

# データマイニング

- テキストマイニング
- 自然言語処理
- 機械学習
- 情報検索

第1週目

# データマイニング 入門

# データマイニングとは

多量のデータ(ビッグデータ)から有用な知識を発掘する技術の総称



- 大容量記憶媒体の低価格化や計算機処理能力の向上によって、膨大な量のデータの収集が容易

データだけでは、有効に活用できない



データから有用な知識を見たい

# 有用な知識とは...



例: 大量のPOS<sup>(※)</sup>データからの知識発見

□ 「紙おむつ」と「ビール」の例

大量のPOSデータを分析したところ、紙おむつを買う人は同時にビールをよく買うという傾向が分かった。



紙おむつの隣にビールを置く。

有用な知識



Xを買えばYも買う。



(※)販売時点情報管理システム。商品の売り上げの情報を、販売の時点でリアルタイムに収集

# データマイニングの定義

- 大量のデータから有用な知識を発掘(マイニング)する
  - 雜多な鉱石を含む鉱脈(大量のデータ)から、貴重な鉱物(有用な知識)を掘り当てるのは困難
  - 貴重な鉱物がどこにあるか、そもそも存在するのかも不明
  - 計算機を使用して発掘(大量のデータを扱えるうえに、それらを人間よりはるかに高速に処理できる)
- 知識を計算機を使って機械的に発掘
  - データに内在する規則や特徴的なパターンをパターン発見アルゴリズムを使用して抽出



# テキストマイニング

大量のテキスト(新聞記事、メール、Web上の掲示板等)を対象としたデータマイニング

- 自由記述アンケートの自動分析
- メールの自動分類(迷惑メールの除去等)
- 有害情報フィルタリング

自然言語処理技術(形態素解析や情報検索等)とデータマイニング技術を結合

# 自然言語処理技術(1)

## ■ 形態素解析

入力文における、意味を担う最小の言語要素  
(形態素)を同定する処理

例:

「昨日学校へ行った」



表記	原型	品詞
昨日	昨日	副詞
学校	学校	名詞
へ	へ	助詞
行つ	行く	動詞
た	た	助動詞

# 自然言語処理技術(2)

## ■ 固有表現抽出

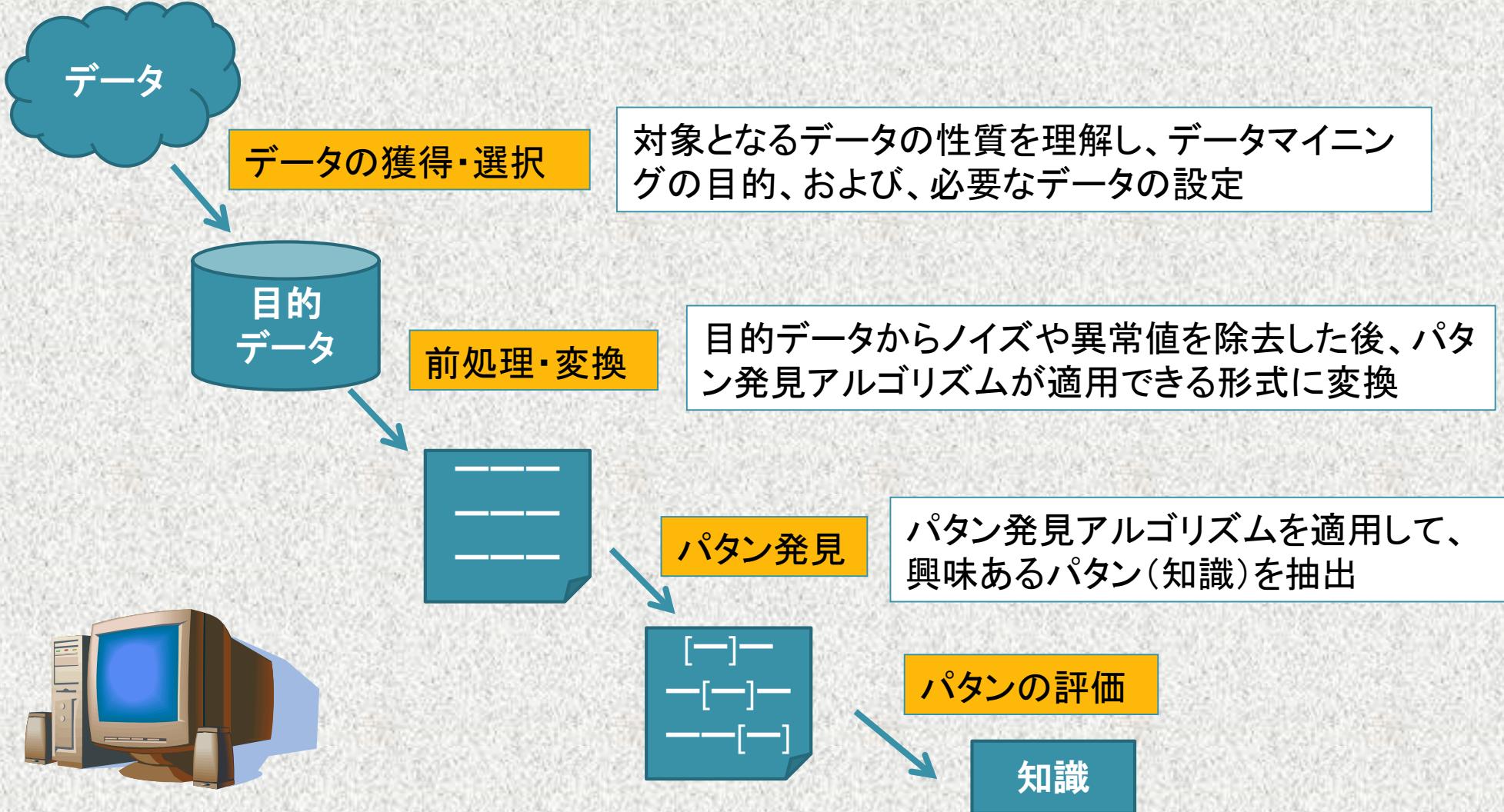
テキストに出現する人名・地名・組織名・日付・時刻などを高精度に同定

例:「小泉首相は二十四日首相官邸で会見し...」



「<PERSON>小泉</PERSON>首相は  
<DATE>二十四日</DATE>  
<LOCATION>首相官邸</LOCATION>で会見し...」

# データマイニングのプロセス

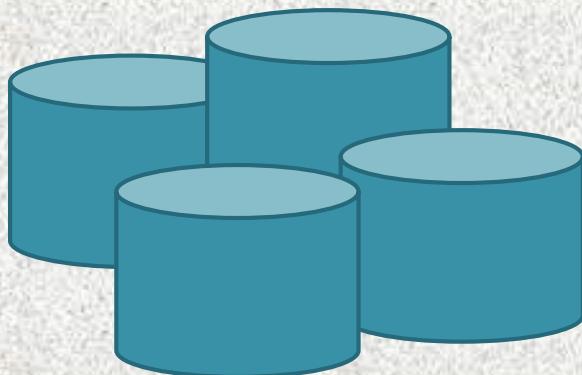


# データの獲得・選択

- 目的の設定
- 目的に沿ったデータの獲得・選択

## 目的(例)

交通事故を扱った新聞記事から、事故原因（「前方不注意」など）に言及している部分を抽出して分析



新聞記事の集合



交通事故を扱った  
記事の集合

# 前処理・変換

- 目的データからノイズや異常値を除去
- パタン発見アルゴリズムが適用できる形式に変換

例： 交通事故を扱った新聞記事を、単語に分割

町田署ではAさんのスピードの出し過ぎが原因とみて調べている。



町田\_\_署\_\_で\_\_は\_\_A\_\_さん\_\_の\_\_スピード\_\_の  
出し\_\_過ぎ\_\_が\_\_原因\_\_と\_\_みて\_\_調べ\_\_て\_\_いる。

# パタン発見

- パタン発見アルゴリズムを適用して、興味ある  
パタン(知識)を抽出

例： 交通事故を扱った新聞記事から、原因について  
言及している部分を抽出

三十一日午前二時四十分ごろ、町田市鶴間の国道16号線内回りで、  
横浜市都筑区すみれが丘、自営業Aさん運転の乗用車が中央分離  
帯に乗り上げ横転、弾みで左側歩道のコンクリート壁に激突した。町  
田署ではAさんのスピードの出し過ぎが原因とみて調べている。

