

テキストマイニングとNLPビジネス

山西健司

NECインターネットシステム研究所

yamanisi@ccm.cl.nec.co.jp

<http://www.labs.nec.co.jp/DTmining/>

2003年10月15日

自然言語処理技術に関するシンポジウム2003

目次

1. はじめに
2. テキスト分類技術とCRM
3. マーケティング知識の発見
4. 評判分析とWebマイニング
5. トピック分析と情報監視
6. テキストマイニング: Challenges
7. おわりに
8. 参考文献

1. はじめに

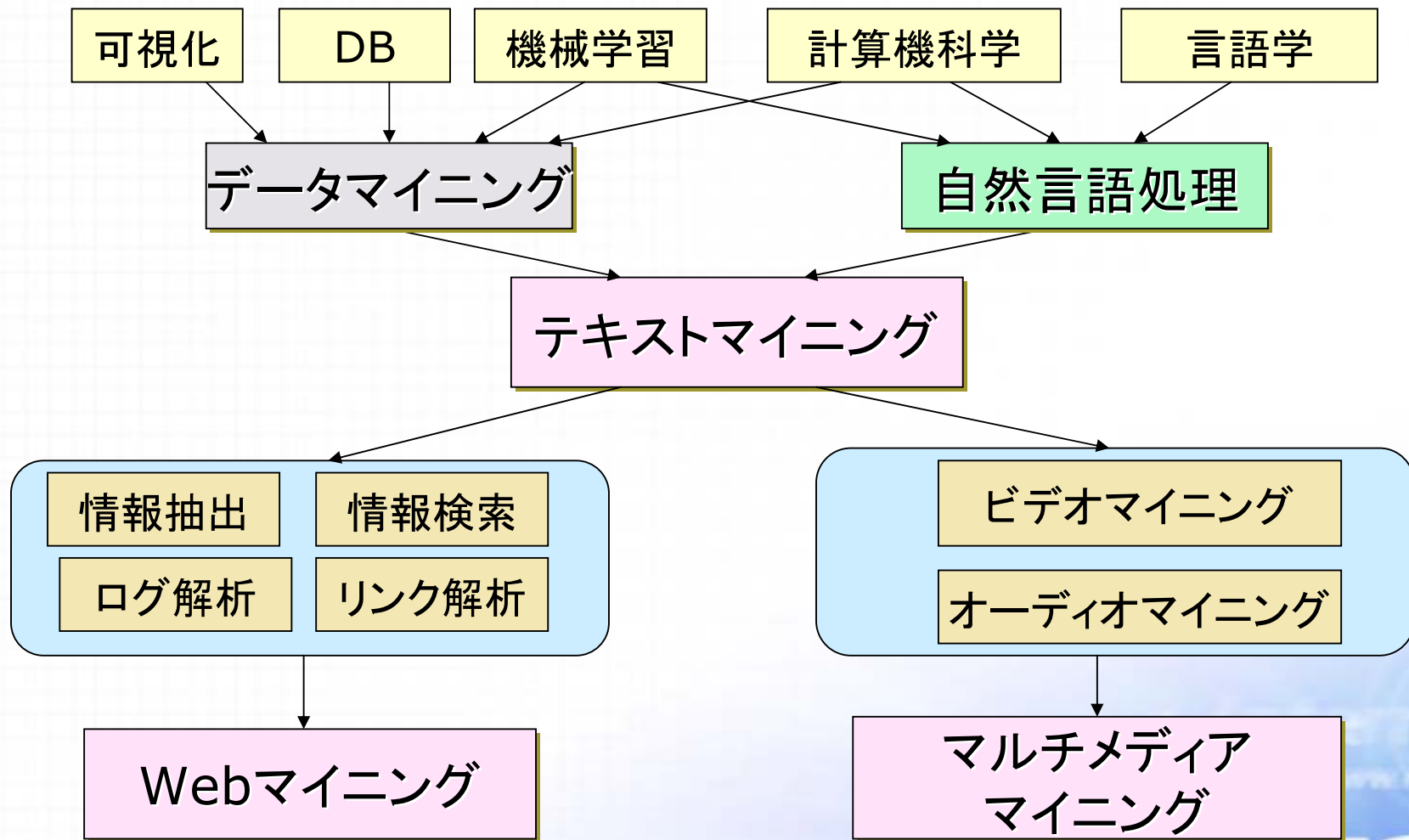
テキストマイニング

大量のテキストデータ(非構造・半構造データ)から新規性のある知識または構造を発見すること

⇒情報的なSurpriseがあること

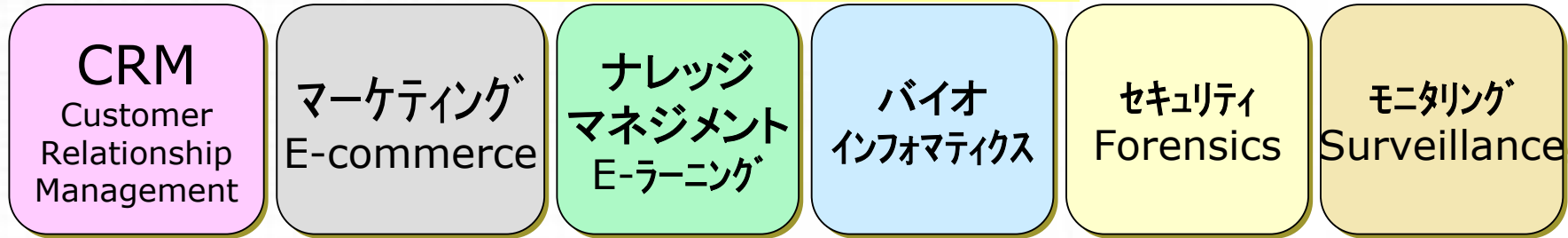
- 指定された条件の情報抽出、情報検索とは区別する
- 言語構造自体の解析(構文解析など)や文書構造自体の解析(情報要約など)とは区別する

テキストマイニングの位置づけ



テキストマイニングの要素技術と応用分野

知識発見+工数削減



メール分類
FAQ自動作成

アンケート分析
傾向分析

ナレッジ構造化
カリキュラム分析

バイオDBからの
知識発見

有害情報フィルタ
Spamフィルタ

情報監視

テキストマイニング

テキスト
分類

テキスト
クラスタリング

相関
分析

共起
分析

対応
分析

代表文
分析

Novelty
Detection

教師あり学習

教師なし学習

単語想起

単語共起

ポジショニング

スコアリング

異常検出

テキストマイニングの環境動向

市場動向

CRM: 2007年にて5000億市場、年率6.2%成長 (IDCジャパン)

ナレッジマネジメント、Forensics分野で新たなニーズが浮上

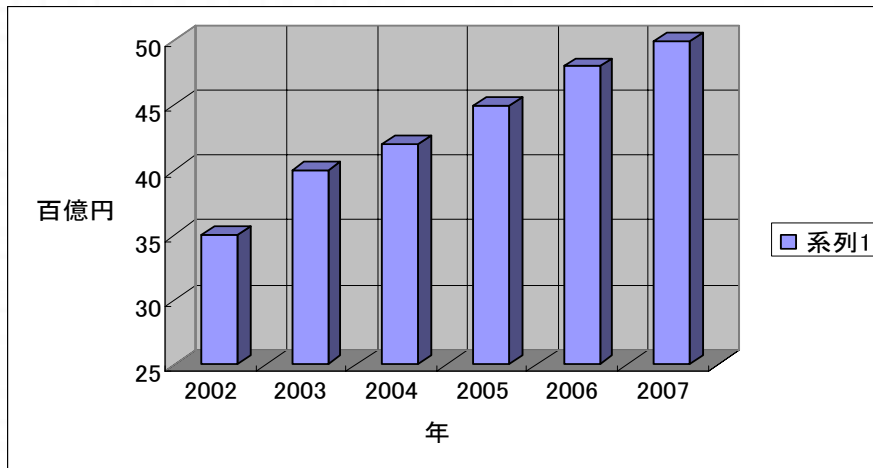
技術動向

IP化、ユビキタス環境がベース⇒リアルタイム、コンテキスト解析

CRM/SCM/KMの統合化

国内CRM市場

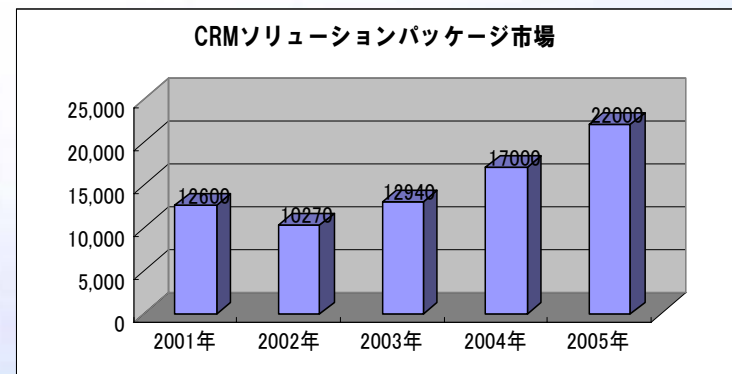
(IDCジャパン予測)



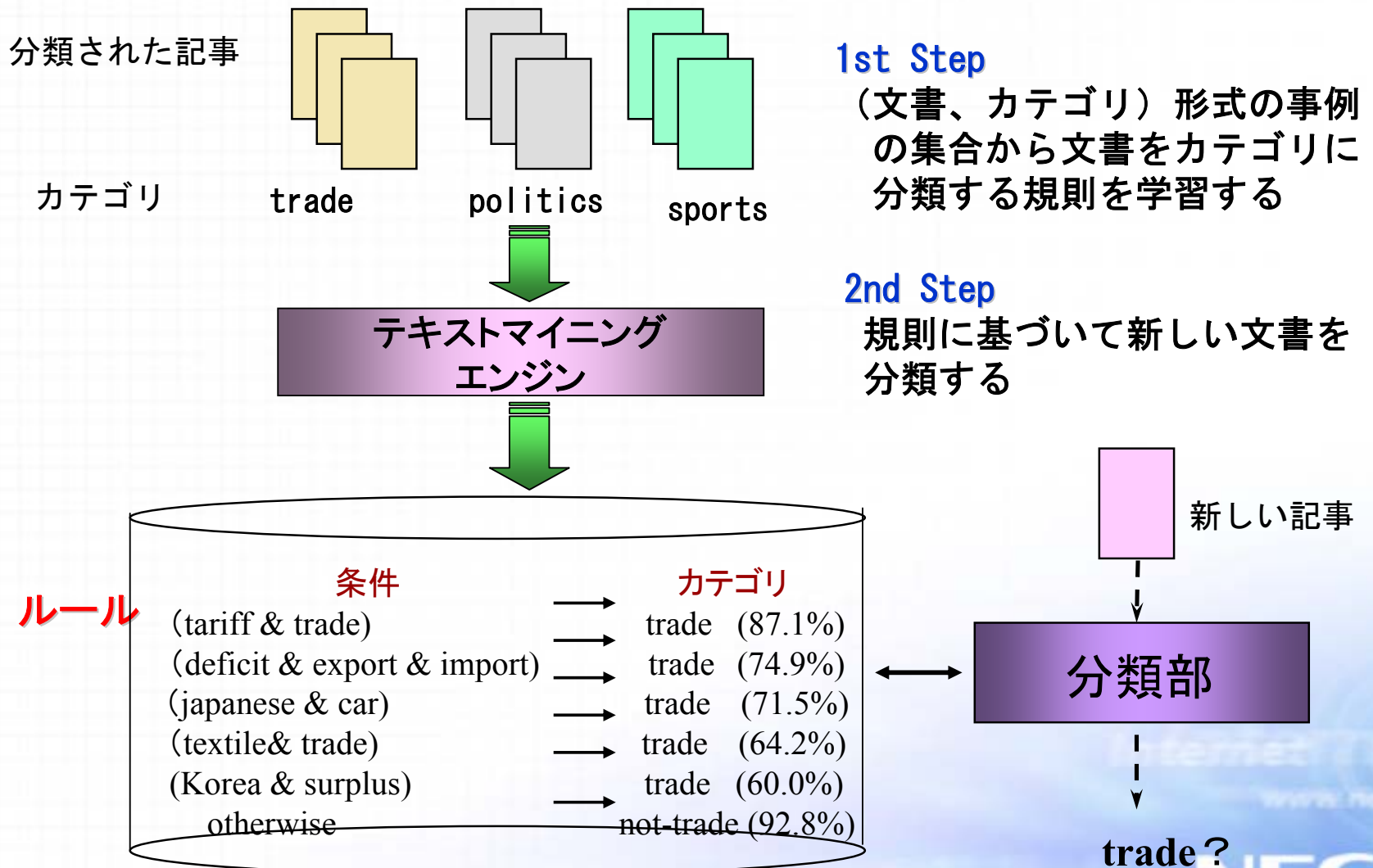
国内CRMパッケージ市場

(矢野経済研究所 2003.4.23)

