頻出パターン列挙に基づいたビジネス応用とパターンの拡張の可能性

ERATO 湊離散構造処理系プロジェクト
第1回　離散構造処理系シンポジウム
2010年5月29日（土）北海道大学

羽室　行信　関西学院大学
森田　裕之　大阪府立大学

キーワード

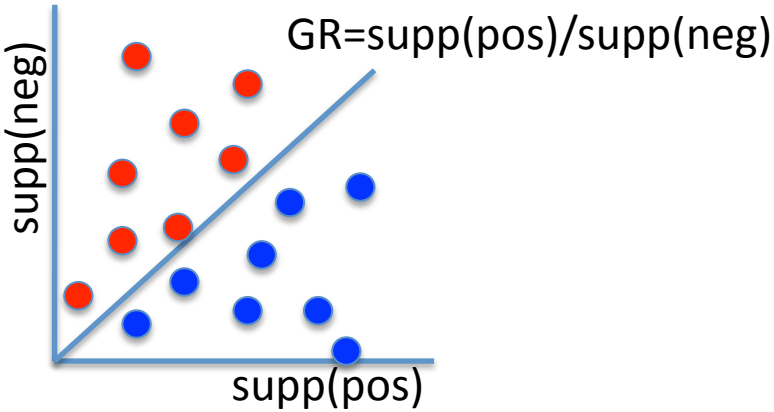
- 時間制約付きアイテム集合シーケンス
  +Emerging Patterns
- Taxonomyの自動生成（高次元データ対応）

既にapriori ベースでの実装は完了→効率性に問題あり。

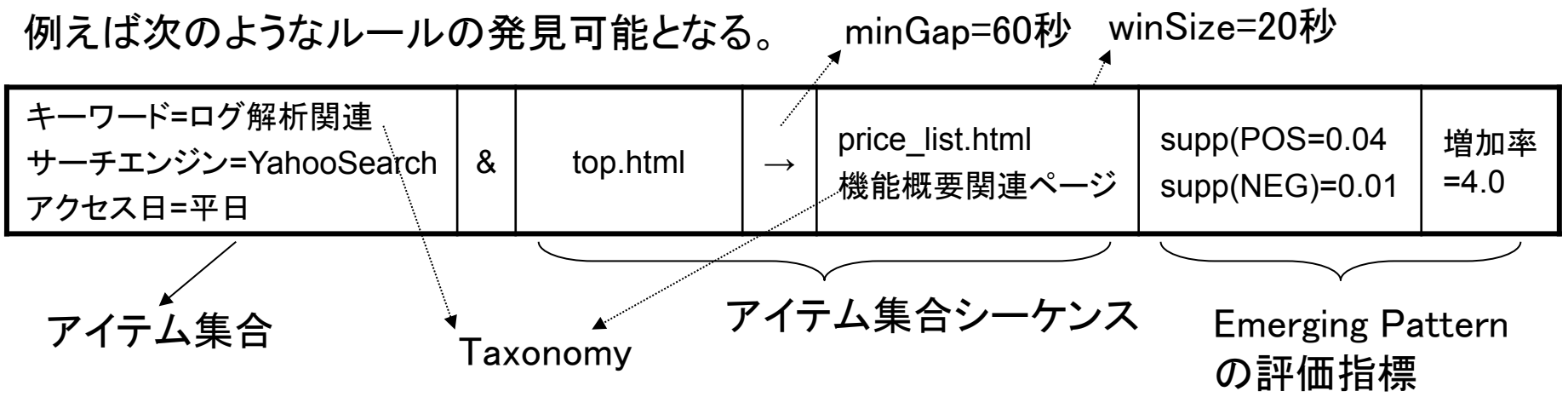
基礎研究担当の皆様：
「ZDDをベースとした より高度な離散構造の
algebraの構築」に非常に期待しております。

パターン列挙に基づくビジネス応用例　その1）
Webログ解析

- Webページの巡回パターンにより、コンバージョンを予測する。
  - アイテム集合
  - アイテム集合シーケンス
  - 時間制約：winSize, minGap, maxGap
  - Taxonomy
  - Emerging Patterns (EP)