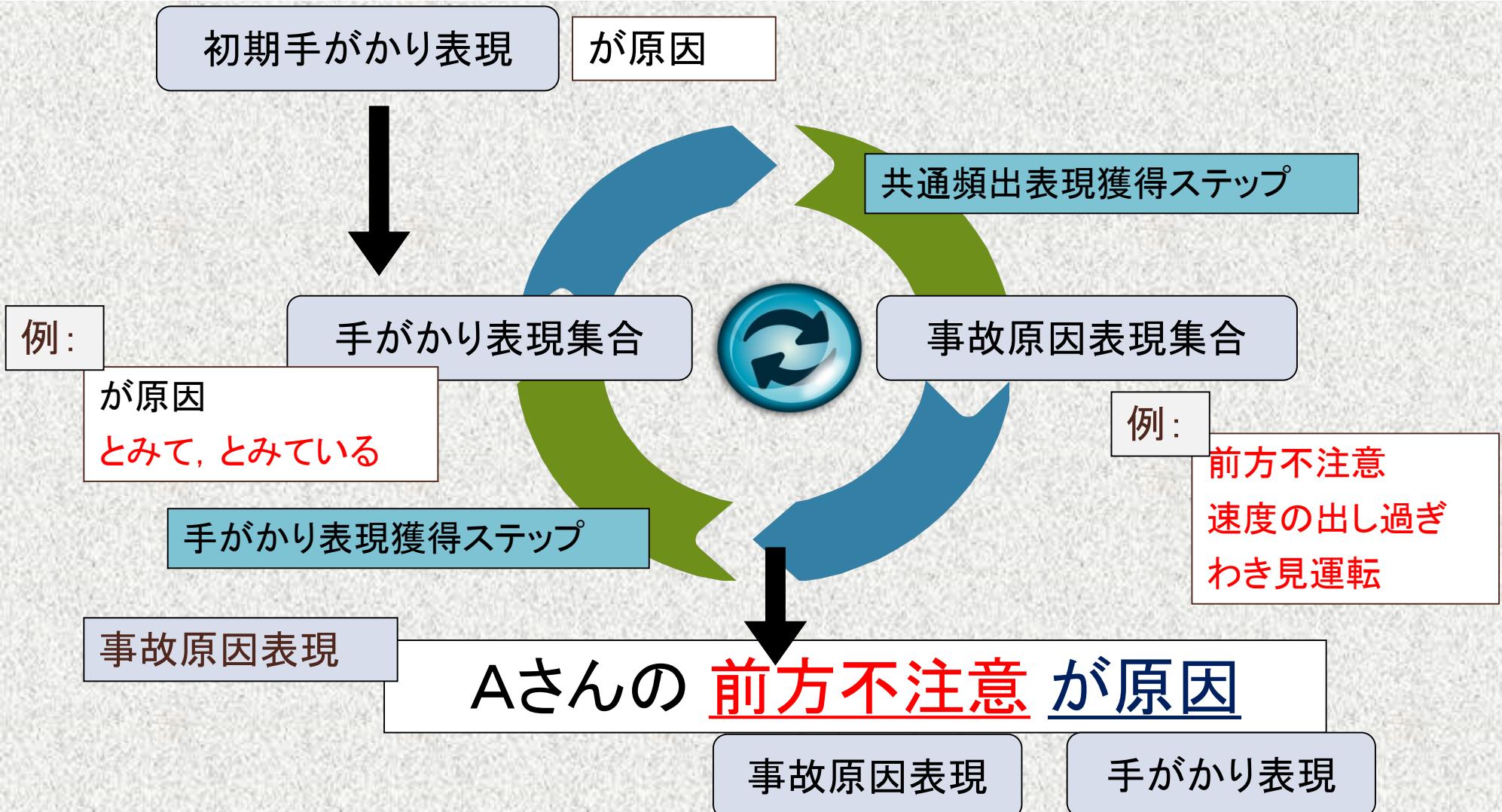


事故原因：スピードの出し過ぎ

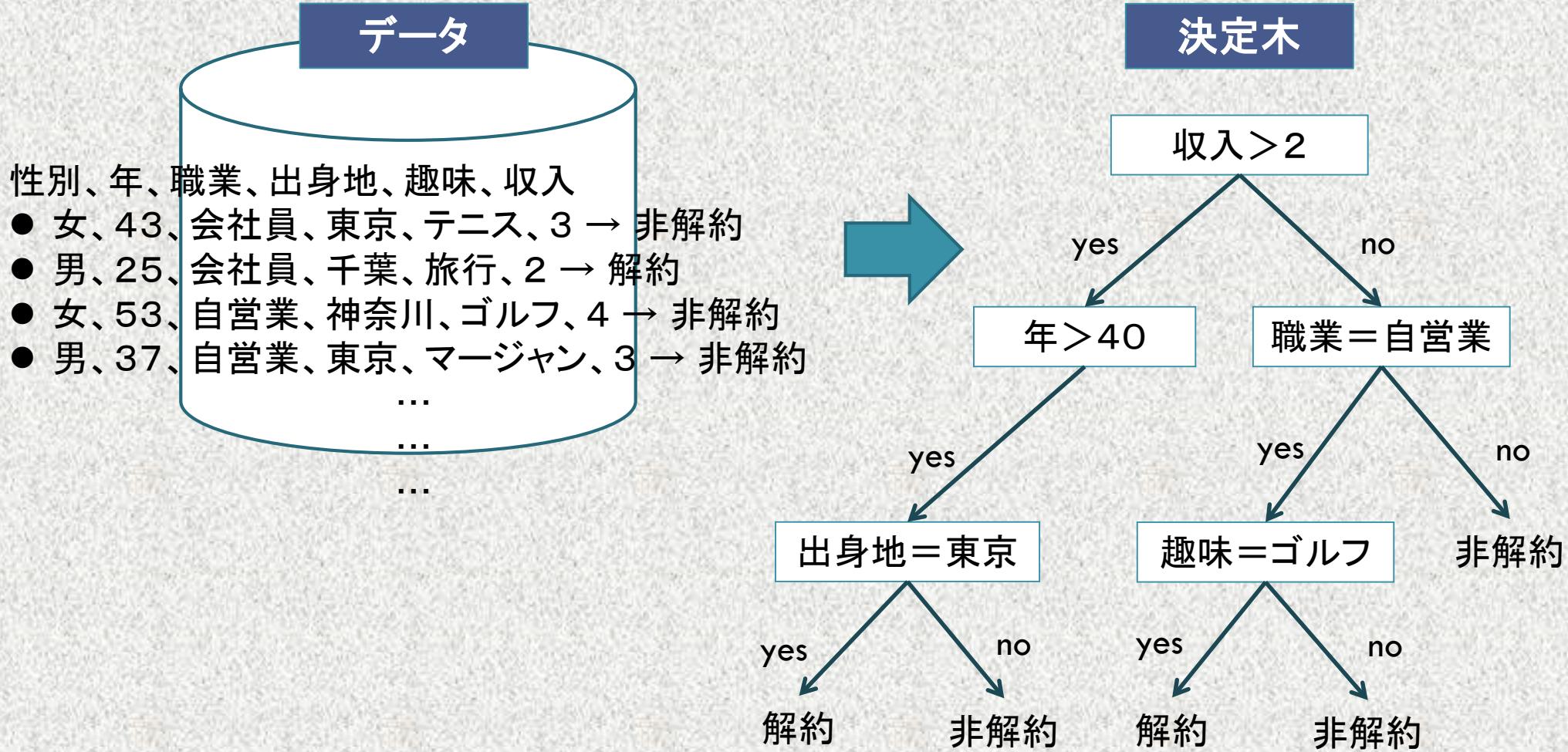
# パターン発見アルゴリズム

- 決定木
- ナイーブ・ベイズ
- 最近傍法
- サポートベクトルマシン
- 深層学習(ディープラーニング)
- k-means法
- 目的に特化した独自手法

# 事故原因表現抽出手法の概要



# 決定木



# ナイーブ・ベイズ(1/3)

- 確率モデルに基づく分類手法
  - 迷惑メールのフィルタ等に使用(メールを正常なメールと迷惑メールに分類)
    - ・文書(メール)  $d = (w_1, w_2, w_3, \dots, w_n)$   $w$ : 単語
    - ・文書のクラス  $c$  = (正常なメール, 迷惑メール)

$$\hat{c} = argmax_c P(c|d)$$

# ナイーブ・ベイズ(2/3)

$$\hat{c} = \operatorname{argmax}_c P(c|d)$$

$$= \operatorname{argmax}_c \frac{P(d|c)P(c)}{P(d)} = \operatorname{argmax}_c \underline{\underline{P(d|c)P(c)}}$$



文書クラスcが与えられたときの文書dの出現確率は、  
様々な文書が存在するため、推定することは困難

- ・文書(メール)  $d = (w_1, w_2, w_3, \dots, w_n)$   $w_i$ : 単語
- ・文書のクラス  $c = (\text{正常なメール}, \text{迷惑メール})$