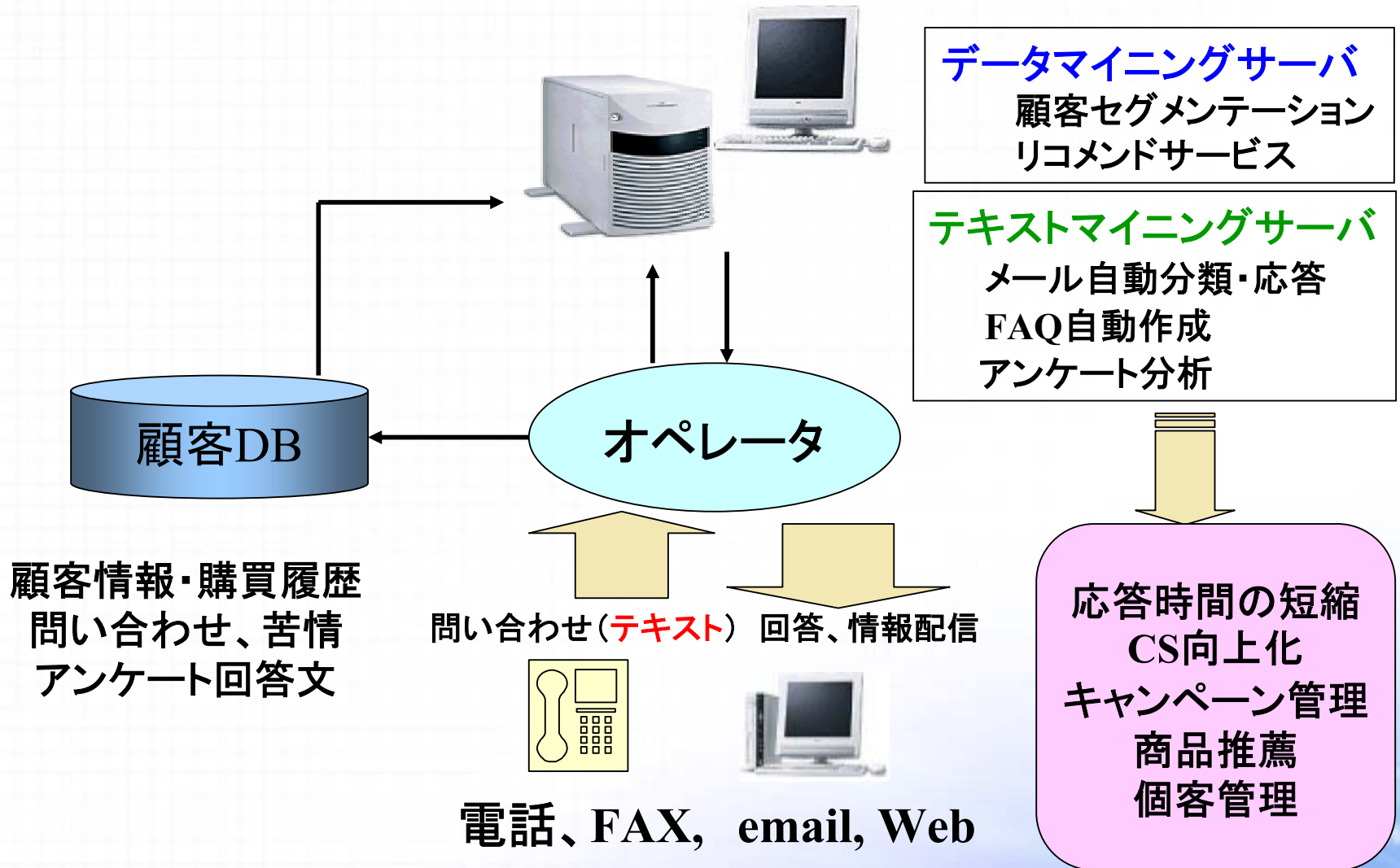
- 2004年以降、毎年130%近い伸張
- 2005年には220億円規模



2. テキスト分類とCRM



テキスト分類のコールセンタ応用



テキスト分類の研究動向

● ルールベースの方法

...高いReadability, modifiability, 知識の融合
やや低いEffectiveness

C4.5, Ripper[Cohen and Singer98]

Bayesian Net[Dumais et.al.98], decision rules[Apte et.al.94]

● 非ルールベースの方法

...低いReadability, 高いEffectiveness

Naïve Bayes[Kar and White 78], cosine法[Rocchio71]

SVM[Joachim98]

課題: ルールベースのreadabilityを保持しながら
高い分類精度を実現する手法の確立

ルールベースのテキスト分類

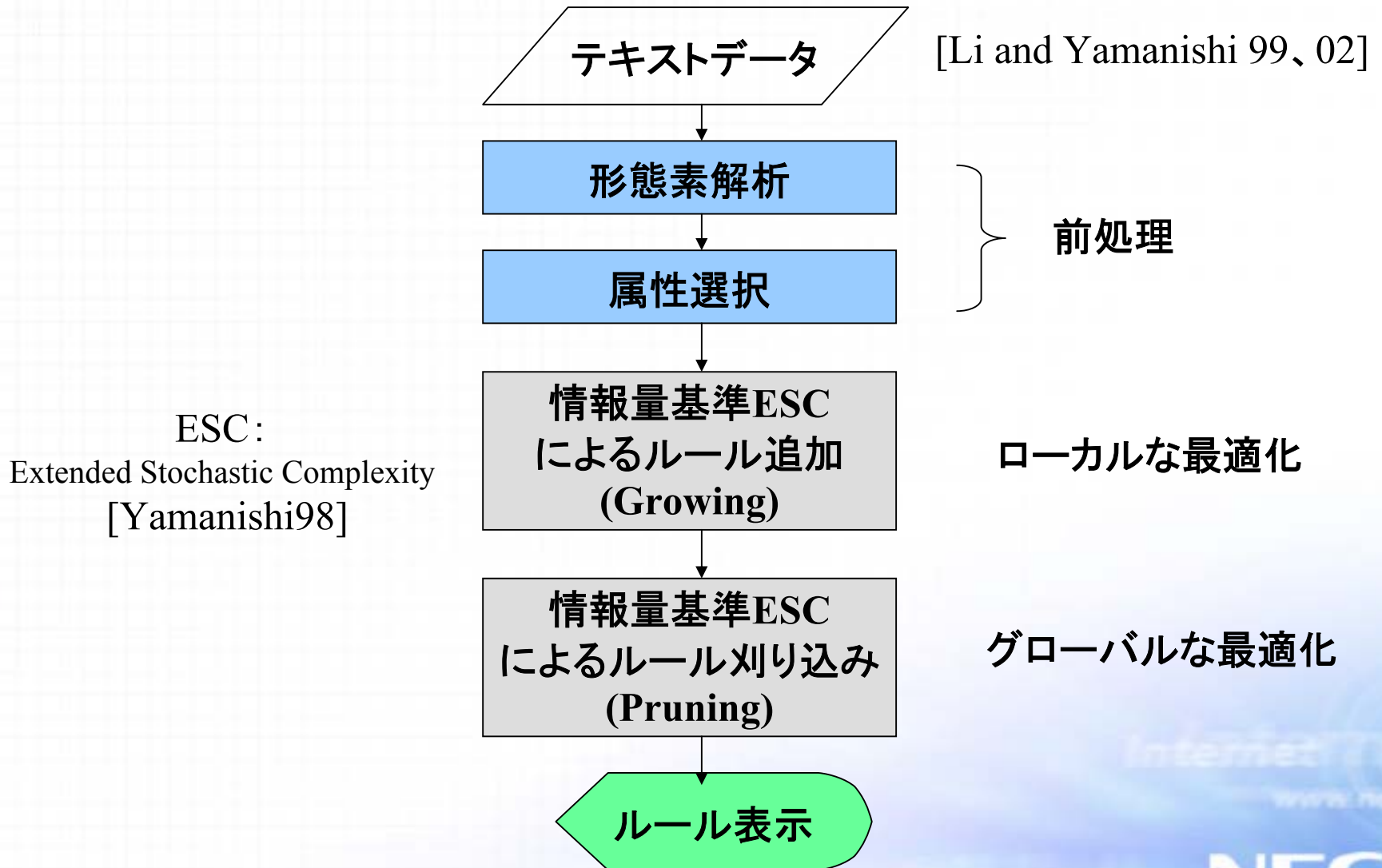
[Li and Yamanshi 99、02]

- **カテゴリ=分類対象** (ex. tradeであるか否か)を指定
- **属性** (=調べる単語)を指定。テキストを属性が現れたか(1)現れないか(0)の**二値ベクトル**で表現
- テキストとカテゴリの対応関係を**分類ルール**として学習

分類ルールの表現.....**確率的決定リスト**

```
if A =1 & B=0 then Text = trade (確率0.8)
else if D=1 then Text = not trade (確率0.9)
.....
else Text = not trade (確率0.75)
```

