプから異音が出るようになった。
ステアリングシャフトが固定されているバルクヘッドの溶接が外れたため、ハンドルを切った後に、ハンドルを戻しても角度が元に戻らず、ステアリングを切ると異音が出るようになった。
後輪のホイールベアリングが損傷したため、異音が出るようになった。

従来技術（キーワードの重み付き共有度を類似度としたRepeated Bisection法）

端的でなく、ポイントがわからない

代表文と無関係の文がグルーピングされている

その後、変速しなくなった。	79
エアコンプレッサーの不良により、エアコンが効かなくなった。	76
エンジン始動後2～3分以内でブレーキをかけると、通常よりブレーキが利かず、停止距離が伸びる事がある。	60
ABSユニットの不具合により、積算距離計の異常が出た。	58
エアフロメーターの内部不良により、ゆるい坂道の信号停止において、ブレーキをかけた際に、エンジンが止まっており、全然ブレーキが利かなかったため、電柱にぶつけて停車した。	49

フライホイールのスプリング部分の不良によりクラッチが正常に機能しなくなった。
燃料ポンプのリターンホースかしめ部の不良により、燃料漏れが発生した。
タイヤの不良により、車両がまっすぐ走らない。
リヤデファレンシャルオイルシールの不良により、オイル漏れが発生した。
キーシリンダーの内部不良により、シリンダーがACC位置よりON側へ回らない。
エアコンのコンプレッサー、コントロールユニット、ソレノイドの不良により、エアコンが利かなくなった。
スロットルセンサーの不良により、エンジンストールを起こす。
シートベルト巻き取り構造の不良により、緩んだ状態からスムーズに巻き取られない。
スロットルアクチュエーターの不良により、アクセルを踏んでも加速しなくなった

まとめ

■ 大量のテキストデータを意味ごとのグループに分類し、グループごとの意味を端的に提示可能な「含意クラスタリング技術」を開発

- 1. どんな話題があるか、もれなく観点を抽出するので安心**
- 2. 分析結果に、コンパクトな代表文がつくので、業務活用しやすい**

■ 本技術は、従来技術では困難であった明確な意味を持つ分類の観pointsの列挙、その観点へのグルーピングが可能になるため、様々な応用が可能

- リスク管理 : 苦情からのリコール製品の早期発見**
- マーケティング : ソーシャルメディアからの新しいトレンドの把握**
- 公安・防災 : 様々な情報源からの事件・災害情報の自動整理**

■ 今後、本技術を活用した製品・サービスを通じ、ビッグデータからのお客様の新たな価値の創出に貢献

Orchestrating a brighter world

世界の想いを、未来へつなげる。

未来に向かい、人が生きる、豊かに生きるために欠かせないもの。
それは「安全」「安心」「効率」「公平」という価値が実現された社会です。

NECは、ネットワーク技術とコンピューティング技術をあわせ持つ
類のないインテグレーターとしてリーダーシップを発揮し、
卓越した技術とさまざまな知見やアイデアを融合することで、
世界の国々や地域の人々と協奏しながら、
明るく希望に満ちた暮らしと社会を実現し、未来につなげていきます。



Empowered by Innovation

NEC