例えば次のようなルールの発見可能となる。



- Webログ解析においては、時間制約を伴ったアイテム集合シーケンスがモデル精度に大きく貢献することが多い。

パターン列挙に基づくビジネス応用例　その2）

株式市場におけるセンチメント分析

方法1：用言の極性辞書に基づく方法

語彙ネットワークの利用、共起情報の利用、周辺文脈情報の利用など

方法2：機械学習に基づく方法

記事番号	単語系列
1	円　上昇　圧力　強まる　・・・
2	弱　含み　実体　経済　見極め　・・・
3	住宅　生産　指標　にらむ　やや　円安　展開　・・・
4	穏やか　円高　基調　決算　受け　・・・
:	:

株価が上昇した時の記事

記事番号	単語系列
1	金融　サミット　焦点　・・・
2	8000円　挟み　一進一退　展開　・・・
3	円　上値　試す　展開　・・・
4	金融　機関　巡る　動き　焦点　・・・
:	:

株価が下落した時の記事

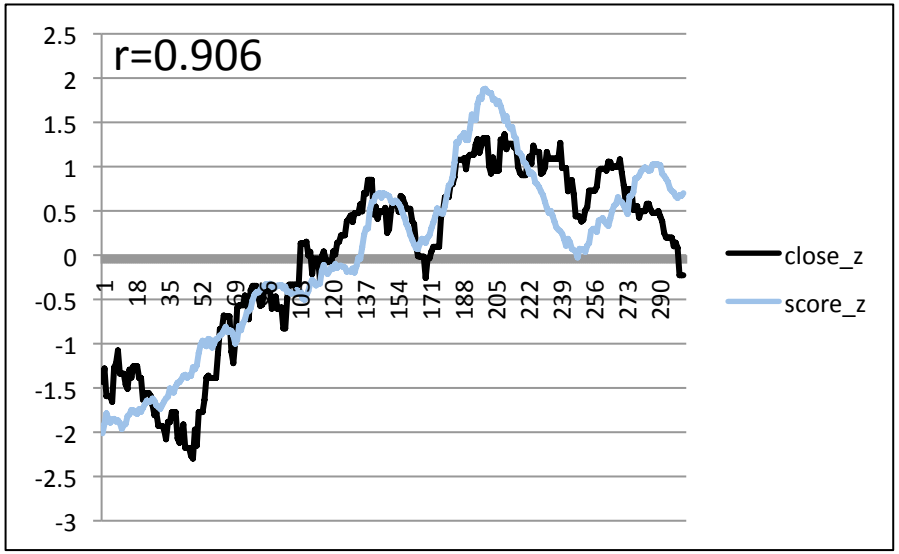
EPに基づくモデル構築

選ばれた顕著なパターンから株価を予測するモデルを構築する。

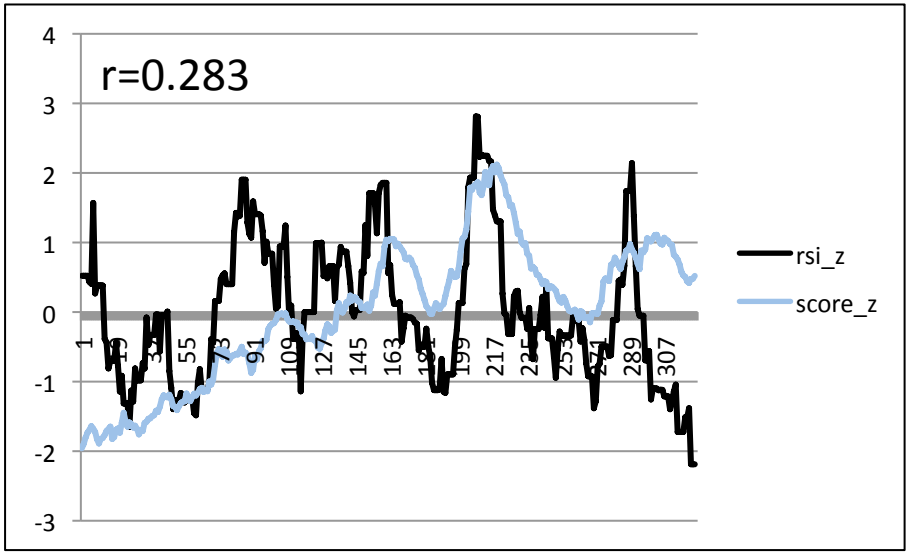
日付	パターン1	パターン2	・・・	TOPIX
1/1	1	1		9043.12
1/2	0	1		9080.84
1/3	0	0		9239.24
1/4	1	1		8876.42
:	:	:		:

・記事を文節もしくは形態素の系列として扱う
・1つの記事の平均文節数107、平均形態素数315 (過去一年の日経新聞による)
・Emerging Patternsの列挙が必要
・文節の位置情報の扱い(Window幅、最大ギャップ、最小ギャップ)
・アイテム数が非常に多い

参考：日経平均とセンチメントスコアの相関



2009/1/21～2009/11/29 日経新聞朝刊



2008/12/31～2009/11/29 日経新聞夕刊

周辺文脈情報の利用など

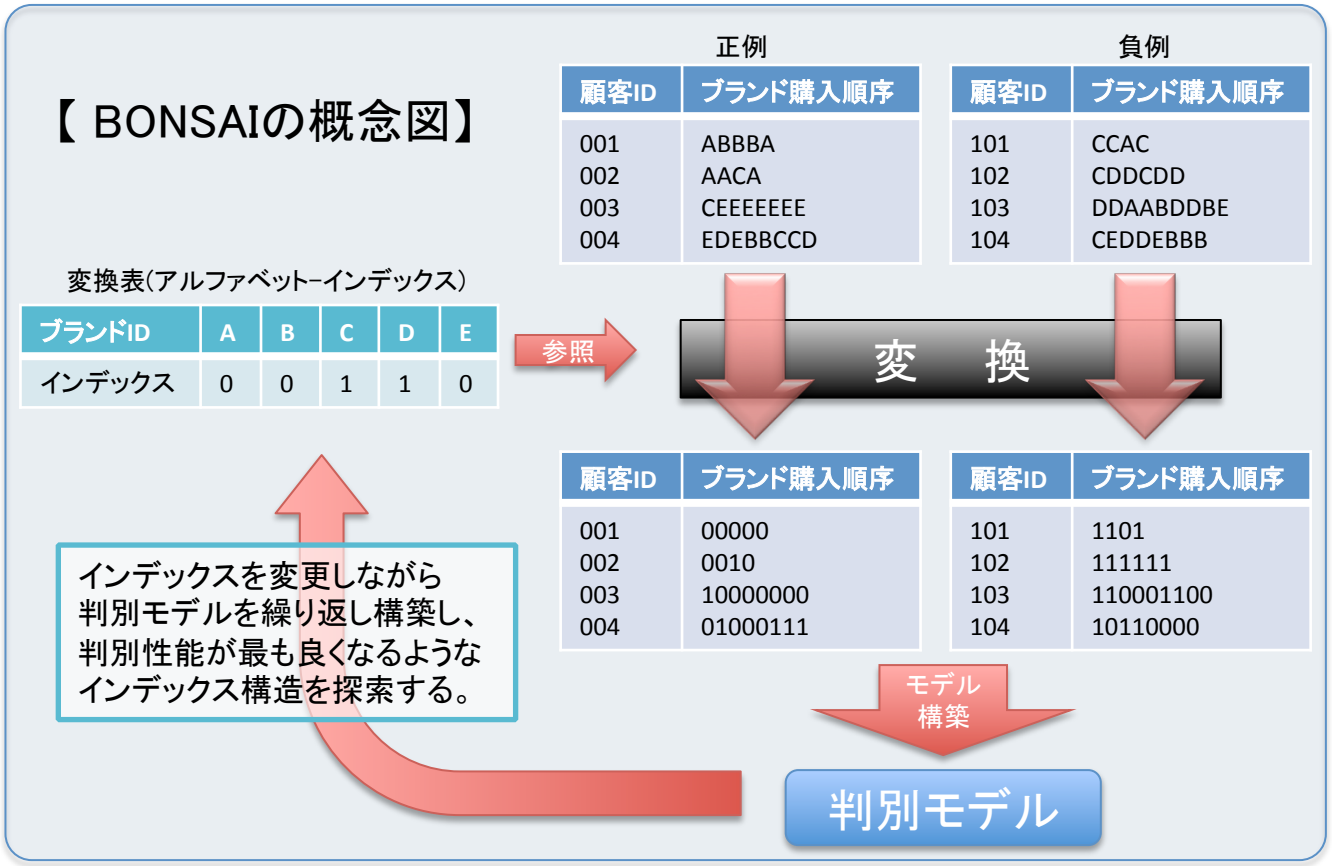
- 極性を持つ種語を分析者が指定する。Ex. 回復する←→悪化する
- 種語の周辺文脈の用言は、種語と同じ極性で使われていると考える。
- 否定語、逆接接続詞を考慮して極性を決定する。
- 新たに登録された極性付き用言を種語として繰り返し処理する。