# ナイーブ・ベイズ(3/3)

- 文書dに含まれる各単語 $w$ が互いに独立に生起すると仮定

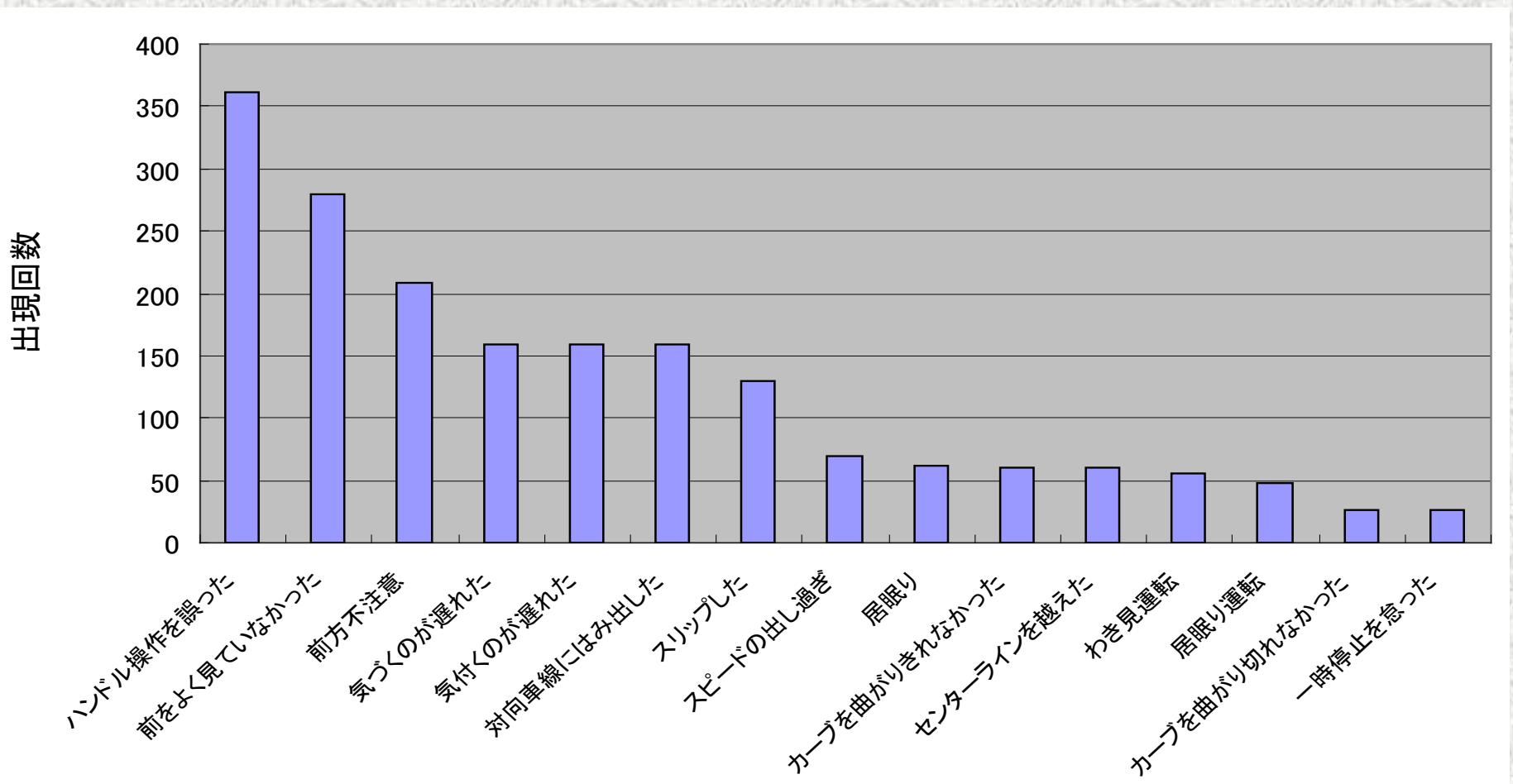
$$P(d|c) = P(w_1, w_2, w_3, \dots, w_n | c) \approx \prod_{i=1}^n P(w_i | c)$$



正常なメールと迷惑メールが識別されたメールの集合  
(学習データ)があれば、推定可能

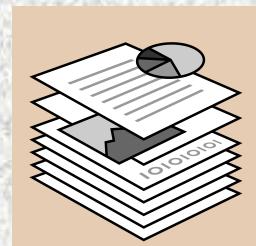
$$\hat{c} = \operatorname{argmax}_c P(c) \prod_{i=1}^n P(w_i | c)$$

# 交通事故原因の傾向分析



# データマイニングの実用化の現状

- POSデータから、販売促進知識(併売パターン)の獲得
- 新製品の売れ行き要因の自動分析(評判情報分析)
- 顧客意見(自由記述形式のアンケート)や製品クレームの自動分析による新製品開発への反映
- スパムメール(迷惑メール)の自動判定
- Webページ閲覧履歴から、ユーザの嗜好の調査やページ自動推薦
- 過去の株価推移からの投資戦略の策定



# 金融とテキストマイニング

21

- ・証券市場における個人投資家の比重が増大しており、個人投資家に対して投資判断の支援をおこなう技術の必要性が増加
- ・投資信託、証券、銀行等の金融系企業においては、日々、大量の金融テキスト（決算短信やアナリストレポート、経済新聞、日銀や政府の金融政策）を分析し、企業調査や経済動向の予測に活用

テキストマイニングを金融市场における様々な場面に応用し、業務支援を行いたい

株式投資で企業についての情報を効率よく集め、  
~~儲けたい~~効率のよい投資がしたい

# 株式投資と企業分析

22

個人投資家にとって、その企業の事業内容を分析し、将来、有望な技術を研究、開発している企業を調べて投資することは、敷居が高い

→ 容易にできるようになれば、企業価値の向上も期待できる

例：

「燃料電池」を研究、開発している企業に投資したい



- ・ 日本触媒 → 燃料電池材料
- ・ 日清紡ホールディングス → 燃料電池用力一ボンセパレータ
- ・ 日邦産業 → 燃料電池用部材

# 金融テキストと企業分析

23

個人投資家にとって、その企業の事業内容を分析し、将来、有望な技術を研究、開発している企業を調べて投資することは、敷居が高い

→ 容易にできるようになれば、企業価値の向上も期待できる



投資判断の支援のための企業分析を行うために情報源として、決算短信、有価証券報告書などの**金融テキスト**が有望

(その他にも、経済新聞記事やWEB上のニュース、企業WEBページ等が有望)

# 金融テキストを用いた研究事例(1)

24

## ➤ 決算短信

- 企業の決算短信からの業績要因の抽出
- 決算短信からの原因・結果表現の抽出
- 決算短信からの業績予測文の抽出
- 決算短信から抽出した原因・結果表現の意外性の判定

## ➤ 有価証券報告書

- 有価証券報告書から、事業セグメントごとの業績要因、業績結果を抽出

# 金融テキストを用いた研究事例(2)

25

- 株主招集通知
  - ・ 株主招集通知における議案タイトルとその分類及び開始ページの推定
- 企業WEBページ
  - ・ 企業Webページを用いた関連企業の推定
- 経済新聞記事
  - ・ 経済テキストからの市況分析コメントの自動生成
- アナリストレポート
  - ・ アナリストレポートからのアナリスト予想根拠情報の抽出と極性付与

# 重要な業績要因の自動抽出・判定

企業の業績発表に関する記事から、業績要因が記述してある部分を自動的に識別、抽出し、その業績要因に対して極性(Positive, Negative)、重要度(★印で表現)を自動付与

例:

- 液晶ディスプレー向けガラス基板の好調が寄与  
(Positive, ★★★) ←「旭硝子」の業績発表記事より
- 収益性が高いカードゲームのブームー巡が響いた  
(Negative, ★★★) ←「コナミ」の業績発表記事より

# 重要な業績要因の自動抽出・判定

企業の業績発表に関する記事から、業績要因が記述してある部分を自動的に識別、抽出し、その業績要因に対して極性(Positive, Negative)、重要度(★印で表現)を自動付与

## この研究が必要な理由

ある企業にとって特に重要な業績要因をコンピュータが自動的に判定し、提示できれば、その企業についての高度な専門知識がない個人投資家に対する投資判断支援を行うための有用な情報源になる

# Causal Expressions Extraction System

http://133.220.112.11/CEES/easy\_search.cgi?0  
CEES(Causal Expression ...)

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)  
× Canon | Easy-WebPrint EX | 印刷 | プレビュー | クリップ | 自動クリップ | クリップ

Causal Expressions Extraction System  
CEES

エアコン 検索 クリア

抽出した業績要因に対してキーワードで検索できるシステム

081114-0054 ノジマ  
↑ 新興——ノジマ4—9月、経常益6%増、家電販売が堅調。

081107-0046 ダイキン工業  
↓ ダイキン純利益19%減、今期610億円、欧州でエアコン不振。

081030-0056 三菱重工業  
↑ 三菱重、経常益横ばい、今期、上期の円安効果大きく。

081010-0054 三菱電機  
↑ 三菱電機、営業益1250億円、4—9月3%減、計画を150億円上回る。

080927-0065 富士通ゼネラル  
↓ 富士通ゼ、営業益11%減、今期91億円。

080812-0049 ラオックス  
↑ ラオックス4—6日 最終赤字23億円

「エアコン」を業績要因とする企業の業績発表記事を検索

↑ は業績良好の業績発表記事、  
↓ は業績悪化の業績発表記事  
(自動判別)

# Causal Expressions Extraction System

http://133.220.112.11/CEES/cees.cgi?070428-0090

CEES(Causal Expression ...)

