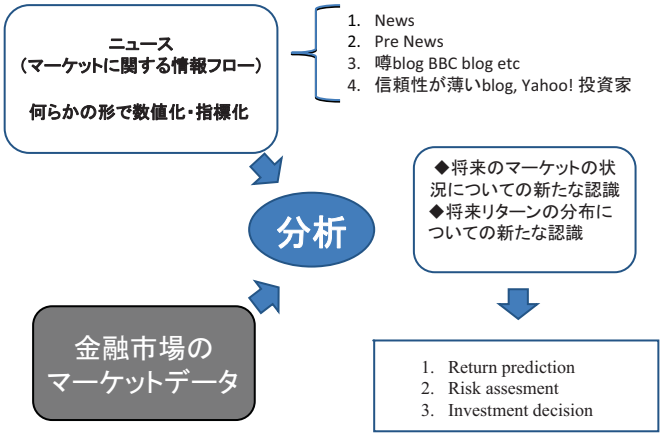
格フレームの概念化手法と株価予測への応用について

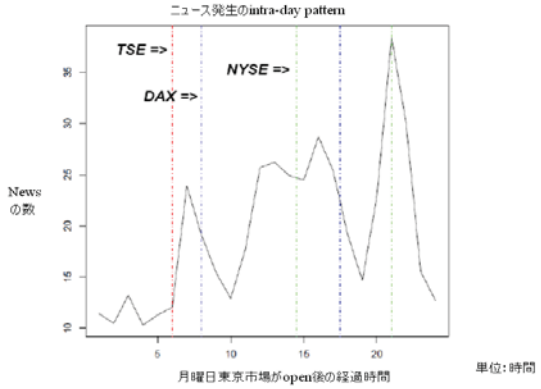
2010年11月30日
(火)
湊ERATO合宿

羽室行信、岡田克彦

高まる期待:
テキストマイニングの株価予想への応用



過去の研究から、株式市場は肯定的なニュースよりも否定的なニュースに強く反応することが分かっている. またマーケットに流れるニュースの量は、positive : negative = 2:1



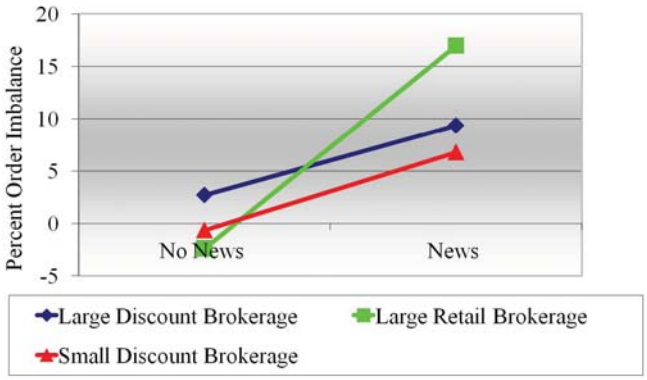
質的なテキスト情報をどのように数値化するか？

- 辞書を使ってセンチメントを指数化(米国ではハーバード辞書を使用)
  しかし、例えば『減配』は辞書では負とみなされるが、金融ニュースにおいては、成長機会の多い企業では新規投資の機会が多いという正の意味を持つ場合がある。
- 投資家が読んでどう感じるかという部分を数値化するため、米国では金融の専門家がセンチメント分類し、数値化し、それを提供している会社も存在する。
  http://www.ravenpack.com/index.html
- Marketの反応をみて、センチメントを評価する方法。
  MITのAndrew LoはEvent studyの手法でニュースの反応を見極め分類。

行動ファイナンスから得られた知見の応用

No News vs. News

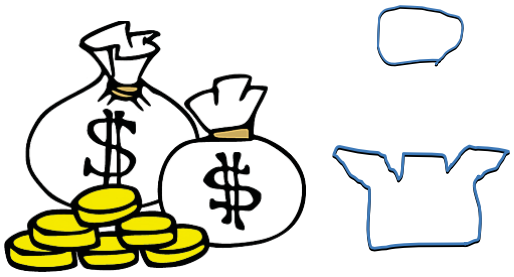
Order imbalance by number of trades



- Newsとmarketの関係についてはとても複雑で、切り口は無数に存在する。
- News releaseについて投資家がどう反応するかを行動ファイナンス、神経経済学の枠組みで捉え、投資家のシステマティックな反応を利用して裁定利益を得ようとするファンドも存在する。
- http://www.marketpsy.com/index.php