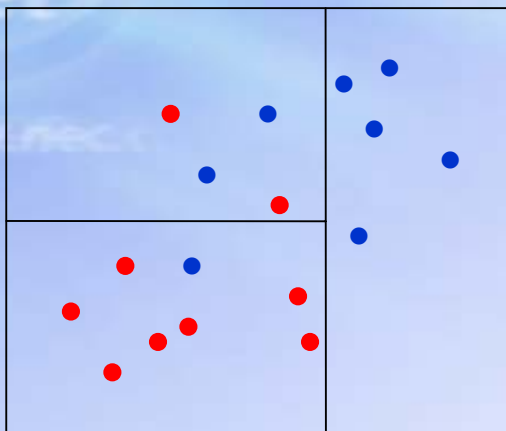
ルール学習アルゴリズムDL-ESC



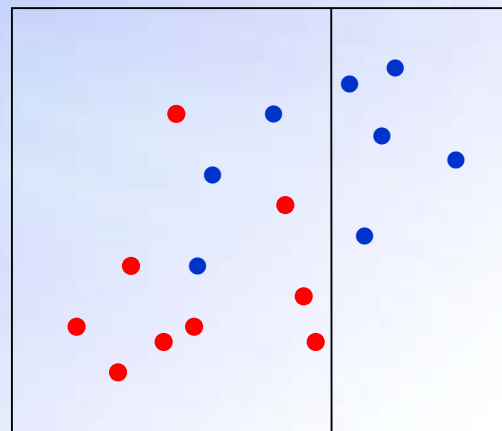
ESCに基づくルール選択

単語空間

- trade
- not trade

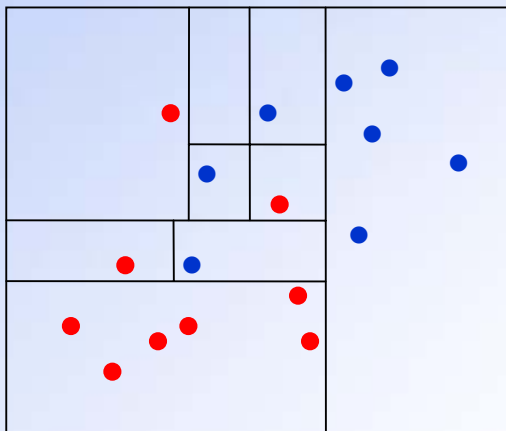


$ESC=15$

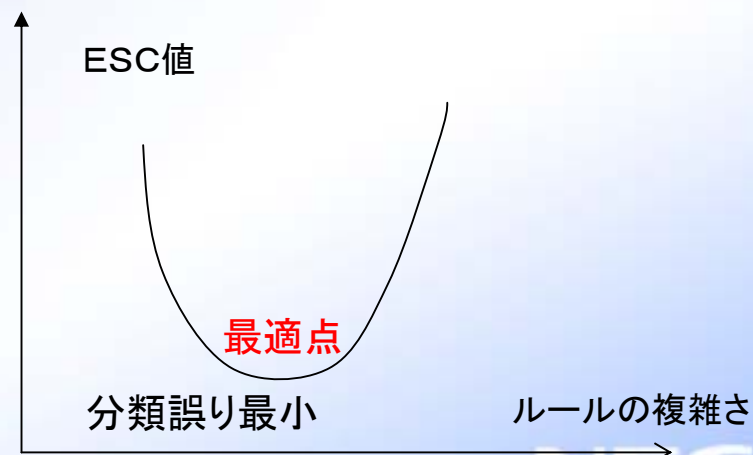


$ESC=22$

(簡単すぎるルール)



$ESC=20$ (複雑すぎるルール)



非ルールベースのテキスト分類

テキスト d のベクトル表現

$$d = (w_1, w_2, \dots, w_n)$$

Tf-idf

$$w_i = \log(1 + \text{テキスト } d \text{ における単語 } i \text{ の頻度})$$

$$\times \log(\text{全テキスト数} / \text{単語 } i \text{ を含むテキスト数})$$

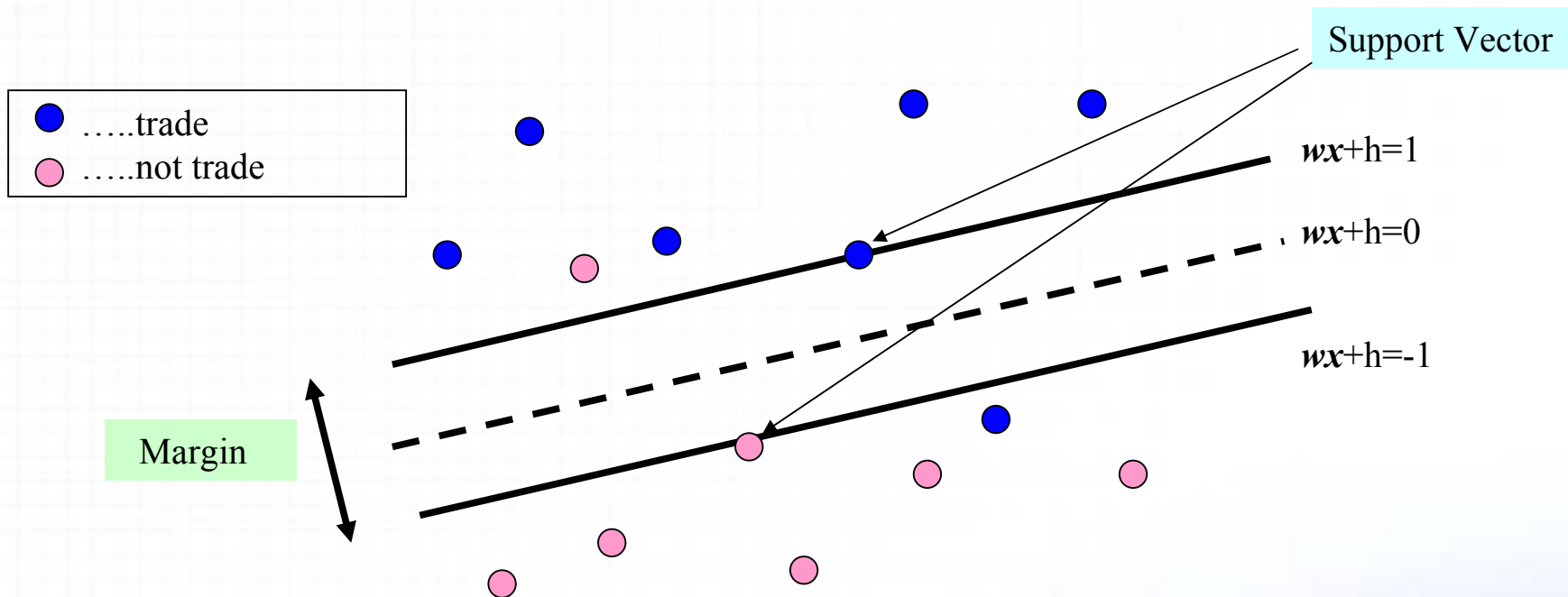
$$d \text{ と } e \text{ の類似度} = \cos(d \text{ と } e \text{ のなす角}) = \frac{d \cdot e}{|d| |e|}$$

- ・コサイン法
- ・k-NN
- ・ニューラルネットワーク
- ・SVM

等等 多数

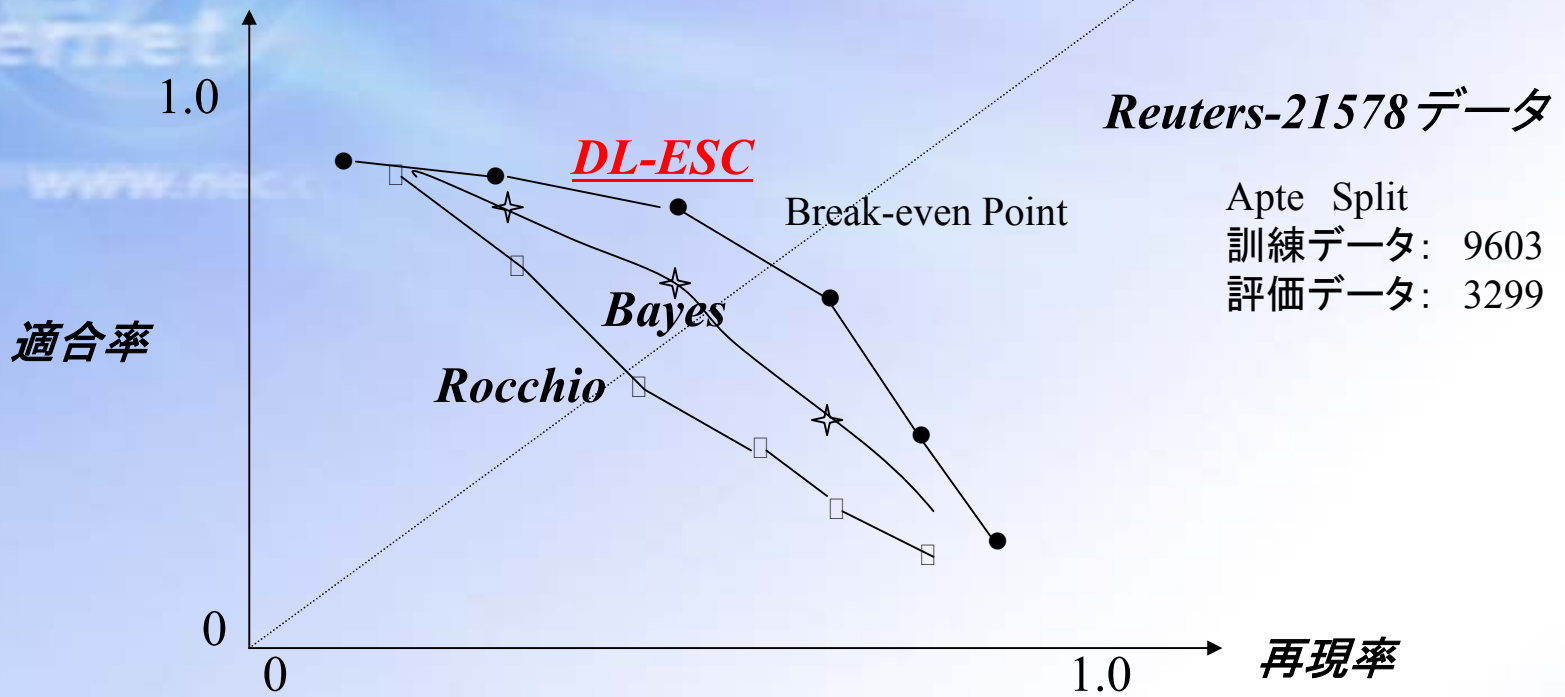
SVM(Support Vector Machine)

[Vapnik95],[Joachims98]



- Margin が最大となる超平面でカテゴリを分離
- 構造的リスク最小化の保証 = 未知データの予測誤差小
- 次元の呪い(次元の指数の計算時間)から逃れられる

テキスト分類エンジンの性能比較



再現率 = 正しく分類できたテキスト数 / 正しく分類すべきテキスト数

適合率 = 正しく分類できたテキスト数 / 分類できたテキスト数

Rule-Based	Break-even Point	Non-rule Based	Break-even Point
DL-ESC	82.0%	SVM	84.1%
DL-SC	78.3%	Bayes	77.3%
BayesNet	80.0%	BIM	74.7%
C4.5	79.4%	Rocchio	62.5%

[Li and Yamanishi 2002]

3. マーケティング知識の発見

ブランド	満足度	年代	イメージの自由記述
A社 セダンA	1	20		高級車の中で最高。
B社 セダンB	2	30		スタイルが良い。
C社 外車C	1	40		お金持ち、値段が高い。
D社 外車D	3	20		ファミリーの中でもスポーティな感じ。
E社 ワゴン車E	1	40		速い。硬い。
F社 ワゴン車F	2	30		重そう。強そう。
A社 セダンA	1	50		普通車。よく見かける。

テキストDB
(例:車のアンケートデータ)

自由記述アンケート分析ツール(CodeName)
 ※TopicScopeとして製品化
<http://www.sw.nec.co.jp/soft/TopicScope>

特徴分析
 (対象物に固有な表現を抽出、単純な頻度分析とは異なる)

A車の特徴は

- “乗り心地がよい”
- “運転しにくい”
- ⋮

心地良い車A	高級感
車C 車B 庶民的	走り重視

対応分析
 (複数の対象物とその特徴語の相互関係をポジショニング)