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

× Canon | Easy-WebPrint EX | 印刷 | プレビュー | クリップ | 自動クリップ | クリップリスト

Causal Expressions Extraction System

## CEES

記事から業績要因を識別し、極性付与(↑, ↓)、重要度付与(★★★)

三菱電の営業益最高、前期17年ぶり、FA機器など寄与。

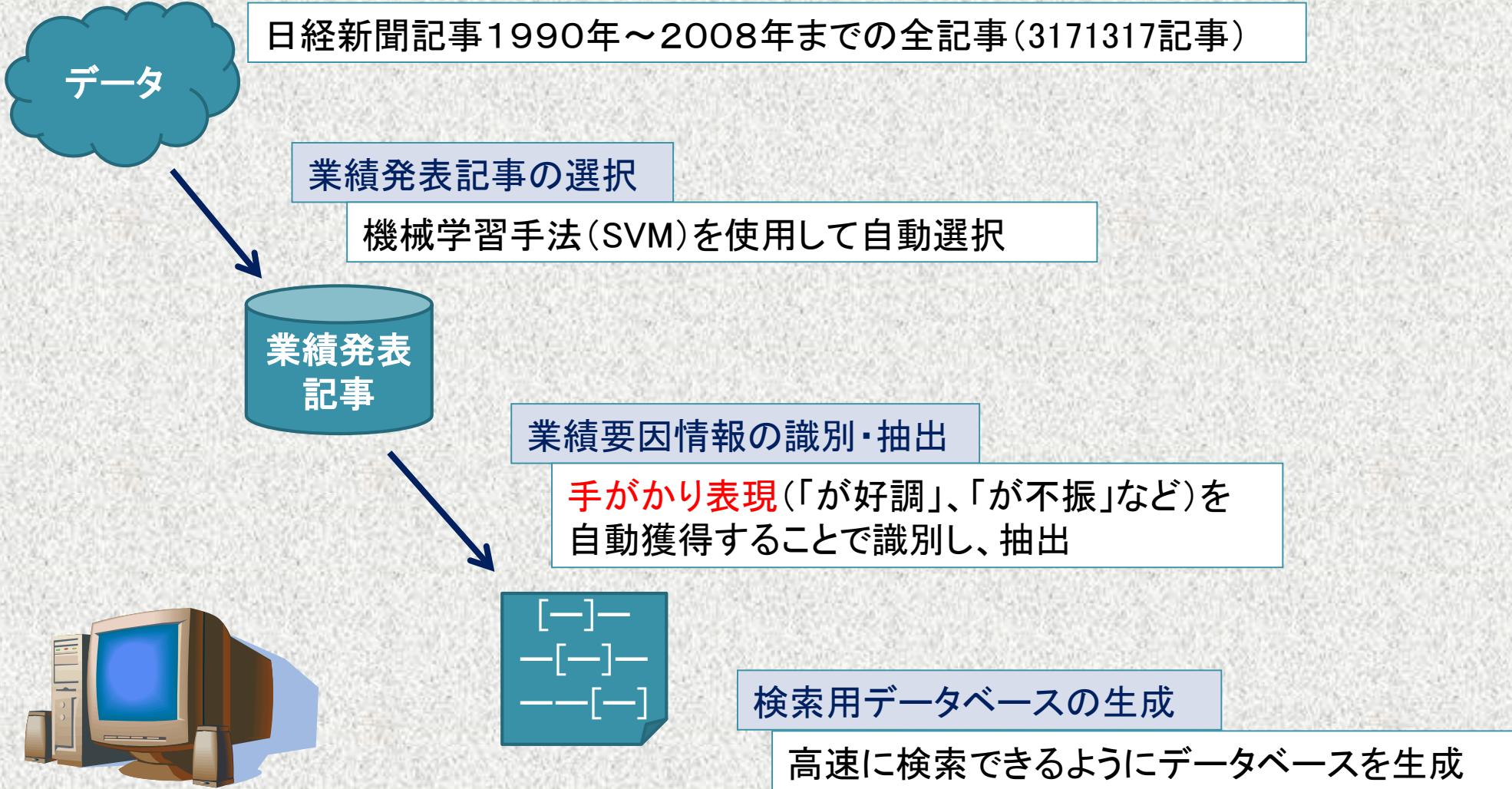
三菱電機が二十七日発表した二〇〇七年三月期の連結決算は営業利益が前の期比四八%増の二千三百三十億円と十七年ぶりに過去最高を更新した。設備投資関連のファクトリーオートメーション機器、エアコンの好調が寄与した。〇八年三月期は薄型テレビの設備投資減速でFAが伸び悩み、営業利益は前期比一四%減る見通し。前期の売上高は七%増の三兆八千五百五十七億円。FAを含む産業メカトロニクス、情報通信システムなど六つの部門すべてが増収となった。売上高営業利益率は六・〇%と中期目標の五%を超えた。独禁法関連費用四百二十一億円を営業外費用に計上したが、税引き前利益は千八百四十七億円と二一%増えた。費用の対象はDRAMと、送電・遮断の調節に使うガス絶縁開閉装置の二分野。三菱電はエレベーターでも欧州委員会から価格カルテルを結んだとし制裁金支払いを求められていたが、会見した佐藤行弘副社長は「制裁金は円換算で二億八千万円。弁護士費用を考えると応じるのが合理的」とし、一ヶ月以内に命令に応じる考えを示した。〇八年三月期は売上高が二%増にとどまる見込み。「薄型テレビなどで投資先送りの動きが増え、昨年後半からFAの受注が減少に転じている」。為替動向や素材価格の上昇も響く。

↑ ★★★ 設備投資関連のファクトリーオートメーション機器、エアコンの好調が寄与した。  
↓ ★★ 〇八年三月期は薄型テレビの設備投資減速でFAが伸び悩み、  
独禁法関連費用四百二十一億円を営業外費用に計上したが、  
★ DRAMと、送電・遮断の調節に使う  
「薄型テレビなどで投資先送りの動きが増え、  
為替動向や素材価格の上昇も響く。

↑ は業績良好の業績要因、  
↓ は業績悪化の業績要因  
(自動判別)

→ ★の数で重要度を表現  
(★★★が最重要)

# CEES のプロセス



# 手がかり表現の自動獲得

## ■ 初期手がかり表現

が好調, が不振



## ■ 研究室で公開している手がかり表現自動獲得プログラム(Clupes)にて獲得

<http://www.ci.seikei.ac.jp/sakai/clupes.html>

# 就職活動支援への応用

「医薬品」で検索した結果

↑ ツムラの純利益、4—9月20%増、子会社売却益13億円。

081107-0061 小野薬品工業  
↓ 小野薬4—9月、純利益19%減に。

081106-2451 バイタルネット  
↓ 決算から——4—9月、バイタルネット、連結子会社離脱で減収。

081103-0171 第一三共  
↓ 7—9月、印ランバクシー、最終赤字39億ルピー。

081101-0065 大日本住友製薬  
↓ 4—9月、大日本住友、純利益21%減。

081029-2270 朝日印刷  
↑ 朝日印刷4—9月、純利益9%増、中間配15円に。

081029-0048 三菱倉庫 → 「三菱倉庫」?

↑ 4—9月、三菱倉、経常益5%増。

081021-0050 武田薬品工業  
↓ 武田、純利益6.8%減、4—9月、予想より減益幅縮小。

081010-0060 日医工

# 就職活動支援への応用

The screenshot shows a web browser window with the URL <http://133.220.112.11/CEES/cees.cgi?081029-0048>. The page title is "CEES(Causal Expression ...)" and the tab title is "Causal Expressions Extraction System". The main content area displays a news article with the headline "4—9月、三菱倉庫、経常益5%増。" (Mitsubishi Kouryou, operating profit up 5% for the April-September period). The text discusses Mitsubishi Kouryou's financial performance, mentioning a 5% increase in operating profit, new warehouse openings in Saitama and Osaka Prefecture, and increased handling of pharmaceuticals and food products.

4—9月、三菱倉庫、経常益5%増。

三菱倉庫の二〇〇八年四—九月期の連結経常利益は、前年同期比5%増の七十五億円強になったようだ。従来予想は4%増の七十四億円。特殊な管理が必要な医薬品など利益率の高い物流受託が伸びた。売上高は1%増の八百五十億円強になったようだ。埼玉県や大阪府で新規に稼働する倉庫が貢献した。医薬品や食品などで倉庫保管・陸上運送を一括で取り扱う貨物が増えた。不動産事業では新規の賃貸物件はなかったものの、テナント入れ替わりなどに伴う工事作業で增收となった。営業利益は10%増の七十億円弱になったようだ。物流事業で港湾作業が伸び悩んだが、好採算の倉庫保管が増えた。

- ↑ ★★★ 特殊な管理が必要な医薬品など利益率の高い物流受託が伸びた。
- ↑ 埼玉県や大阪府で新規に稼働する倉庫が貢献した。
- ↑ ★★ 医薬品や食品などで倉庫保管・陸上運送を一括で取り扱う貨物が増えた。
- ↑ ★ 物流事業で港湾作業が伸び悩んだが、好採算の倉庫保管が増えた。

「三菱倉庫」と  
「医薬品」が  
関連がある

# CEESの問題点

- 日経新聞記事における業績発表記事から業績要因を抽出して検索対象としているので、著作権の問題で一般公開できない。
- ↓
- **決算短信**など、企業WEBサイト等で一般に公開している情報から業績要因を抽出して、それを検索対象にすれば、一般公開できる。

# 決算短信

- 決算短信とは、上場会社が決算発表及び四半期決算発表を行う際に、決算内容の要点をまとめた書類の名称
- 記者クラブが決算発表内容の標準化を目的として上場会社に要請したことから始まり、現在は取引所が様式を定め、全ての上場会社が作成
- 決算短信はPDF形式で企業のWebサイトなどで配布され、誰でも閲覧可能