目的

- 実運用に耐えうる株価予測モデルを構築する。
- 構築したモデルに基づいた投資ファンドを実際に運用する。



変数

目的変数

収益率 × { 期の単位(数時間～数ヶ月) / 個別銘柄 or NIKKEI225 }

$$r_t = \frac{p_t}{p_{t-1}} - 1 \qquad r_t: \text{期 } t \text{ の収益率},\ p_t: \text{期 } t \text{ の株価終値}$$

説明変数

- ファンダメンタル変数
  - ・株価終値
  - ・Volume(取引量)
- テキストマイニング変数
  - ・センチメント指数
    - a) 極性付き概念格フレーム
    - b) 顕在概念格フレーム
  - ・アナリスト評価
    - 評価のばらつき
    - 格付けの変更
    - 楽観度、悲観度
  - ・新規トピック
- その他anomality変数
  - ・月曜日ダミー
  - ・一月ダミー

× { 時系列 / 分野別 / 個別銘柄 or NIKKEI225 }

テキストデータの整備

日経新聞(→全記事のダウンロードができなくなっている)
Bloomberg(→現在、操作自動化ソフトにより取得中)

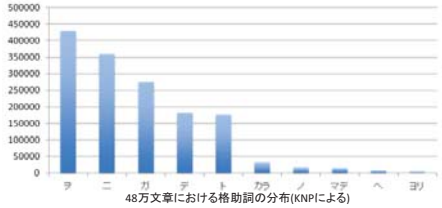
各種辞書
- ・EDR概念辞書
- ・京大格フレーム辞書
- ・類語.jpシソーラス(DB化済み)
- ・日本語大シソーラス(DB化済み)
- ・日本語語彙体系(購入予定)

格フレームとは(復習)

格フレーム(case frame)
1968年に言語学者チャールズ・フィルモアによって提唱された格文法理論。
用言句(動詞を基準として、取り得る格とその値に関する制約を記述したもの。

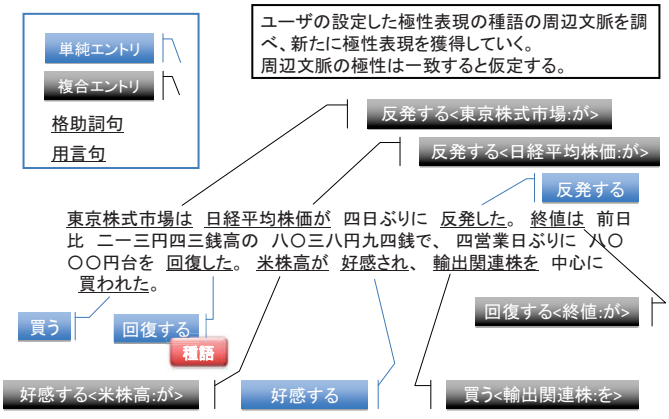
格の種類
表層格: ガ、ヲ、ニ、カラ、ヘ、ト、ヨリ、マデ、デ
深層格: 動作主、経験者格、道具格、対象格、源泉格、目標格、場所格、時間格

ex.
「買う」は、動作主と対象格が必須で、場所格と時間格をとることができる。
私は、昨日スーパーで納豆を買った。



48万文章における格助詞の分布(KNPによる)

周辺文脈法(復習)

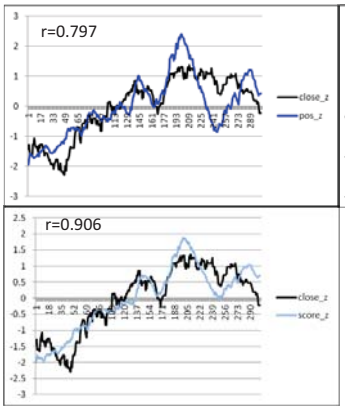


得られたエントリー覧(復習)