肯定極性	否定極性
好転する、上方修正する、回復する、回復、底堅い、下げ止まる、利益確定売る、改善、続伸、続伸する、改善する、好調	悪化、悪化する、下方修正する、下落する

パターン列挙に基づくビジネス応用例　その3）

高次元アンケートデータ

- メディア接触、消費スタイル、商品購入実態を中心とした10,000次元を超えるアンケートデータの解析。
- 異なる二つのブランドを選好する顧客のEPを列挙し分類モデルを構築
- あまりに次元数が高いので、あらかじめ次元縮約をおこなう。
- BONSAIの考え方を応用



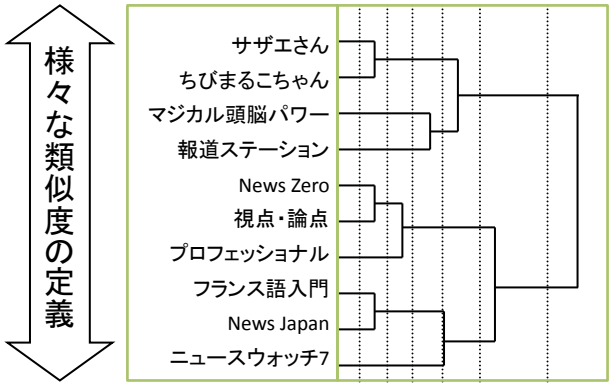
BONSAIの課題

アルファベットのサイズとインデックスのサイズによっては、実時間内に求解できない。

Taxonomyの自動構築

① クラスに関係なく全データからアイテム間の類似度を求め階層型クラスタリングを実行する。

階層型クラスタリング



② 類似度の定義とカットするレベルを固定しクラスタ数を決定する。

クラスタA	クラスタB
サザエさん ちびまるこちゃん	マジカル頭脳パワー 報道ステーション
クラスタ C	**クラスタ D**
News Zero 視点・論点 プロフェッショナル	フランス語入門 News Japan ニュースウォッチ

参照

クラス1

Cyber-shot

サンプルID	質問&回答 アイテム
001	サザエさん、ちびまるこちゃん、News Japan
002	視点・論点、プロフェッショナル
003	マジカル頭脳パワー、ニュースウォッチ、サザエさん
004	News Zero

クラス2

FinePix

サンプルID	質問&回答 アイテム
101	News Zero、フランス語入門、ニュースウォッチ
102	サザエさん、視点・論点
103	News Zero、フランス語入門
104	News Zero、マジカル頭脳パワー