平成26年3月期 決算短信【米国基準】(連結)																																																										
上場会社名 ソニー株式会社					平成26年5月14日 上場取引所 東																																																					
コード番号 6758 URL <a href="http://www.sony.co.jp/">http://www.sony.co.jp/</a>					TEL 03-6748-2111(代表) 平成26年6月3日																																																					
代表者 (役職名) 代表執行役員 問合せ先責任者 (役職名) 財務部 VP 定時株主総会開催予定日 平成26年6月19日 有価証券報告書提出予定日 平成26年6月26日					(氏名) 平井 一夫 (氏名) 村上 敦子 配当支払開始予定日																																																					
決算補足説明資料作成の有無 : 有 決算説明会開催の有無 : 有 (投資家・アリスト向け)					(百万円未満四捨五入)																																																					
1. 平成26年3月期の連結業績(平成25年4月1日～平成26年3月31日) (1) 連結経営成績 (%表示は対前期増減率)																																																										
<table border="1"><thead><tr><th></th><th>売上高及び営業収入</th><th>営業利益</th><th>税引前当期純利益</th><th>当社株主に帰属する当期純利益</th></tr><tr><th></th><th>百万円</th><th>百万円</th><th>百万円</th><th>百万円</th></tr></thead><tbody><tr><td>26年3月期</td><td>7,767,266</td><td>14.3</td><td>26,495</td><td>△88.3</td></tr><tr><td>25年3月期</td><td>6,795,504</td><td>4.7</td><td>226,503</td><td>—</td></tr><tr><td>(注)当期包括利益</td><td>26年3月期 121,978百万円 (62.6%)</td><td>25年3月期 325,798百万円 (—%)</td><td>26年3月期 25,741</td><td>△89.4</td></tr><tr><td></td><td></td><td></td><td>242,084</td><td>—</td></tr><tr><td></td><td></td><td></td><td>41,540</td><td>—</td></tr></tbody></table>											売上高及び営業収入	営業利益	税引前当期純利益	当社株主に帰属する当期純利益		百万円	百万円	百万円	百万円	26年3月期	7,767,266	14.3	26,495	△88.3	25年3月期	6,795,504	4.7	226,503	—	(注)当期包括利益	26年3月期 121,978百万円 (62.6%)	25年3月期 325,798百万円 (—%)	26年3月期 25,741	△89.4				242,084	—				41,540	—														
	売上高及び営業収入	営業利益	税引前当期純利益	当社株主に帰属する当期純利益																																																						
	百万円	百万円	百万円	百万円																																																						
26年3月期	7,767,266	14.3	26,495	△88.3																																																						
25年3月期	6,795,504	4.7	226,503	—																																																						
(注)当期包括利益	26年3月期 121,978百万円 (62.6%)	25年3月期 325,798百万円 (—%)	26年3月期 25,741	△89.4																																																						
			242,084	—																																																						
			41,540	—																																																						
<table border="1"><thead><tr><th></th><th>1株当たり当社株主に帰属する当期純利益</th><th>潜在株式調整後1株当たり当社株主に帰属する当期純利益</th><th>株主資本当社株主に帰属する当期純利益</th><th>総資産税引前当期純利益率</th><th>売上高営業利益率</th></tr><tr><th></th><th>円 銭</th><th>円 銭</th><th>円 銭</th><th>%</th><th>%</th></tr></thead><tbody><tr><td>26年3月期</td><td>△124.99</td><td>△124.99</td><td>△5.8</td><td>0.2</td><td>0.3</td></tr><tr><td>25年3月期</td><td>41.32</td><td>38.79</td><td>2.0</td><td>1.8</td><td>3.3</td></tr></tbody></table>											1株当たり当社株主に帰属する当期純利益	潜在株式調整後1株当たり当社株主に帰属する当期純利益	株主資本当社株主に帰属する当期純利益	総資産税引前当期純利益率	売上高営業利益率		円 銭	円 銭	円 銭	%	%	26年3月期	△124.99	△124.99	△5.8	0.2	0.3	25年3月期	41.32	38.79	2.0	1.8	3.3																									
	1株当たり当社株主に帰属する当期純利益	潜在株式調整後1株当たり当社株主に帰属する当期純利益	株主資本当社株主に帰属する当期純利益	総資産税引前当期純利益率	売上高営業利益率																																																					
	円 銭	円 銭	円 銭	%	%																																																					
26年3月期	△124.99	△124.99	△5.8	0.2	0.3																																																					
25年3月期	41.32	38.79	2.0	1.8	3.3																																																					
<table border="1"><thead><tr><th>(参考)持分法投資損益</th><th>26年3月期 △7,374百万円</th><th>25年3月期 △6,946百万円</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></tr><tr><th></th><th>円 銭</th><th>円 銭</th><th>百万円</th><th>百万円</th><th>百万円</th><th>百万円</th><th>百万円</th><th>百万円</th><th>百万円</th></tr></thead><tbody><tr><td></td><td></td><td></td><td>26年3月期</td><td>25年3月期</td><td>26年3月期</td><td>25年3月期</td><td>26年3月期</td><td>25年3月期</td><td>26年3月期</td></tr></tbody></table>										(参考)持分法投資損益	26年3月期 △7,374百万円	25年3月期 △6,946百万円									円 銭	円 銭	百万円	百万円	百万円	百万円	百万円	百万円	百万円				26年3月期	25年3月期	26年3月期	25年3月期	26年3月期	25年3月期	26年3月期																			
(参考)持分法投資損益	26年3月期 △7,374百万円	25年3月期 △6,946百万円																																																								
	円 銭	円 銭	百万円	百万円	百万円	百万円	百万円	百万円	百万円																																																	
			26年3月期	25年3月期	26年3月期	25年3月期	26年3月期	25年3月期	26年3月期																																																	
<table border="1"><thead><tr><th>(2) 連結財政状態</th><th>総資産</th><th>資本合計(純資産)</th><th>株主資本</th><th>株主資本比率</th><th>1株当たり株主資本</th></tr><tr><th></th><th>百万円</th><th>百万円</th><th>百万円</th><th>%</th><th>円 銭</th></tr></thead><tbody><tr><td>26年3月期</td><td>15,333,720</td><td>2,783,141</td><td>2,258,137</td><td>14.7</td><td>2,163.63</td></tr><tr><td>25年3月期</td><td>14,211,033</td><td>2,672,004</td><td>2,192,262</td><td>14.7</td><td>2,168.62</td></tr></tbody></table>										(2) 連結財政状態	総資産	資本合計(純資産)	株主資本	株主資本比率	1株当たり株主資本		百万円	百万円	百万円	%	円 銭	26年3月期	15,333,720	2,783,141	2,258,137	14.7	2,163.63	25年3月期	14,211,033	2,672,004	2,192,262	14.7	2,168.62																									
(2) 連結財政状態	総資産	資本合計(純資産)	株主資本	株主資本比率	1株当たり株主資本																																																					
	百万円	百万円	百万円	%	円 銭																																																					
26年3月期	15,333,720	2,783,141	2,258,137	14.7	2,163.63																																																					
25年3月期	14,211,033	2,672,004	2,192,262	14.7	2,168.62																																																					
<table border="1"><thead><tr><th>(3) 連結キャッシュ・フローの状況</th><th>営業活動によるキャッシュ・フロー</th><th>投資活動によるキャッシュ・フロー</th><th>財務活動によるキャッシュ・フロー</th><th>現金及び現金同等物期末残高</th></tr><tr><th></th><th>百万円</th><th>百万円</th><th>百万円</th><th>百万円</th></tr></thead><tbody><tr><td>26年3月期</td><td>664,116</td><td>△710,502</td><td>207,877</td><td>1,046,466</td></tr><tr><td>25年3月期</td><td>476,165</td><td>△705,280</td><td>88,528</td><td>826,361</td></tr></tbody></table>										(3) 連結キャッシュ・フローの状況	営業活動によるキャッシュ・フロー	投資活動によるキャッシュ・フロー	財務活動によるキャッシュ・フロー	現金及び現金同等物期末残高		百万円	百万円	百万円	百万円	26年3月期	664,116	△710,502	207,877	1,046,466	25年3月期	476,165	△705,280	88,528	826,361																													
(3) 連結キャッシュ・フローの状況	営業活動によるキャッシュ・フロー	投資活動によるキャッシュ・フロー	財務活動によるキャッシュ・フロー	現金及び現金同等物期末残高																																																						
	百万円	百万円	百万円	百万円																																																						
26年3月期	664,116	△710,502	207,877	1,046,466																																																						
25年3月期	476,165	△705,280	88,528	826,361																																																						
<table border="1"><thead><tr><th>2. 配当の状況</th><th colspan="5">年間配当金</th><th>配当金額(合計)</th><th>配当性向(連結)</th><th>株主資本配当(連結)</th></tr><tr><th></th><th>第1四半期末</th><th>第2四半期末</th><th>第3四半期末</th><th>期末</th><th>合計</th><th>円 銭</th><th>百万円</th><th>%</th><th>%</th></tr><tr><th></th><th>円 銭</th><th>円 銭</th><th>円 銭</th><th>円 銭</th><th>円 銭</th><th>円 銭</th><th>百万円</th><th>%</th><th>%</th></tr></thead><tbody><tr><td>25年3月期</td><td>—</td><td>12.50</td><td>—</td><td>12.50</td><td>25.00</td><td>25,181</td><td>58.4</td><td>1.2</td><td></td></tr><tr><td>26年3月期</td><td>—</td><td>12.50</td><td>—</td><td>12.50</td><td>25.00</td><td>26,016</td><td>—</td><td>1.2</td><td></td></tr></tbody></table>										2. 配当の状況	年間配当金					配当金額(合計)	配当性向(連結)	株主資本配当(連結)		第1四半期末	第2四半期末	第3四半期末	期末	合計	円 銭	百万円	%	%		円 銭	円 銭	円 銭	円 銭	円 銭	円 銭	百万円	%	%	25年3月期	—	12.50	—	12.50	25.00	25,181	58.4	1.2		26年3月期	—	12.50	—	12.50	25.00	26,016	—	1.2	
2. 配当の状況	年間配当金					配当金額(合計)	配当性向(連結)	株主資本配当(連結)																																																		
	第1四半期末	第2四半期末	第3四半期末	期末	合計	円 銭	百万円	%	%																																																	
	円 銭	円 銭	円 銭	円 銭	円 銭	円 銭	百万円	%	%																																																	
25年3月期	—	12.50	—	12.50	25.00	25,181	58.4	1.2																																																		
26年3月期	—	12.50	—	12.50	25.00	26,016	—	1.2																																																		
平成27年3月期の配当予想額については未定です。																																																										
3. 平成27年3月期の連結業績予想(平成26年4月1日～平成27年3月31日) (%表示は、対前期増減率)																																																										
<table border="1"><thead><tr><th></th><th>売上高及び営業収入</th><th>営業利益</th><th>税引前当期純利益</th><th>当社株主に帰属する当期純利益</th></tr><tr><th></th><th>百万円</th><th>百万円</th><th>百万円</th><th>百万円</th></tr></thead><tbody><tr><td>通期</td><td>7,800,000</td><td>0.4</td><td>428.4</td><td>405.0</td></tr><tr><td></td><td></td><td></td><td>130,000</td><td>△50,000</td></tr></tbody></table>											売上高及び営業収入	営業利益	税引前当期純利益	当社株主に帰属する当期純利益		百万円	百万円	百万円	百万円	通期	7,800,000	0.4	428.4	405.0				130,000	△50,000																													
	売上高及び営業収入	営業利益	税引前当期純利益	当社株主に帰属する当期純利益																																																						
	百万円	百万円	百万円	百万円																																																						
通期	7,800,000	0.4	428.4	405.0																																																						
			130,000	△50,000																																																						