目的・用途

- マーケティングリサーチ
- CS調査

効果

- 分析工数の劇的削減
- 知識発見

テキストマイニングの機能

■ 特徴語分析

- ・・・カテゴリ特有に現れる単語/フレーズを抽出

■ 共起語分析

- ・・・特徴語と共起する単語/フレーズを抽出

■ 典型文解析

- ・・・カテゴリを代表するテキストを順にリストアップ

■ 対応分析

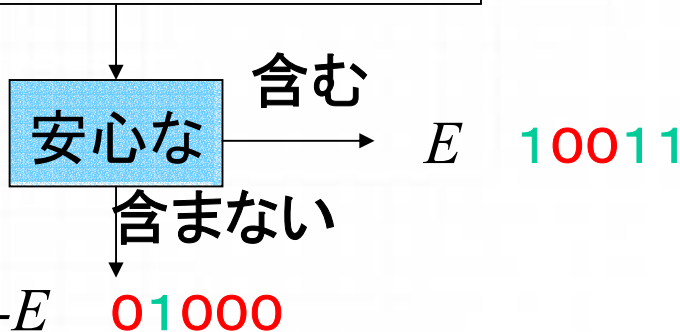
- ・・・複数のカテゴリ間の相対位置をマッピング

特徴語分析

[Li and Yamanishi 98, 01]

D : 1010000110
1: A商品 0: その他

10個のテキスト
データ



情報量規準

$I(E) + I(D-E) \rightarrow$ 小 \Rightarrow “安心な” はカテゴリ“A商品”の特徴語

$I(x) = mH(m_1/m) + (1/2)\log(m\pi/2)$ m : データ数、 m_1 : 1の出現数

.... 確率的コンプレキシティ(SC) \sim データ圧縮の規準

$I(x) = \min\{m_1, m - m_1\} + \lambda(m \log m)^{1/2}$

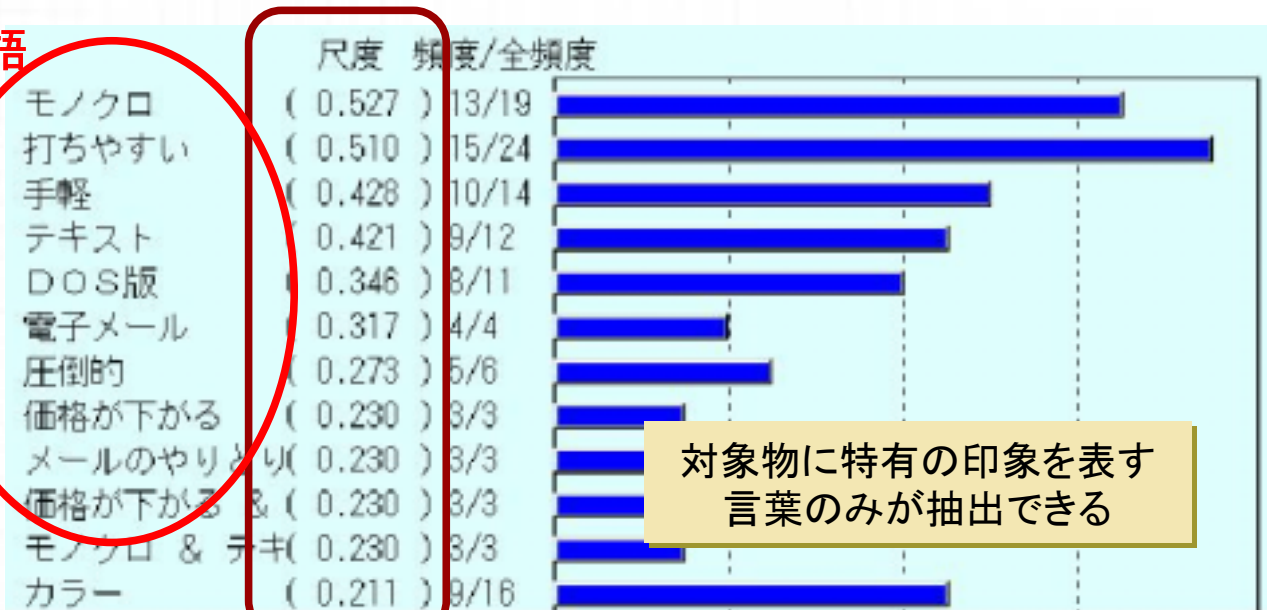
.... 拡張型確率的コンプレキシティ(ESC) \sim 予測誤差最小の規準

“安心な” の情報利得: $\Delta I = I(D) - (I(E) + I(F))$

特徴語分析の例

「PDA製品A」に関する
肯定意見における特徴語

ESCに基づく特徴語抽出結果 ↓



対象物に特有の印象を表す
言葉のみが抽出できる

特徴語は「拡張型確率的コンプレキシティ」に基づく
情報利得を計算することで求めている

この尺度は、全意見セットと比べて、着目意見セット(この
例では「PDA製品A」の肯定意見)に偏って出現する単語
について値が大きくなる

[参考]単純頻度による特徴語抽出結果 ↓

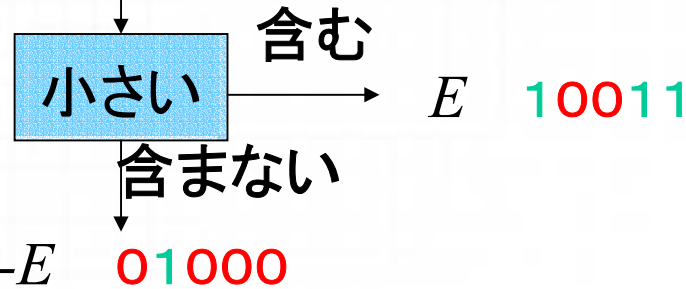


いろいろな対象物に共通する
言葉も抽出されてしまう

共起語分析

D: 1010000110
 1: “キーボード”を含む
 0: “キーボード”を含まない

10個のテキスト
データ



情報量規準

$I(E) + I(D-E) \rightarrow$ 小 \Rightarrow “キーボード”と“小さい”の共起性大

例

単語	共起単語
キーボード	打ちやすい
キーボード	小さい
キーボード	入力

典型文分析

[Morinaga, Yamanishi, Tateishi, Fukushima 02]

テキスト $s = w_1, \dots, w_N$ (w_i : 単語またはフレーズ)

$$\text{Score}(s) = \frac{p(c) \prod_{i=1}^N p(w_i | c)}{\sum_c p(c) \prod_{i=1}^N p(w_i | c)}$$