		肯定極性 (64エントリ)	否定極性 (27エントリ)
妥当なエントリ	SE	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 (12エントリ)	悪化、悪化する、下方修正する (3エントリ)
	CE	持ち直す<効果:で>, 縮小<減少幅:が>, 伸びる<売り上げ:が>, 取り戻す<金融市場:が>, 付ける<高値:を>, なる<改善:と>, なる<終値:が>, なる<上昇:が>, いう<改善:と>, 上回る<市場予想:を>, 更新する<日経平均株価:が>, 上回る<前年実績:を>, 超える<上昇率:が>, ある<値ごろ感:の>, なる<黒字:に>, 縮小する<赤字幅:が>, なる<プラス成長:と>, 上昇する<前月:より>, 上昇する<指数:が>, 上回る<前月水準:を>, 更新する<高値:を>, 減る<輸入:が>, 反発する<日経平均株価:が>, 上昇する<鉱工業生産指数:が>, 反発する<東京株式市場:で>, 好感する<こと:を>, 転換する<営業黒字:に>, が上昇する<株価:が>, なる<増加:と>, なる<黒字:と>, 進む<在庫調整:>, 更新する<年初来高値:を>, 買う<中心:に>, 取り戻す<市場:が>, 縮小する<赤字:が>, 進展する<在庫調整:が> (36エントリ)	広がる<金融危機:が>, 落ち込む<景気:が>, する<悪化:と>, 広がる<雇用調整:が>, 冷え込む<個人消費:が>, 伸び悩む<需要:が>, なる<上場廃止:と>, 落ち込む<需要:が>, なる<懸念:が>, 直撃する<金融危機:が>, 差す<水:を>, なる<失業率:が> (11エントリ)
妥当でないエントリ	SE	曇る (1エントリ)	蒙る (1エントリ)
	CE	なる<支え:と>, かかる<悪化:に>, 発表する<1日:に>, 示す<消費者心理:を>, よる<景気調査:に>, 大きい<面:が>, なる<販売台数:が>, 除く<金融機関:を>, 差し引く<割合:を>, ある<株価:が>, 低い<格付け:が>, 終える<上海総合指数:が>, 示す<実感:を>, 示す<景気動向:を>, なる<営業損益:が>, 緩やか<テンポ:が> (15エントリ)	よる<雇用統計:に>, 発表する<米労働省:が>, する<帳消し:に>, 算出する<ため:に>, ある<高水準:に>, 探る<将来動向:を>, 続く<雇用情勢:が>, 発表する<上昇:と>, 発表する<欧州連合統計局:が>, ある<今週:が>, 高い<利益率:の>, 広がる<正社員:に> (12エントリ)

* α =0.1, minSupp_{cs}=10にて作成された辞書を利用
* 3回のiterationで得られた内容を掲載(4回目のiterationで収束)
* 下線は種語

株価推移とセンチメント指数の相関(復習)



2009/1/21〜2009/11/29
α=0.01
$minSupp_{ce}$=10
移動平均=WMA
terms=50

しかし、収益率予測に大して
はほとんど説明力がない。

認識している周辺文脈法の問題点

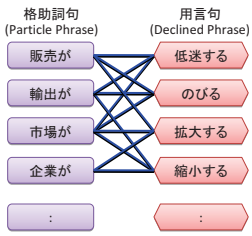
- 極性値が0-1値で定義され、中間を表現できない。
- 単純エントリでは意味が広すぎる。
  ex. 上昇する (株価が上昇する vs. 金利が上昇する)
- パラメータに敏感(収束／発散)。
- 収束する場合は大抵、得られた評価表現数が少ない。
- 目的の極性(株価収益率)と記事の極性が必ずしも一致しているとは限らない。