ソニーの平成26年3月期決算短信

# 決算短信

- PDF形式ではあるが、業績情報だけでなく、事業環境や業績要因(なぜ業績が好調、あるいは不振であったかの理由)など、多くのテキスト情報が存在



個人投資家への投資判断の支援  
を行うための情報源として有望の  
はず



テキストマイニング技術を活用して分析

ソニー株式会社(6758) 2013年度 決算短信

営業損失は、前年度に比べ588億円縮小し、255億円となりました。この損益改善は、主に、液晶テレビの製品ミックスの改善及び費用の削減によるものです。また、当年度の構造改革費用（純額）は、前年度に比べ108億円減少し、16億円となりました。

なお、テレビについては、売上高は、前年度比 29.7%増加の 7,543 億円となりました。営業損失\*は前年度に比べ、439 億円縮小し、257 億円となりました。

\* 分野全体に含まれる構造改革費用は製品カテゴリーには配賦されておらず、テレビの営業損失には含まれていません。

## デバイス分野

	2012年度 億円	2013年度 億円	増減率 %
売上高	8,486	7,942	△6.4
営業利益（損失）	439	△130	-

デバイス分野には、半導体カテゴリー及びコンポーネントカテゴリーが含まれます。半導体カテゴリーにはイメージセンサー、コンポーネントカテゴリーには電池、記録メディア、データ記録システムなどが主要製品として含まれています。

デバイス分野の売上高は、前年度比 6.4%減少し、7,942 億円となりました（前年度の為替レートを適用した場合、19%の減収）。当年度において、為替の好影響及びモバイル機器向けの需要増加によるイメージセンターの大幅な増収がありましたが、主に PS3®向けシステム LSI の減収や前年度にはケミカルプロダクツ関連事業の売上が含まれていたことなどにより、分野全体で減収となりました。なお、外部顧客に対する売上高は、前年度比 0.9%増加しました。

営業損益は、前年度の 439 億円の利益に対し、当年度は 130 億円の損失となりました。この大幅な損益悪化は、主に電池事業において 321 億円の長期性資産の減損を計上したこと、及び、2011 年度に発生したタイの洪水による損害や損失に対する保険収益（純額）が前年度に比べ減少したことによるものです。なお、当年度の構造改革費用（純額）は、前年度に比べ 102 億円減少し、89 億円となりました。

\* \* \* \* \*

ソニーの平成26年3月期決算短信

# 決算短信PDFファイルの自動取得

企業Web ページ  
(3,821社)

※ 決算短信検索システムを一般に公開するために、企業  
Webページから取得(決算短信PDFへのリンクも得る)



IR 情報ページかそれ以外のページであるかを  
SVM(機械学習手法のひとつ)によって自動分類



精度はいいが、  
網羅率がよくない

IR情報ページのURL  
によく出現する文字列

IR情報ページによく出現するURL の文字列を抽出し、その文字列  
をURLに含むページにあるPDFファイルを全てダウンロード

ir, library, company, investor, calendar.html,  
financial, library.html, calendar, investors,  
finance, IR, report,tanshin.html, irinfo ...



ダウンロードしたPDFファイルから、  
決算短信PDFを選別



# 企業WEBページの自動収集

38

- 上場企業の数は約4,000社あり、1社あたりのWEBサイトは、1,000ファイル～10,000ファイルのHTMLファイルで構成されている → 人手で集めることは不可能
- クローラーと呼ばれるWEBページ自動収集プログラムを作成して、自動的にWEBページを収集

# 企業WEBページの自動収集

39

## クローラーの動作

Step 1: 企業Webサイトのトップページのファイル(index.html)をダウンロードし、その中のHTMLファイルへのリンクを得る。

Step 2: リンク先のHTMLファイルをダウンロードし、その中のHTMLファイルへのリンクを得る

Step 3: Step 2の処理を再帰的に繰り返し、企業WEBサイトにおける全てのHTMLファイルをダウンロードしたら終了

# 企業WEBページの自動収集

40

## クローラー作成時の注意点

- 企業側のWebサイトに負担を掛けてはいけないので、3秒に1ファイルのダウンロードしかないように、アクセスに制限をかける
- 上記のようなことをしていては膨大な時間がかかるので、複数企業(50くらい)に対して同時にアクセスをかける

# 決算短信からの業績要因文の抽出

41

- 業績要因文を抽出するのに有効な手がかり表現（「好調でした」等）と企業ごとの重要なキーワード（企業キーワード）を使用して、業績要因文を抽出

例：

- 電子表示デバイス用ガラス基板の好調が寄与しました。
- 「遊戯王トレーディングカード」シリーズが堅調に推移いたしました。

「赤字：企業キーワード」、「青字：手がかり表現」

手がかり表現

堅調でした、好調でした、伸び悩んだ、寄与しました、低調でありました、低迷した

企業ごとの企業キーワード、有効な手がかり表現は数多く、  
全て人手で用意することは困難

手がかり表現、企業キーワードを自動獲得

# 手がかり表現の自動獲得

42

## ■ 初期手がかり表現

が好調, が不振



## ■ 当研究室で公開している手がかり表現自動獲得プログラム(Clupes)にて獲得

<http://www.ci.seikei.ac.jp/sakai/clupes.html>

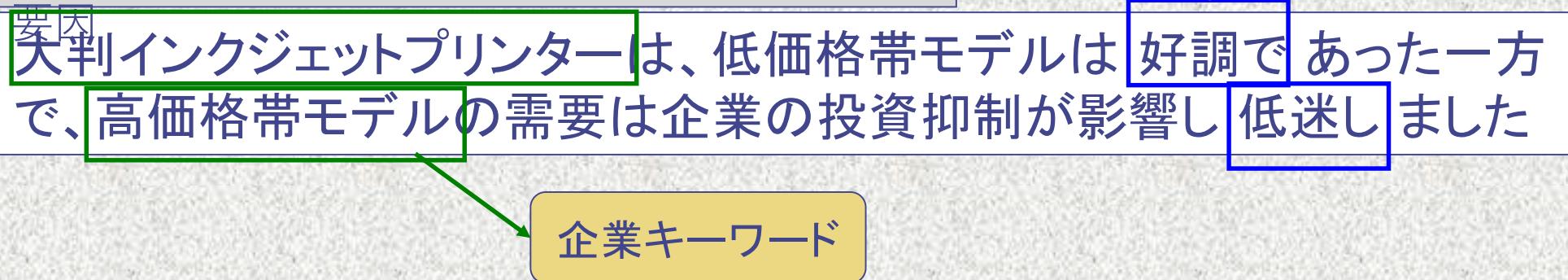
# 決算短信からの業績要因抽出

決算短信PDFから企業ごとに重要なキーワード(企業キーワード)を抽出し、企業キーワードと手がかり表現を含んでいる文を抽出

例：ソニーの決算短信PDFに含まれる業績



例：エプソンの決算短信PDFに含まれる業績



# 企業キーワードの獲得

- 企業  $t$  の決算短信PDF集合  $S(t)$  に含まれる名詞  $n_i$  に対して、重み  $W(n_i, S(t))$  を計算

$$W(n_i, S(t)) = Tf(n_i, S(t)) \times H(n_i, S(t)) \times idf(n_i)$$

$S(t)$ : 企業  $t$  のWebサイトからダウンロードした決算短信PDFの集合

$Tf(n_i, S(t))$ :  $S(t)$ において名詞  $n_i$  が出現する頻度.