カテゴリCにおける
テキストsの典型文スコア

$$p(c) = \frac{N_c + \beta}{\sum_c N_c + |C| \beta}$$

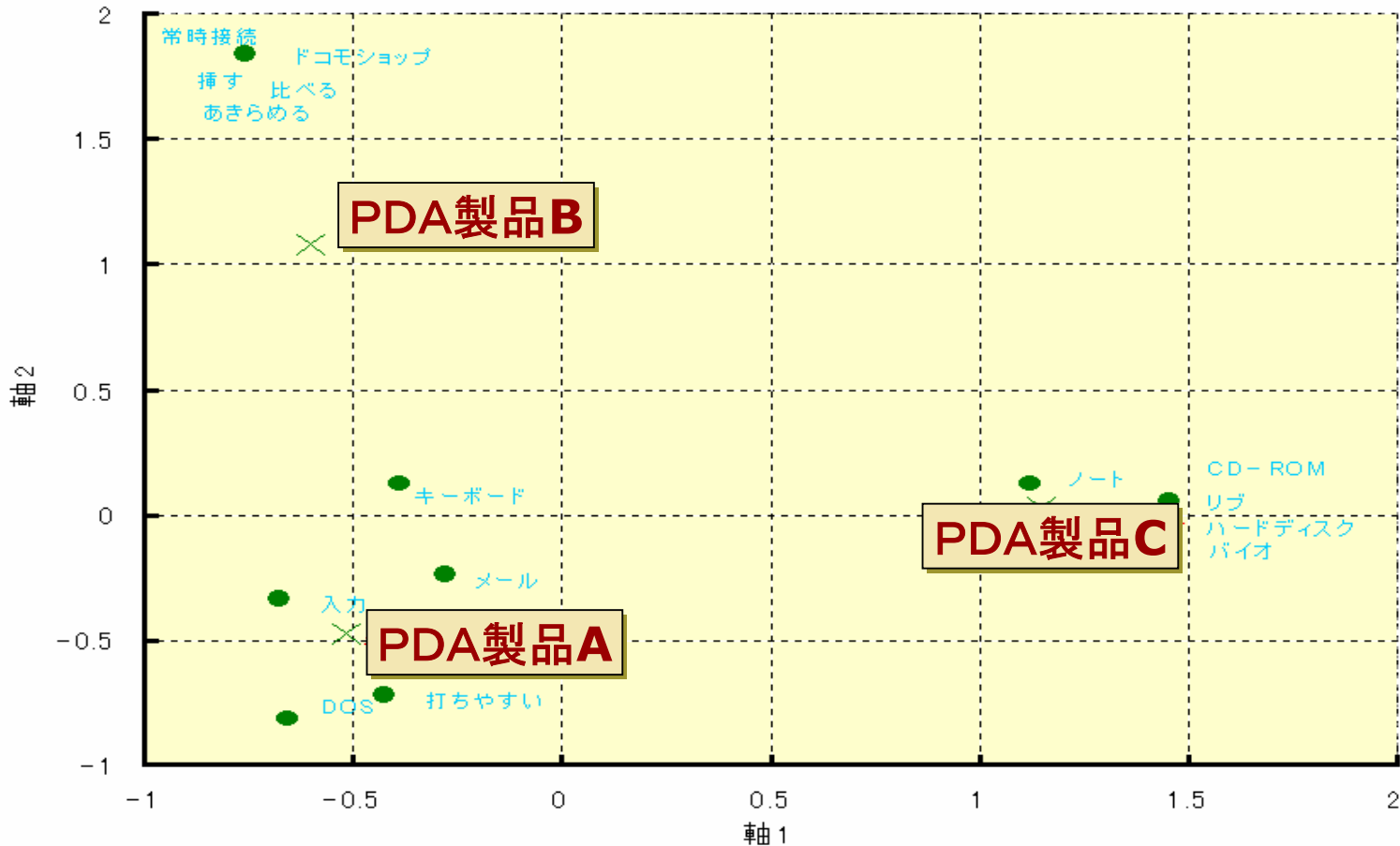
N_c : カテゴリCに属するテキスト数

$$p(w_i | c) = \frac{m_w + \beta}{\sum_w m_w + |W| \beta}$$

m_w : カテゴリCに属するテキスト
の中に含まれる単語wの数

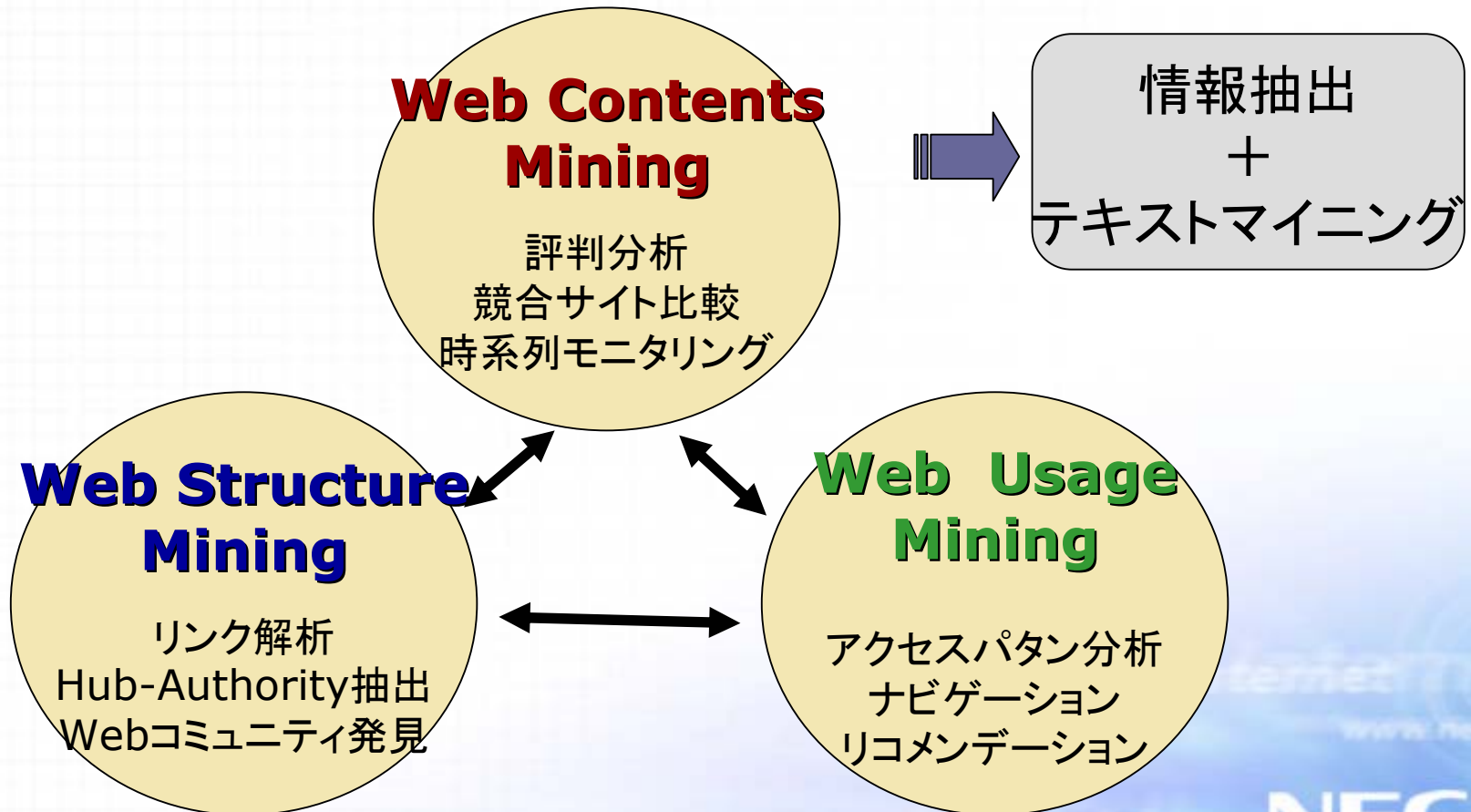
対応分析

各商品名と関連の強い特徴語を2次元マップ上に配置



4. 評判分析とWebマイニング

Web マイニングの分類 [Kosala and Blockeel 2000]



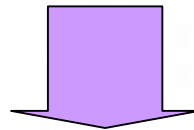
評判分析

Web上の意見の収集から分析までを自動化

評判検索.... Webからの評判検索・意見抽出
[立石、石黒、福島 01]

+

テキストマイニング (SurveyAnalyzer)
.... 特徴単語抽出によるテキスト分類
[Li and Yamanishi 01]



評判分析....評判の原因を分析

[Morinaga, Yamanishi, Tateishi, Fukushima 02]

～ マーケティング⇒工数激減・知識発見

掲示板, レビューサイト,
個人サイト, 日記サイト等



意見収集対象:
携帯端末A・B・C

分析対象:
携帯端末Aの肯定意見

① 評判検索

② テキスト
マイニング

携帯端末Aに関する意見:

携帯端末Aのキーボードは 打ちやすい	○肯定
携帯端末Aは乾電池で長時間駆動が 魅力	○肯定
.....	

携帯端末Bに関する意見:

携帯端末Bのデザインが 最高	○肯定
携帯端末Bなんて 大嫌い	×否定
.....	

携帯端末Cに関する意見:

携帯端末Cは 安い	○肯定
携帯端末Cは重いので 嫌い	×否定
.....	

携帯端末Aの
肯定意見の特徴語:

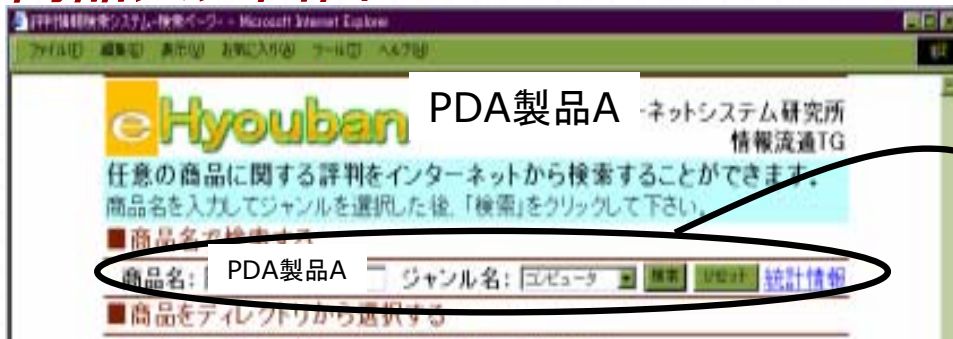
電子メール
キーボード
長時間
.....

Internet
www.nec

NEC

評判検索：結果出力

商品入力画面



商品名: PDA製品A
ジャンル: コンピュータ

検索結果画面

適正值

評価

抽出した意見

- 5.338 ☺ > 幸い、PDA製品 A はそこそこ速いようですし、キーも結構打ちやす
6.338 ☺ れないと思うのです。>> 幸い、PDA製品 A はそこそこ速いようですし、キー
も結構打ちや
[モバイルギア使用記](#)
http://plaza6.mbn.or.jp/~yath_yas/mgdaily.htm
- 7.338 ☹ PDA製品A は結構キーが重い・インタートップのキーボ
[リレー連載 第21講](#)
<http://watch.impress.co.jp/pc/docs/article/970114/relay21.htm>
- 8.313 ☺ これはイカすぞ PDA製品A はじめにお断りしておきたいが、私はNECのちい
さいモノが大好きである。古くはHANDY-98を所有

評判検索：ラベル化

■ 商品分野ごとの評価表現の辞書を作成

商品カテゴリ	評価表現リスト
共通	好き、良い、良くない、勧め、最高、満足だ…
書籍	面白、名作、読みにく、分かりやす、違和感
コンピュータ	速い、壊れやす、うるさ、不安定、信頼で…

■ 構文的特徴を考慮して意見らしさを判定

ID	適正值判定ルール(正規表	ルールの意味
1	<u>商品名</u> .*(は が も).* <u>評価表現</u>	格助詞が存在
2	<u>商品名</u> .*(。 . ? !).* <u>評価表現</u>	別の文に存在
3	<u>評価表現</u> _{0,12} <u>商品名</u>	接近して存在
4	<u>評価表現</u> .*(¥? ?)	文末が疑問符

評判検索からテキストマイニングへ

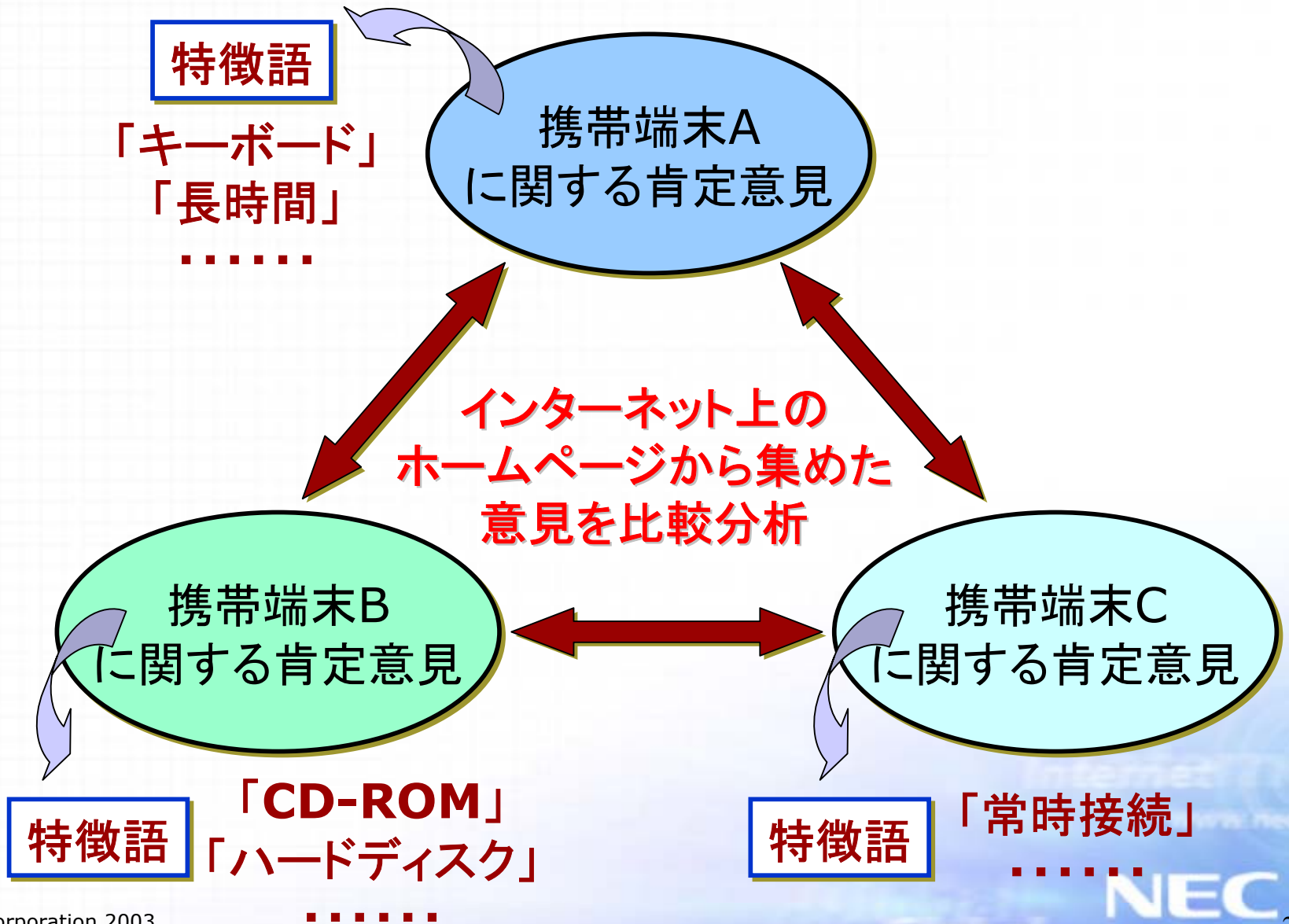
	ラベル		テキスト	
	商品名	肯定/否定	適性値	評判・意見
正例	PDA 製品A	肯定	0.75	* * は使いやすくて素晴らしい * * *

負例	PDA 製品A	否定	0.82	* * は重くて持ち運びに困る * * *



特定のラベルの組み合わせ(カテゴリ)を識別する特徴的表現を
マイニング！

評判分析例



評判分析の応用

企業における マーケットリサーチ

新商品開発
・商品改良

当社の現商品について
ユーザはどんな不満をも
っているのだろう

競合
調査

当社の競合商品の
評判はどうだろう

今度の新CMは好評だ
ろうか、あの俳優の好
感度はどうだろう

広告・キャンペーン等の
効果把握

掲示板で悪評が
立っていないか

誹謗中傷・悪評対策

一般ユーザ向けの アドバイス提供

この商品とあの商
品ではどちらが評
判が良いかな

商品購入
支援

欧州に旅行に行く
のにどこが評判
が良いだろう

旅行計画
・行動支援

今度訪問するあ
の会社はどんな
評判なのだろう

会社・個人の
信用調査



5. トピック分析と情報監視

テキストストリームからのトピックの抽出

TDT: Topic Detection and Tracking

~DARPAの研究プログラムの一つ。年1回のCompetition

A Topic is.....

a seminal event or activity, along with all directly related events and activities

A Story is.....