特定の分野における記事やブログであれば、極性の軸がぶれる事は少ない。
→周辺文脈法の前提。
新聞記事には、様々な極性の軸が存在する。

そこで、二つに分けて考える。
・株価の収益率に影響を与えそうな格フレームをできる限り正確に取得する。
→ 極性付き概念格フレーム、顕在概念格フレーム
・より一般的な極性(良い/悪い、きれい/きたない)を扱う。経済ニュースだけでなく、スポーツや芸能も含める。

2(n)部グラフとして格フレーム

格助詞句と用言句を2部グラフとして表現



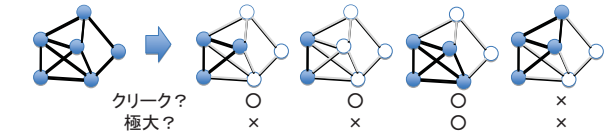
枝が密な部分グラフに注目する。

今回の実験での点数と辺数
辺数 :175,165
用言句数 :23,095
格助詞句数(ガ格) :62,334
枝密度 :0.00012

2部グラフの極大クリーク

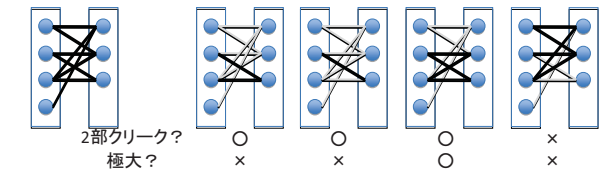
一般グラフの場合

クリーク: 任意の2点間に辺が存在するような部分グラフ
極大クリーク: クリークの点集合が他のどのクリークの点集合にも真に含まれないようなクリーク



2部グラフの場合

2部クリーク: 同じ点集合の点ど同士に辺はなくてよい。
極大2部クリーク: 2部クリークの点集合が他のどの2部クリークの点集合にも真に含まれないような2部クリーク



疑似クリークの定義(宇野 2007)

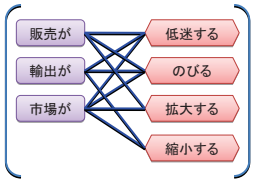
一般グラフの場合 $G = (V, E)$

$$\text{枝密度} p(G) = \frac{|E|}{|V| \times (|V| - 1)/2}$$

2部グラフの場合 $G = (V_1 \cup V_2, E)$

$$\text{枝密度} p_b(G) = \frac{|E|}{|V_1| \times |V_2|}$$

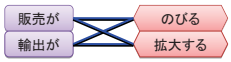
疑似クリーク列挙問題:最小枝密度θ以上の枝密度を持つ疑似クリークを全列挙する。



枝密度: 10 / (3 × 4) = 0.833

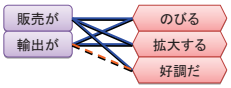
極大2部クリーク、極大2部疑似クリークを用いる意味は？

- クリークを列挙する事により、似た用法の格フレームを集めることができ、種語を一気に増やせることが期待できる。
  → 概念格フレーム。



種語に「販売が のびる」があれば、その他の
格助詞句と用言句の全組合わせも強制的に種語に組み入れる。

- 疑似クリークを列挙する事により、コーパスに出現しない格フレームを補完する事が期待できる
  → 格フレーム補完。



「輸出が 好調だ」の格フレームがなかったとしても、あるものとして扱う。

クリーク関係列挙ソフトウェア (宇野先生)

	クリーク列挙	疑似クリーク列挙
一般グラフ	MACE (MAximal Clique Enumerater)	PCE (Pseudo Clique Enumerater)
2部グラフ	LCM	AFIM (Ambigious Frequent Itemset Minor)

出所) http://research.nii.ac.jp/~uno/codes-j.htm

重複問題

- クリークであっても、疑似クリークであっても、同じような(疑似)クリークが多数列挙されてしまう(多くの点が重複したクリークが列挙される)。

例)