$H(n_i, S(t))$ :  $S(t)$ の要素である決算短信PDF  $d$  に名詞  $n_i$  が出現する確率に基づくエントロピー(後述)

重みが大きい名詞を企業キーワードとして抽出

# エントロピー $H(n_i, S(t))$

$S(t)$ の各決算短信PDF  $d$  に名詞 $n_i$  が出現する確率  $P(n_i, d)$  に基づくエントロピー

$$H(n_i, S(t)) = - \sum_{d \in S(t)} P(n_i, d) \log_2 P(n_i, d)$$

名詞 $n_i$  が企業 $t$  のWebサイトからダウンロードした決算短信PDF の集合 $S(t)$  に、まんべんなく出現している場合に大きな値

# 決算短信の名詞の重み

- 企業 $t$  の決算短信PDFにおける名詞 $n_i$  に対して,  $idf(n_i)$  を計算

$$idf(n_i) = \log_2 \frac{|N|}{df(n_i)}$$

$df(n_i)$ : 名詞 $n_i$  を含む決算短信PDFを持つ企業の数

$N$ : 決算短信PDFを収集した企業の集合

$W(n_i, S(t))$  が大きい名詞

- 企業 $t$ のWebサイトからダウンロードした決算短信PDFの集合中に多く, かつ, まんべんなく出現
- 企業 $t$ の決算短信PDFに出現しており, 他の企業の決算短信PDFには出現していない名詞

# 抽出された企業キーワードの例

企業名称	企業キーワード
東芝	電子デバイス部門, インフラ部門, デバイス部門, ストレージ
大日本印刷	エレクトロニクス, 印刷事業, 液晶カラーフィルター
力ゴメ	野菜飲料, 野菜生活, 果美食品, 生鮮トマト, 飲料事業
エーザイ	医薬品, アリセプト, パリエット, 医薬品事業, 抗がん剤

# Causal Expressions Extraction System (決算短信検索システム)

48

CEES (Causal Expression Extraction System) - Google Chrome  
133.220.112.11/cees/cees\_search.cgi  
アメブロ News Financial Shopping Seikei SNS Google Yahoo  
Language Information Laboratory's company search system  
CEES  
Japanese Site  
エアコン 検索 クリア  
1. 富士通ゼネラル  
平成26年3月期決算短信【日本基準】(連結) 平成26年04月25日  
・米州では、北米において、天候にも恵まれエアコン需要が伸長するなか、政府や電力会社の補助金対象となる省エネ性能に優れたルームエアコンの拡販に努めるとともに、寒冷地向け機種のラインアップ強化による暖房需要の取り込みが進展し、売上が増加しました。  
・ブラジルでは、大型機種やマルチエアコンの販路拡大に取り組み、売上が増加しました。  
・オセアニアでは、天候不順の影響を受け市況が停滞するなか、下半期の需要期に向け省エネ性能を大幅に高めたルームエアコンを投入するなど拡販に努め、前年度並みの売上を確保しました。  
平成26年3月期第3四半期決算短信【日本基準】(連結) 平成26年01月24日  
・中国では、猛暑により需要が前年を上回るなか、ルームエアコンの販売間口および地域の拡大を進めるとともに、VRFでも営業体制強化による販売網拡大に取り組み、売上が増加しました。  
・なお、需要が増加している家庭用マルチエアコンの販売拡大に向け、室外機の小型化による設置性向上と同時に高い省エネ性能を実現

抽出した業績要因を対象にし、キーワードで  
決算短信を検索できるシステム

「エアコン」を業績要因とする  
企業の決算短信を検索



「エアコン」と関連が深い企業を  
検索可能

研究室WEBページにて公開  
<http://hawk.ci.seikei.ac.jp/cees/>

# 決算短信検索システムを使用した 企業検索

「太陽電池」を業績要因とする企業  
=「太陽電池」と深く関連がある企業

「太陽電池」で検索される企業

東洋炭素、クレハ、カネカ、旭ダイヤモンド工業、日清紡  
ホールディングス、SUMCO、ユシロ化学工業、リンテック、  
島津製作所...

現在の投資テーマに関連している企業を検索可能

# 決算短信検索システムを使用した企業分析

例:「東洋炭素」の決算短信PDFを検索



## 1. 東洋炭素

複合材, 等方性黒鉛材料, 黒鉛, コンポジット製品, 半導体, 単結晶シリコン, 一般産業分野, 太陽電池, 単結晶シリコン製造用, 太陽電池用, 小型モーター用, カーボン,

■ 平成26年12月期第1四半期決算短信〔日本基準〕(連結) 平成26年05月14日

### 【業績要因】

- ・また、LED市場が堅調であることに加え、一般産業用市場においては景気回復を背景に需要が増加する等、全体としては緩やかながらも回復の傾向をたどりました。
- ・米国一般産業用等の一部用途は底堅く推移したものの、半導体用等のエレクトロニクス関連の不振により、総じて低調に推移いたしました。
- ・欧州工業炉用や放電加工電極用等の一般産業用の拡販が進んだこと等により、収益の改善が進みました。
- ・アジア中国を中心に太陽電池用型モーターブラシも堅調を維持する。
- ・また、LED市場が堅調であることは景気回復を背景に需要が増加するが続くものと想定しております。

業績要因を表示



企業Webサイトで公開している決算短信PDFを閲覧可能

業績要因に多く出現しているキーワード  
= その企業が力をいれている内容

# 業績要因への極性付与

- より有効な情報として利用するために、抽出した業績要因に対して業績に対する極性（「ポジティブ」、「ネガティブ」）を付与する必要

## 例

半導体製造装置の受注が好調でした。



ポジティブ

世界的な太陽電池市況の低迷により太陽電池製造装置の販売が減少しました。



ネガティブ

業績要因を使用した景気動向予測、および、業績要因に基づいて株取り引きを行うコンピュータトレーディングにも応用が期待

# 業績要因への極性付与

- 獲得した121種類の手がかり表現へ極性を付与し、付与した極性を使用して業績要因へ極性を付与

極性付与した 手がかり表現	手がかり表現	極性
	堅調でした	positive
	好調でした	positive
	伸び悩んだ	negative
	寄与しました	positive
	低調でありました	negative
	低迷した	negative

半導体製造装置の受注が好調でした。



ポジティブ

# 極性付与された手がかり表現を使用した業績要因への極性付与

## 例

海外売上高は、中国向け昇降機事業が堅調に推移した社会・産業システム部門等が 増加したものの、ハードディスクドライブ事業を売却したことや、電子装置・システム部門等が前年同期を 下回った ことから、前年同期に比べ11%減少し、8,677億円となりました