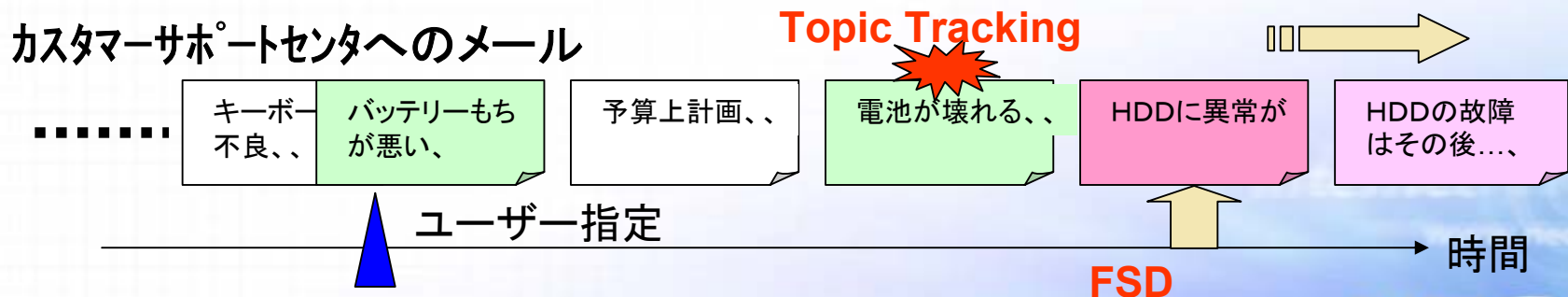
a topically cohesive segment of news that includes two or more DECLARATIVE independent clauses about a single event

- 異なるテキストのストリームからのトピック分析
- 同一テキスト内でのトピック分析

トピック分析の5大問題

- トピック： 特定のイベント。（例： 首相訪朝、不審船引き上げ、、、等）
- ストーリー： 単一のトピックについて述べている文章。（例：新聞の記事一つ）

- Story Segmentation: 長いテキストをストーリーごとに分割する
- Link Detection: 二つのストーリーが同じトピックかどうかを判定する
- Topic Detection: ストーリー集合を、トピックに関してクラスタリングする
- Topic Tracking: 指定されるストーリーと同じトピックのものをトラッキングする
- First Story Detection (FSD): 新しいストーリーの出現を検出する



Topic Tracking の現行技術

基本アルゴリズム:

- テキストを単語の集合とみなしベクトルで表現

$$d = (w_1, w_2, \dots, w_n)$$

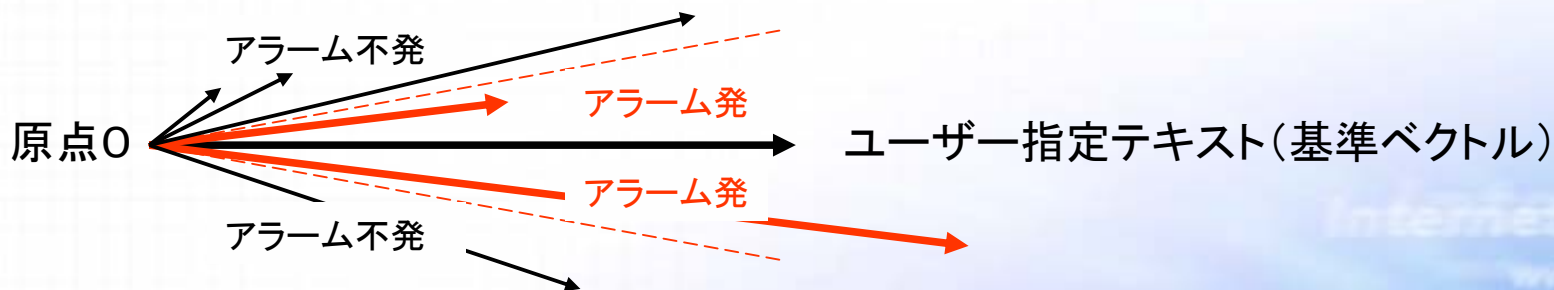
Tf-idf $w_i = \log(1 + \text{テキスト } d \text{ における単語 } i \text{ の頻度})$

$\times \log(\text{全テキスト数} / \text{単語 } i \text{ を含むテキスト数})$

- テキスト間の類似度をベクトルの角度(コサイン)で定義

$$d \text{ と } e \text{ の類似度} = \cos(d \text{ と } e \text{ のなす角}) = \frac{d \cdot e}{|d| |e|}$$

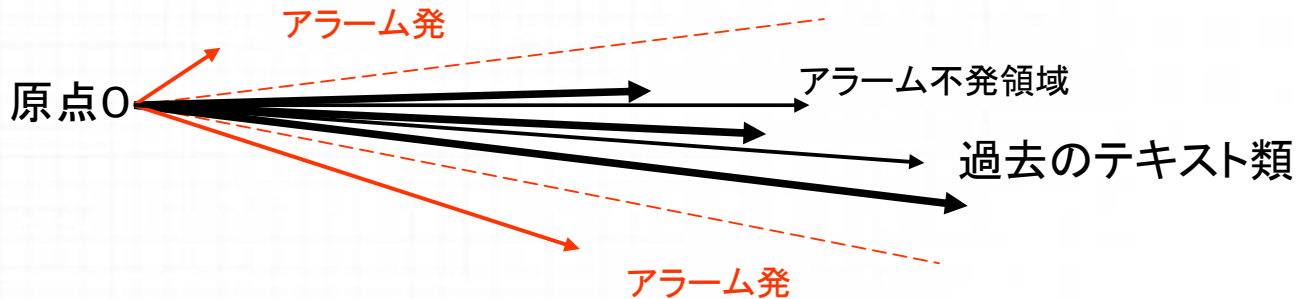
- ユーザー指定テキストとの類似度が閾値以上の新テキストが来たらアラーム



FSD の現行技術

基本アルゴリズム:

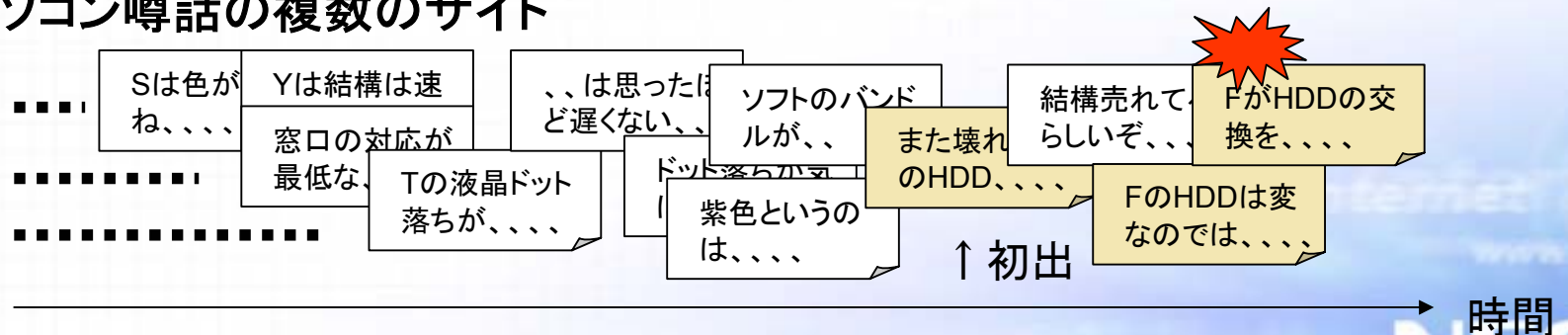
- Topic Tracking と同様の文書ベクトル表現に対して、過去のどのテキストとも類似度が閾値以下であるテキストにアラームを出す



改良アルゴリズム:

- 初出トピックのその後の出現回数が閾値を越えたらアラーム → 情報潮流発見

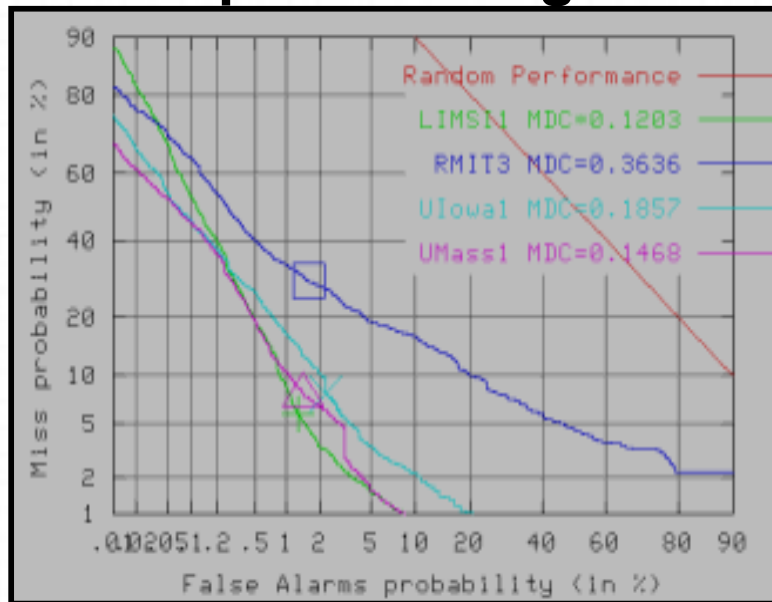
パソコン噂話の複数のサイト



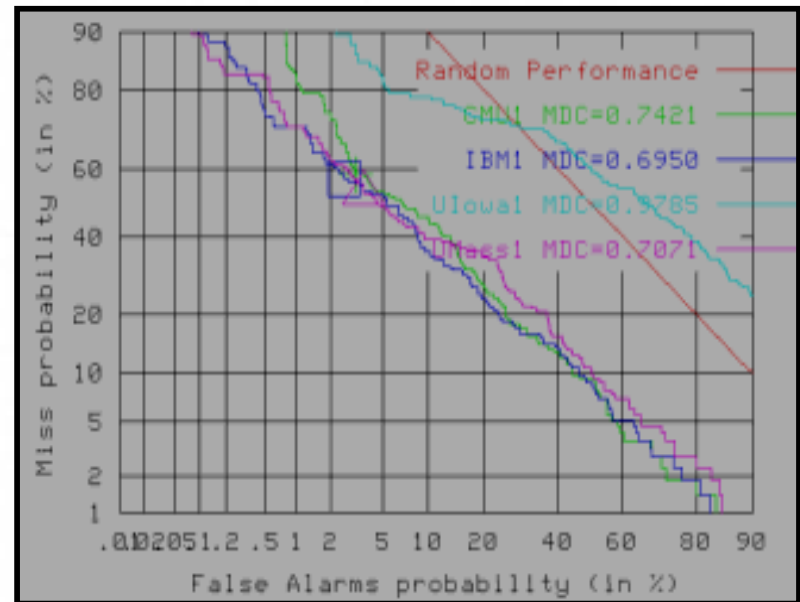
トピック分析のレベル

TDT evaluation 2001の結果

Topic Tracking



FSD



J.Fiscus: Overview of the TDT 2001 Evaluation and Results

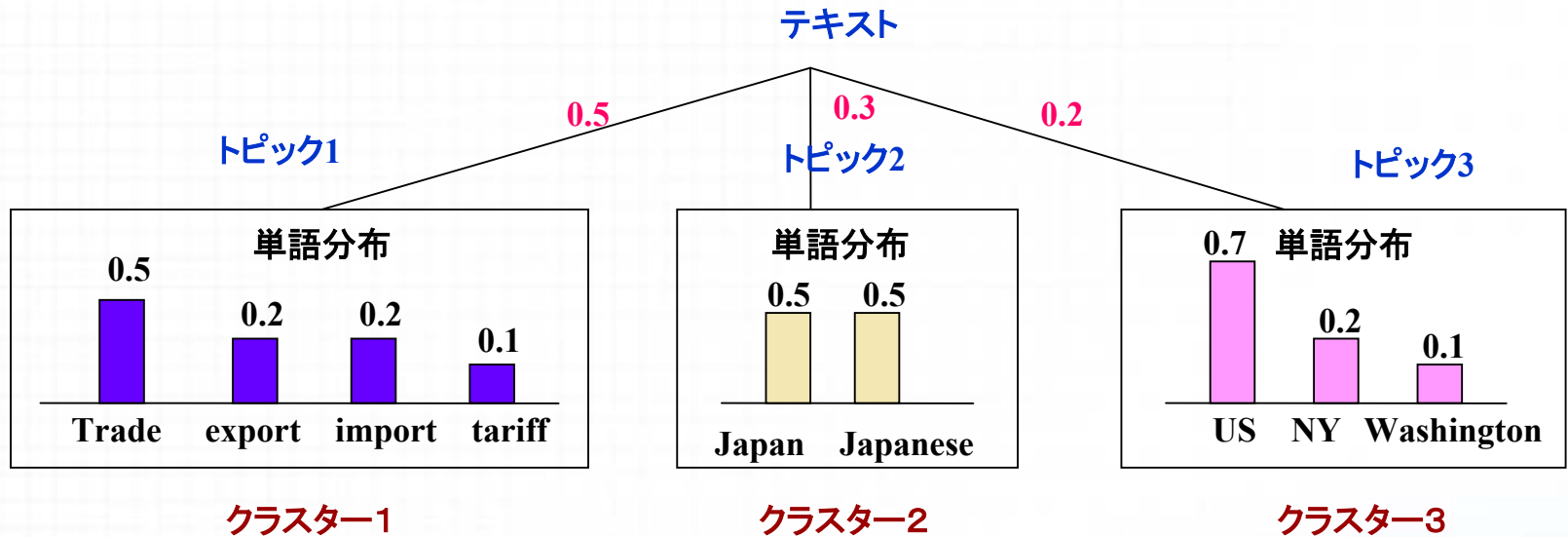
- ・Breakeven-pointにて95%超の精度
- ・データは主にニュース原稿。～比較的簡単な課題(似たトピックが少ないテキスト集合)

- ・Breakeven-pointは約70%の精度にとどまる (FSDは五大問題中最難問)

テキスト内でのトピック分析

●トピック: 話項目...単語クラス(クラスタ)で表現する

例: **trade**: export, import, tariff, GATT, protectionist



K: トピックの集合

$P(k)$: **K**上の確率分布

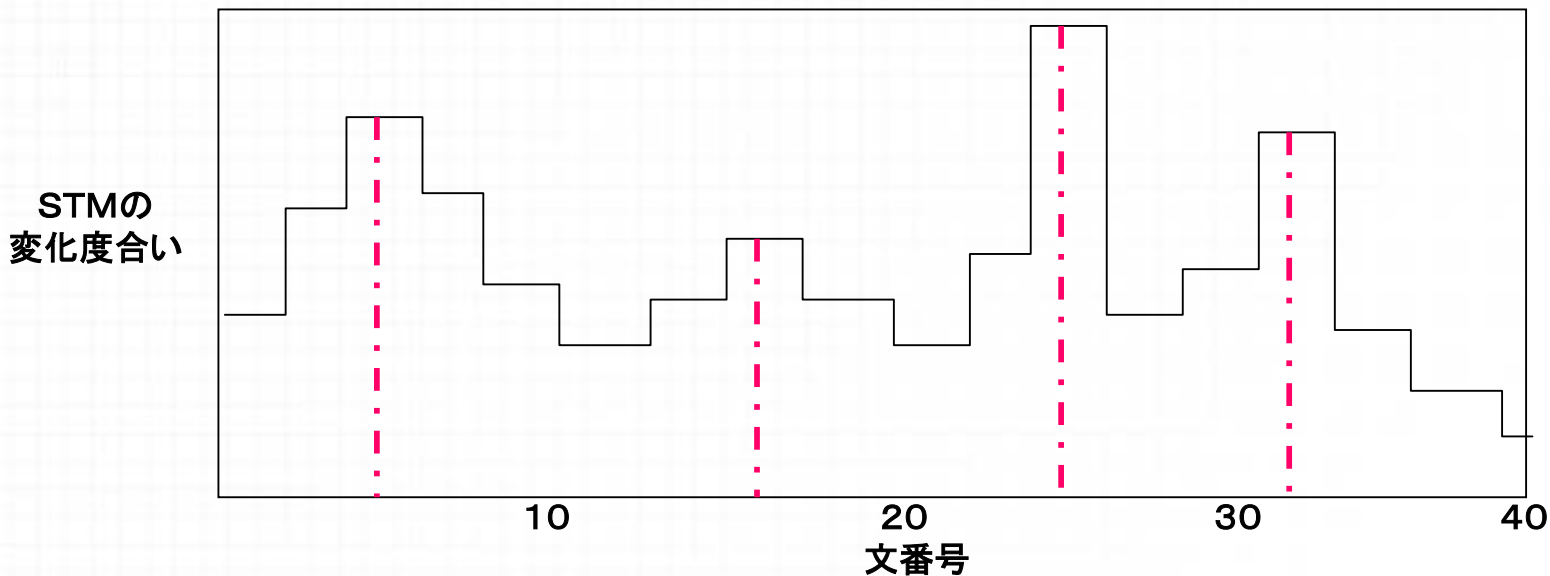
$P(w | k)$: トピック(クラスタ) k 内の単語の確率分布

確率的トピックモデル [Li and Yamanishi 00,03]

....Finite Mixtureを用いた単語分布の表現

$$P(w) = \sum_{k \in K} P(w | k) P(k)$$

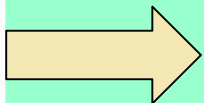
テキストセグメンテーション



文XにおけるSTMの変化度合い

= Xの**前**の文章のSTM P_L と Xの**後**の文章 P_R のSTM
の統計的距離

$$D(x) = \sum_{\omega} |P_L(\omega) - P_R(\omega)|$$



$D(x)$ が極大になる文xで分割

テキストセグメンテーションの例

ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RIFT (25-MAR-1987)

block 0 ----- **trade-export-tariff-import(0.12)** **Japan-Japanese(0.07)** **US(0.06)**

- 1 They told Reuter correspondents in Asian capitals a U.S. move against Japan might boost ...
- 2 But some exporters said that while the conflict would hurt them in the long-run, in the ...
- 3 The U.S. has said it will impose 300 mln dlrs of tariffs on imports of Japanese electronics ...
- 4 Unofficial Japanese estimates put the impact of the tariffs at 10 billion dlrs and spokesmen ...
- 5 "We wouldn't be able to do business," said a spokesman for leading Japanese electronics ...
- 6 "If the tariffs remain in place for any length of time beyond a few months it will mean the ...

block 1 ----- **trade-export-tariff-import(0.17)** **US(0.09)** **Taiwan(0.05)**

- 7 In Taiwan, businessmen and officials are also worried.
- 8 "We are aware of the seriousness of the U.S. threat against Japan because it serves as a ...
- 9 Taiwan had a trade surplus of 15.6 billion dlrs last year, 95 pct of it with the U.S.
- 10 The surplus helped swell Taiwan's foreign exchange reserves to 53 billion dlrs, among the ...
- 11 "We must quickly open our markets, remove trade barriers and cut import tariffs to allow ...
- 12 A senior official of South Korea's trade promotion association said the trade dispute between ...
- 13 Last year South Korea had a trade surplus of 7.1 billion dlrs with the U.S., up from 4.9 ...
- 14 In Malaysia, trade officers and businessmen said tough curbs against Japan might allow ...

block 2 ----- **Hong-Kong(0.16)** **trade-export-tariff-import(0.10)** **US(0.04)**

- 15 In Hong Kong, where newspapers have alleged Japan has been selling below-cost semiconductors, ...
- 16 "That is a very short-term view," said Lawrence Mills, director-general of the Federation of ...
- 17 "If the whole purpose is to prevent imports, one day it will be extended to other sources...
- 18 The U.S. last year was Hong Kong's biggest export market, accounting for over 30 pct of ...

block 3 ----- **trade-export-tariff-import(0.14)** **Button(0.08)** **Japan-Japanese(0.07)**

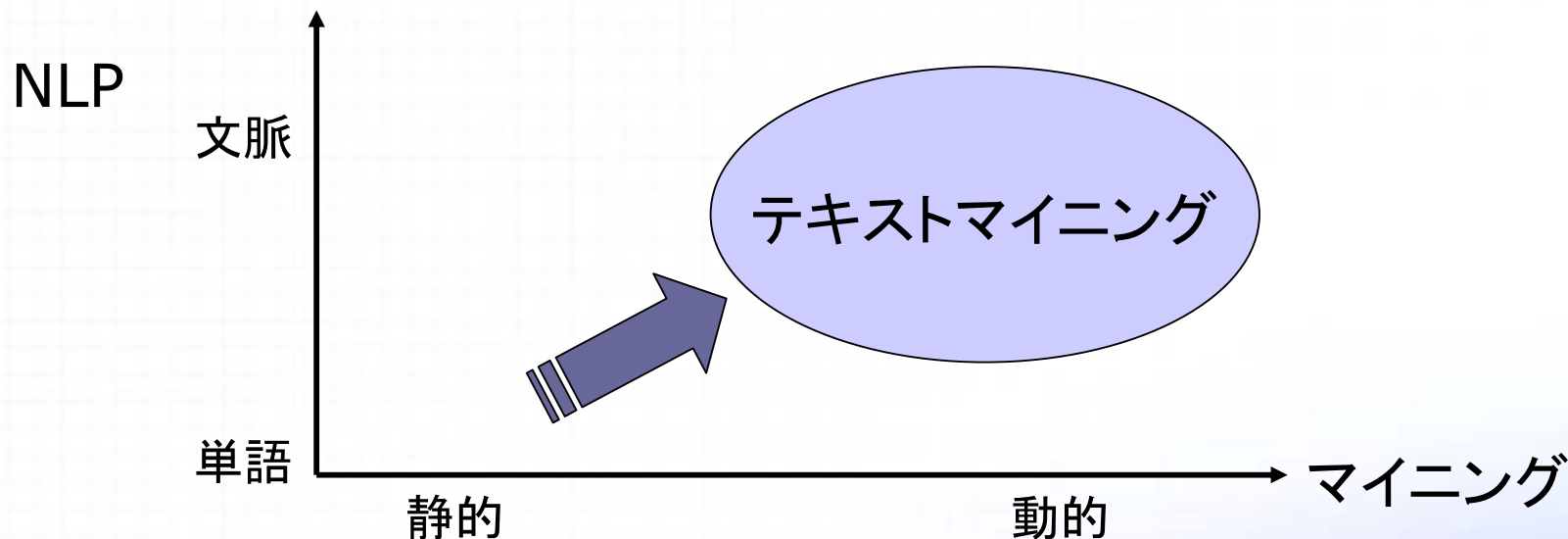
- 19 The Australian government is awaiting the outcome of trade talks between the U.S. and Japan ...
- 20 "This kind of deterioration in trade relations between two countries which are major trading ...
- 21 He said Australia's concerns centered on coal and beef, Australia's two largest exports to ...
- 22 Meanwhile U.S.-Japanese diplomatic manoeuvres to solve the trade stand-off continue.