期待_が 反発_が 動き_が 見方_が 懸念_が 声_が 不満_が 影響_が 危機感_が 批判_が	あう 広がる 強い

期待_が 反発_が 動き_が 見方_が 懸念_が 声_が 不満_が 影響_が 議論_が 批判_が	あう 広がる 出る

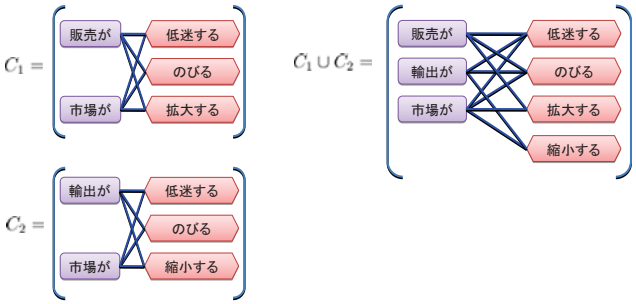
クラスタリング

極大2部クリーク$C_1, C_2, ..., C_m$をボトムアップでクラスタリングする。

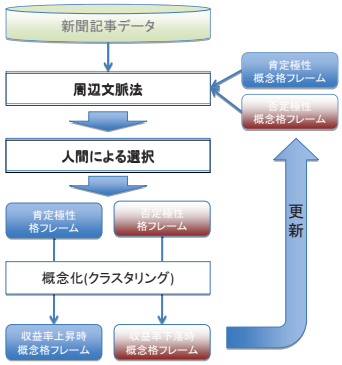
$sim(C_i, C_j) =$ 枝密度$p_b(C_i \cup C_j)$

$sim(C_i, C_j) \geq \theta$となるようなクラスタがなくなれば終了する

得られたクラスタ(極大2部疑似クリーク)を概念格フレームと呼ぶことにする。



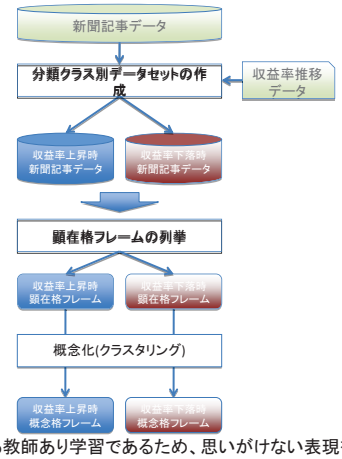
手法の概略1: 極性付き概念格フレーム



初期の種語として数百の格フレームを用意する。

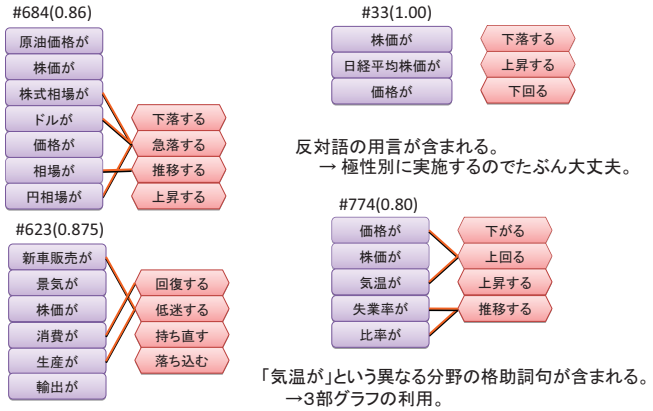
☞ 改良点のポイント:
- 概念格フレームを導入する事で、より多くの評価表現を得ることが期待できる。
- 人間による選択を組み込む事で妥当でない極性表現を省くことで、発散を抑える。

手法の概略2: 顕在概念格フレーム



☞ 特徴:収益率による教師あり学習であるため、思いがけない表現を得る事ができる。

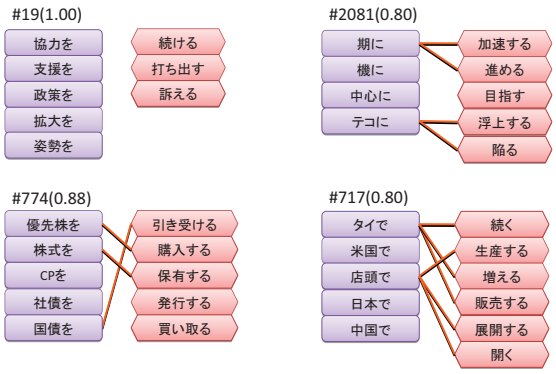
2部疑似クリーク(ガ格)