青字：ポジティブ手がかり表現

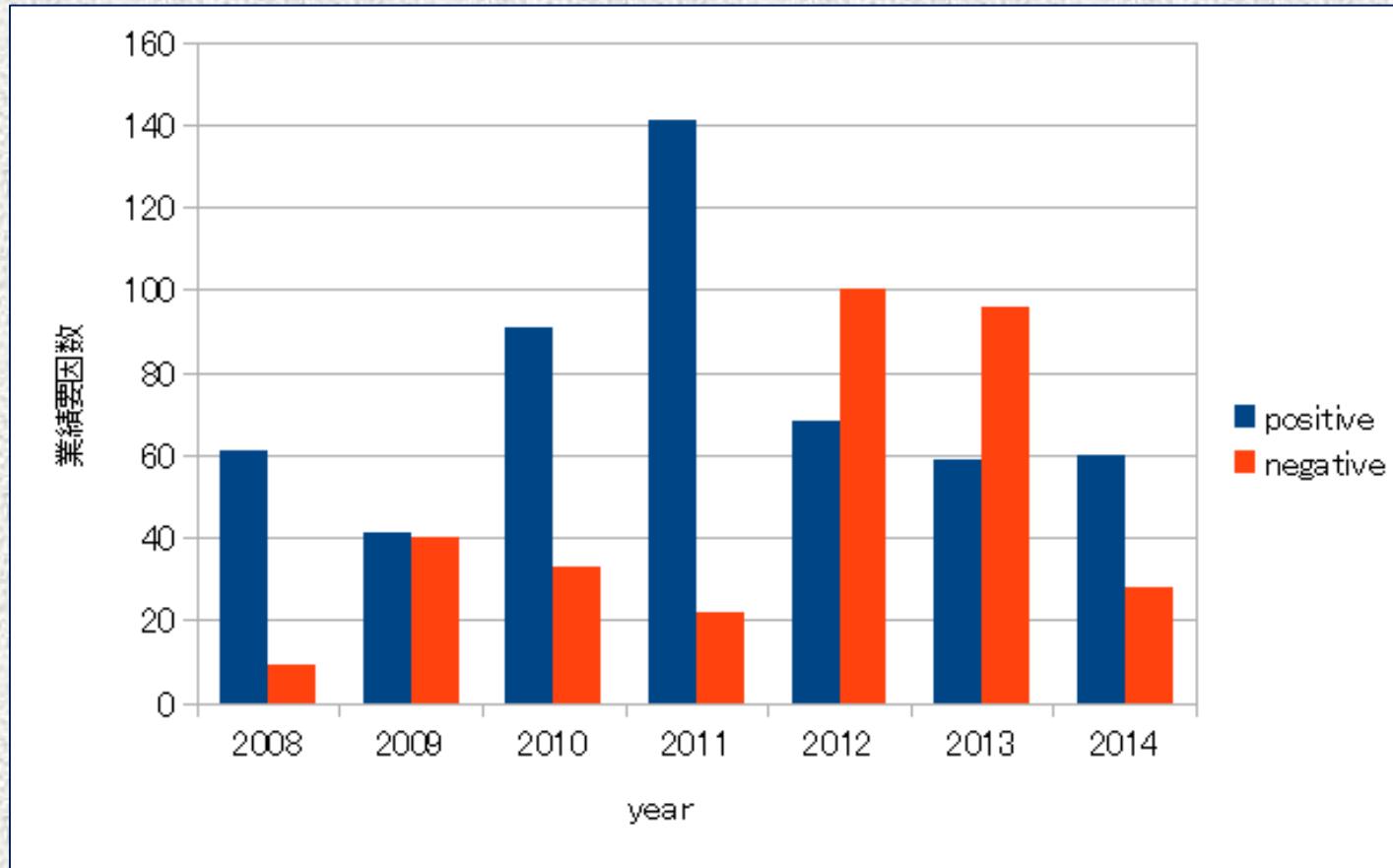
赤字：ネガティブ手がかり表現

- 業績要因の極性は最後の手がかり表現（例では「減少」）の極性（ネガティブ）に従う。

→ 「ネガティブ」が付与

# 極性付とした業績要因の応用

## ■ 「太陽電池」を含む業績要因の数の推移



# 極性付与された業績要因の例

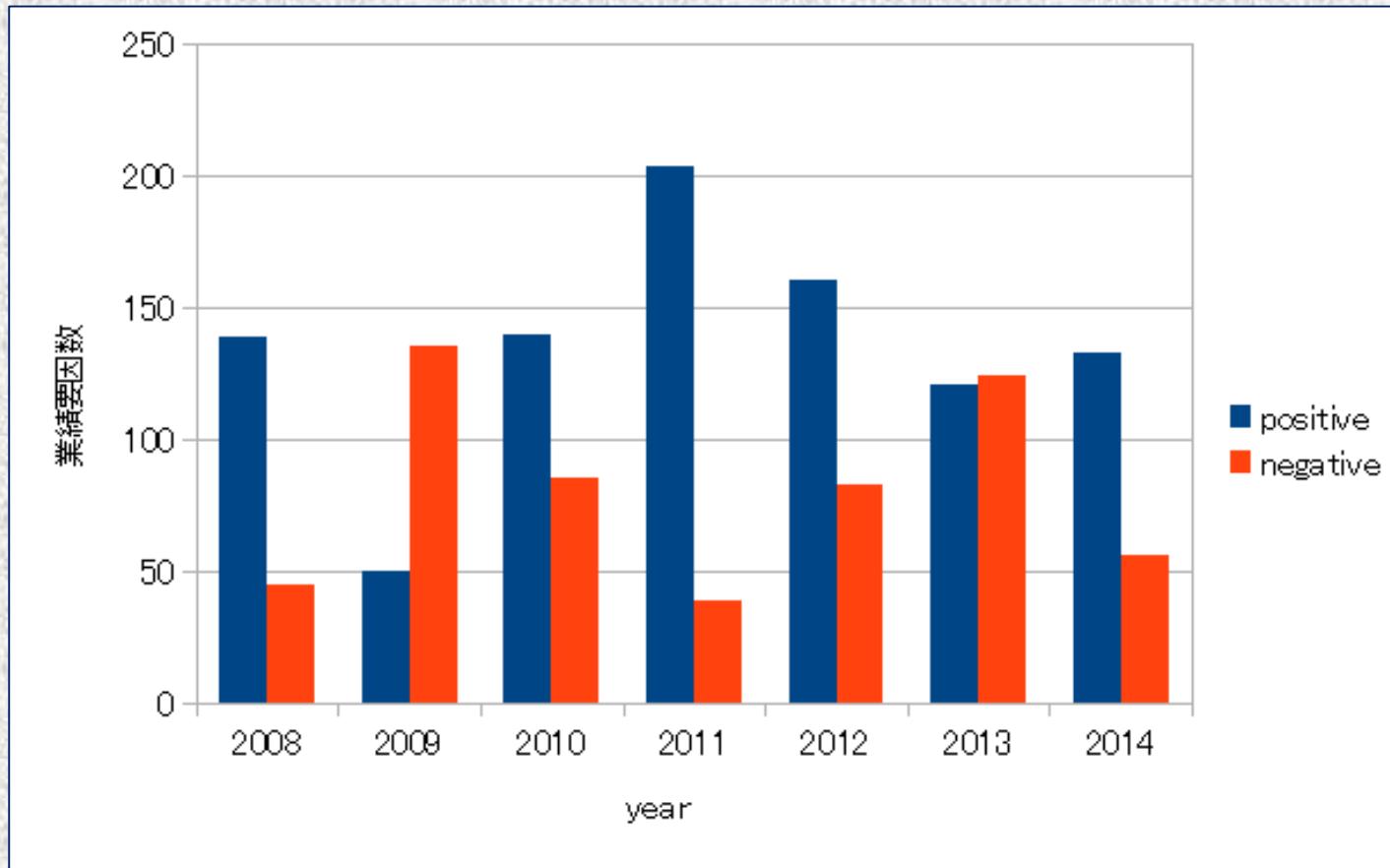
- 2012年の「太陽電池」を含むネガティブな業績要因

## 例

多結晶シリコンは、**太陽電池**パネルの供給過剰とパソコンの販売不振等を背景にした半導体ウエハーの在庫調整に伴う、販売数量の減少及び販売価格の下落により大幅な減収減益となりました

# 極性付与した業績要因の応用

## ■ 「建設機械」を含む業績要因の数の推移



# 極性付与された業績要因の例

- 2009年の「建設機械」を含むネガティブな業績要因

## 例

中国向け建設機械やロシア向け道路建設用資材プラントは  
世界的な金融危機の影響を受け、取扱いは減少しました。

# プログラミング言語 Perl

- Perl(パール)とは、ラリー・ウォールによって開発されたプログラミング言語である。
- 実用性と多様性を重視しており、C や シェルスクリプトなど他のプログラミング言語の優れた機能を取り入れている。
- Webアプリケーション、システム管理、テキスト処理などのプログラムを書くのに広く用いられている。

※Wikipediaより抜粋

これまで紹介したシステムも全てPerlで書かれている。

本講義では、講義内容をPerlで実装したコードを公開するとともに、簡単な機械学習(ナイーブ・ベイズを予定)プログラムをPerlで実装する

# データマイニング実用化の現状(1)

- 金融分野 → 膨大な顧客リストを分析
  - 生命保険の潜在的解約候補顧客の推定
  - 効果的なダイレクトメール宛先候補顧客の推定
  - 過去の消費者ローン審査データからのルール発見
  - 膨大なクレジットカード使用記録からの不正利用パターンの発見
  - 決算短信などの企業決算情報を自動分析



# データマイニング実用化の現状(2)

## ▶ 情報通信分野 ※

- 膨大な通信ログから不正使用や不正アクセスの特徴パターンを検出
- スパムメール(迷惑メール)の自動判定
- ブログやTwitter等からユーザの評判や興味の調査

## ➤ 製薬・医療分野

- 化学化合物分子構造と人体への生理的影響



※ IBM, NTT, 富士通, NEC, ディー・エヌ・エー, ドワンゴ...等

# 参考文献

- 元田浩, 津本周作, 山口高平, 沼尾正行,  
“データマイニングの基礎”, オーム社, 2006.
- 金明哲, 村上征勝, 永田昌明, 大津起夫, 山西健司,  
“言語と心理の統計”, 岩波書店, 2003.
- 高村大也, 奥村学,  
“言語処理のための機械学習入門”, コロナ社, 2010.
- 北研二, 津田和彦, 獅々堀正幹,  
“情報検索アルゴリズム”, 共立出版, 2002.
- 酒井浩之, 梅村祥之, 増山繁,  
“交通事故例に含まれる事故原因表現の新聞記事からの抽出”,  
自然言語処理, vol.13, no.4, pp.99–124, 2006.