トピックの
確率分布同定

文書自動
分割

6. テキストマイニング: Challenges

- 文脈マイニング (単語/句から文脈へ)
- オンラインピック分析 (初出表現、Novelty Detection)



● Multi-Mediaとの融合 ⇒ Multi-Media マイニング

● リンク解析、ログ解析との融合

⇒ Webマイニング、Relational マイニング

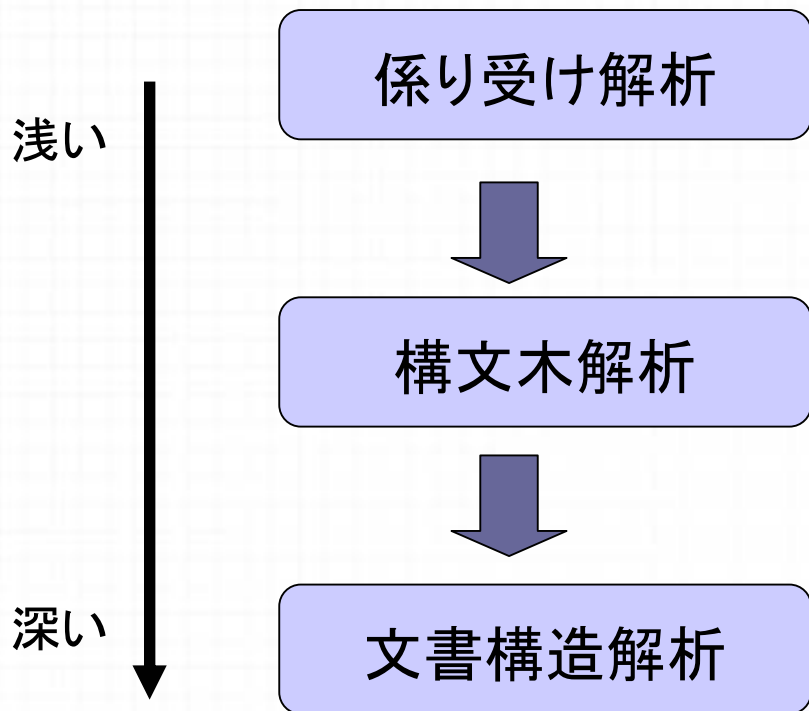
文脈マイニング

文脈解析+マイニング

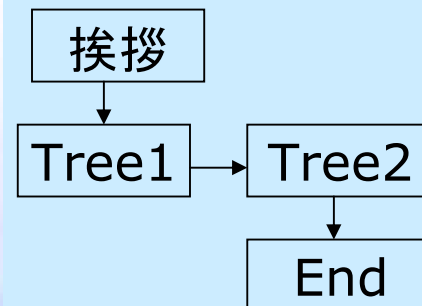
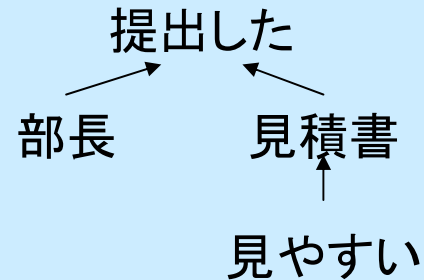
文章

拝啓、昨日A商事を訪問して、...
その後、...商談がまとまった。以上

部長に見易い見積書を提出した。



部長→提出した
見易い→見積書
見積書→提出した



7. おわりに

- 成熟したテキスト分類、これからのトピック分析
- テキストマイニング応用分野が急拡大 (CRM, マーケティング市場、Forensics, etc.)
- テキストマイニング技術は今後、文脈とダイナミクスを取り入れて発展するだろう
- Webマイニング、Relationalマイニング、マルチメディアマイニングの中で技術融合の可能性

8.参考文献

【全般】

1. 金、村上、永田、大津、山西:「データとテキストのマイニング」岩波書店「統計科学のフロンティア」シリーズ10、2003.
2. 山西健司:情報論的学習理論の現状と展望、情報処理、vol.42, No.1, pp:9--15, 2001.
3. 山西健司:データ・テキストマイニングの最新動向ー外れ値検出と評判分析を例にー、応用数理、vol.12, No.4,p.7-22,2002..

【情報理論、学習の基礎】

1. J.Rissanen: Fisher information and stochastic complexity, *IEEE Trans.on Information Theory*, 42(1), pp:40-47 (1996).
2. K.Yamanishi: "A Decision-theoretic Extension of Stochastic Complexity and Its Applications to Learning," *IEEE Trans. on Information Theory*, vol.44, 4, p.1424-1439, 1998.
3. 麻生、津田、村田:「パターン認識と学習の統計学」岩波書店「統計科学のフロンティア」シリーズ6、2003
4. 韓、小林:「情報と符号化の数理」岩波講座応用数学 対象11
5. 山西、韓: MDL入門: 情報理論の立場から、人工知能学会誌, p.427-434, vol 7(3), May 1992.
6. 山西健司: 拡張型確率的コンプレキシティと情報論的学習理論, 応用数理、vol.8, No.3, p.14-29, 1998.
7. 山西健司: 統計的モデル選択と機械学習, 計測と制御、vol.38, p.420-426, 1999.

【テキストマイニング一般】

1. R.Feldman: Mining unstructured data, Tutorial notes for ACM SIGKDD 1999 International Conference on Knowledge Discovery and Data Mining (KDD1999)
(<http://www.acm.org/pubs/citations/proceedings/ai/312179/p182-feldman/>)
2. M.A.Hearst: Untangling text data mining, in Proc.of the 37th Annual Meeting of the Association for Computational Linguistics(ACL99) (1999).
(<http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html>)
3. SIGKDD: <http://www.acm.org/sigkdd/>
4. 人工知能学会誌 Vol.16, No.2 (2001年3月) 特集「テキストマイニング」

【テキスト分類関連】

1. C.Apte, F.Damerau, S.M.Weiss: Towards language independent automated learning of text categorization models in *Proc. of Annual ACM SIGIR Conference on Research and Development on Information Retrieval(SIGIR94)*, pp.24-30,1994.
2. W.Cohen and Y.Singer: Context-sensitive learning methods for text classification, in *Proc.of SIGIR96*, pp:307-315 (1996).
3. S.Dumais, J.Platt, D.Heckerman, and M.Shami: Inductive learning algorithm and representation for Text categorization, in *Proc.of the 7th Int'l Conf. on Information and Knowledge Management(CIKM98)*, pp:148-155 (1998)
- 4.T.Joachims: Text categorization with support vector machines: Learning with many irrelevant features, in *Proc. ECML '98* (1998).
- 5.G.Kar and L.J.White: A distance measure for automatic document classification by sequential analysis, *Information Processing and Management*, 14, pp:57-69 (1978).
- 6.H.Li and K.Yamanishi: Text classification using ESC-based stochastic decision lists, in *Proc. of 8th International Conference on Information and Knowledge Management (CIKM'00)*, pp: 122-130, (2000).
7. .H.Li and K.Yamanishi: "Text classification using ESC-based decision lists," *Information Processing and Management*, .Vol. 38/3, pp 343-361, 2002.
8. Reuters21578 Text Categorization Collection: <http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>
9. J.Rocchio: Relevance feedback information retrieval, in Gerard Salton Editor, *The Smart Retrieval System -Experiments in Automatic Document Processing*, pp:313-323, Prentice-Hall (1971).
- 10.R.E.Schapire,Y.Singer,andA.Signal: Boosting and rocchio applied to text filtering, in *Proc. of SIGIR98*, pp:215-223,1998.
- 10.佐藤、池田、中田、長田:CRM分野へ向けた日本語処理機能のミドルウェア化 言語処理学会第9回年次大会発表論文集 pp.109-112,2003年3月
- 11.永田、平田: テキスト分類—学習理論の「見本市」—、情報処理、vol.42(1), pp:32-37 (2001).
- 12.李: テキスト分類、計測と制御, Vol.38,pp:456-460 (1999).

【マーケティング知識の発見】

1. .H.Li and K.Yamanishi: "Mining from Open Answers in Questionare Data ," *Proc. of the 7th ACM Int'l. Conf. on Knowledge Discovery and Data Mining(KDD2001)*, ACM Press, pp:443-449, 20
2. TopicScope: <http://www.sw.nec.co.jp/soft/TopicScope>
3. 森永、山西: "テキストマイニングによる自由記述アンケート分析" 計測と制御、第41巻第5号、pp:354-357,2002.
4. Yamanishi: and H.Li: "Mining Open Answers in Questionare Data," *IEEE Intelligent Systems*, pp:58-63、Sept/Oct, 2002

【評判分析とWebマイニング】

1. G.W.Flake, S.Lawrence, and C.L.Giles: Efficient identification of web communities, in *Proc. of the 6th ACM Int'l Conf. on Knowledge Discovery and Data Mining(KDD2000)*, pp:150-160, ACM Press, 2000.
2. R.Kosala and H.Blocheel: Web mining research: A survey. *ACM SIGKDD Explorations*, vol.2, No.1, pp:1-15, 2000.
3. B.Liu, Y.Ma, and P.S.Yu: Discovering unexpected information from competitors' web sites. in *Proc. of the 7th ACM Int'l Conf. on Knowledge Discovery and Data Mining(KDD2001)*, pp:144-153, ACM Press, 2001.
4. S.Morinaga, K.Yamanishi, K.Tateishi, and T.Fukushima: "Mining Product Reputations on the Web," in *Proc. of the 8th ACM Int'l. Conf. on Knowledge Discovery and Data Mining (KDD2002)*, pp:341-349 ACM Press, 2002.
5. 立石、石黒、福島: インターネットからの評判検索. 情報処理学会研究報告, NL153-14, pp:105-112, 2003.
6. 山西健司: Webマイニングと情報論的学習理論、 2002年情報学シンポジウム講演論文集、 pp:9-16, 2002.

【トピック分析関連研究】

1. The 2001 topic detection and tracking (tdt2001) task definition and evaluation plan. <http://www.nist.gov/speech/tests/tdt/tdt2001/evalplan.htm> 2001
2. D.Beeferman, A.Berger, and J.Lafferty: Statistical models for text segmentation, *Machine Learning*, 34, pp:177-210, 1999.
3. L.Baker, and A.McCallum: Distributional clustering of words for text classification. in *Proc. of ACM-SIGIR98*, 1998.
4. M. Hearst: Texttiling: Segmenting text into multi-paragraph subtopic passages, *Computational Linguistics*, 23(1), pp:33-64, 1997.
5. G.Salton and C.S.Yang: On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4), pp:351-372, 1973.
6. H.Li and K.Yamanishi: Topic analysis using a finite mixture model, in *Proc. of ACL Workshop on Very Large Corpus*, pp:35-44, 2000.
7. H.Li and K.Yamanishi: Topic analysis using a finite mixture model, *Information Processing and Management*, Vol.39/4, pp 521-541, 2003.