変　換

クラス1

サンプルID	クラスタ化アイテム
001	A D
002	C
003	B D A
004	C

クラス2

サンプルID	クラスタ化アイテム
101	C D
102	A C
103	C D
104	C B

類似度の定義とカットするレベルを変更しながら、判別モデル(CAEP)を**繰り返し構築**し、判別性能が最も良くなるようなクラスタを探索する。

CAEPによる精度推定

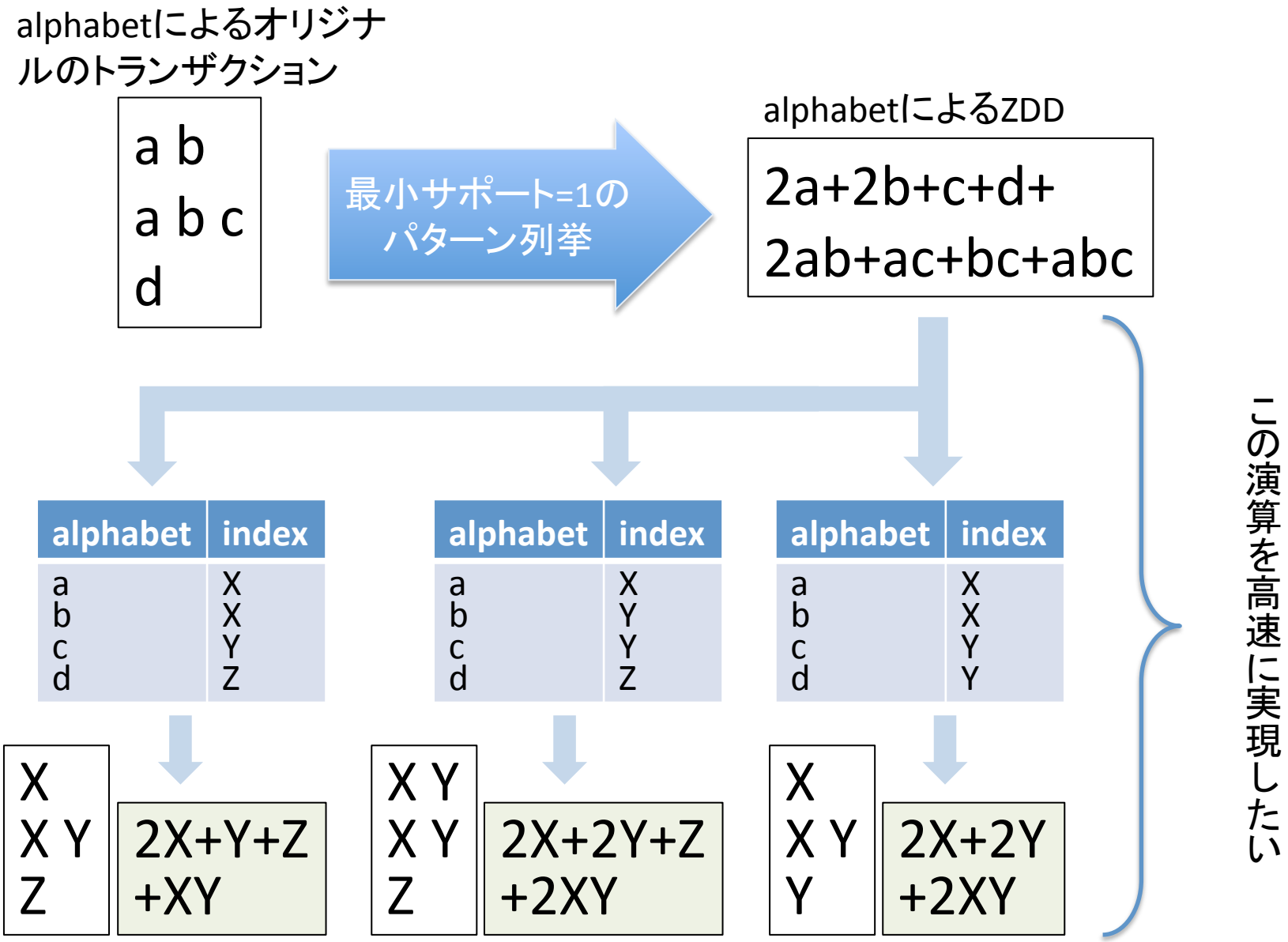
EPの列挙

Cyber-shot EP

Fine Pix EP

統合モデル：各変数群毎に得られた概念（具体的には複数のアンケート項目＋回答で構成されるクラスタ）をアイテムとしてとらえ、さらにデモグラフィック属性をもアイテムに加え、CAEPを用いてブランド選好モデルを構築する。

Taxonomyの自動構築におけるZDDの利用



ERATOで予定しているビジネス応用

- スポーツ用品問屋のWeb受発注データ解析
- 登山用品販売データの解析
- レセプトデータに基づく傷病や投薬の時系列解析：副作用のシグナル検知
- スーパーマーケットにおける購入品の時系列分析
- 企業情報の時系列データ
- 処理履歴の効率的蓄積と有効活用

まとめと課題

- 以下のような拡張パターンを高速に列挙できればビジネスで応用できる問題は多数ある。
  - アイテム集合
  - アイテム集合シーケンス
  - 時間制約：minGap, maxGap, winSize
  - Taxonomy
  - Emerging Patterns

- パターン列挙の大きな欠点：大量のパターン
  →可読性が悪い
  →Taxonomyの自動構築におけるZDDの利用