反対語の用言が含まれる。
→ 極性別に実施するのでたぶん大丈夫。

「気温が」という異なる分野の格助詞句が含まれる。
→3部グラフの利用。
→シソーラスの利用。

2部疑似クリーク(ヲ/ニ/デ格)



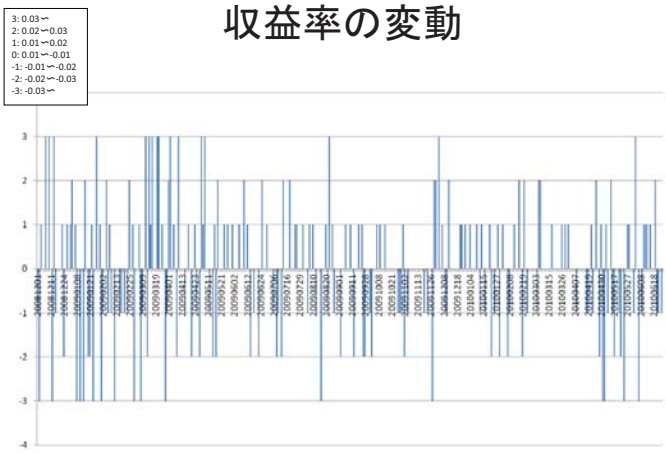
顕在格フレームの抽出

1つの用言とそれに係る(複数の)格助詞をトランザクションの単位とした顕在パターンの抽出

sid	sid	日付	用言句	格助詞句	格助詞	クラス
350	9	27	20081202	回帰する	先進国	に	up
350	9	27	20081202	回帰する	製造業	が	up
343	9	10	20081202	回復する	機能	を	up
384	13	2	20081202	固める	イエスマン	で	up
384	13	2	20081202	固める	周囲	を	up
365	6	6	20081202	固める	方針	を	up
376	0	6	20081202	国際協調志向	「現実路線」	で	up
376	0	6	20081202	国際協調志向	外交	が	up
569	3	4	20081203	参加する	全党	が	down
567	8	2	20081203	参加する	十六人	が	down
588	9	11	20081203	参加できる	代表者	が	down
588	9	11	20081203	参加できる	組織運営	に	down
581	28	15	20081203	及ぶ	わが国	に	down
581	28	15	20081203	及ぶ	高まり	が	down

TID	ITEMS	CLASS
350_9_27	回帰する 先進国_に 製造業_が	up
343_9_10	回復する 機能_を	up
384_13_2	固める イエスマン_で 周囲_を	up
365_6_6	固める 方針_を	up
376_0_6	国際協調志向「現実路線」_で 外交_が	up
569_3_4	参加する 全党_が	down
567_8_2	参加する 十六人_が	down
588_9_11	参加できる 代表者_が 組織運営_に	down
581_28_15	及ぶ わが国_に 高まり_が	down

収益率の変動



変動のより激しいイベントのみを予測した方が当たりやすいかも。
ということで、クラス定義を変えて実験してみました。

実験結果要約

クラス	min supp. (%)	増加率	格フレーム数 (up)	格フレーム数 (down)
1	0.006	4.0	151	177
		1.5	5	16
2	0.01	4.0	18	16
		1.5	82	99
3	0.02	4.0	3	9
		1.5	33	46

DOWN	件数	UP	件数
打撃を 与える	8,0	弾力性を 高める	7,0
ため息を つく	9,0	最小限に 抑える	7,1
圧力を かける	12,0	販売が 好調だ	10,2
土地購入を めぐる	11,0	不安が 広がる	9,2
犠牲に なる	8,1	軌道に 乗る	8,2
金融危機が 深刻だ	11,2	期待が ある	8,2
消費税を 含む	11,2	全力を 挙げる	19,6

クラス=2, 増加率=4.0において列挙された顕在格フレーム

課題

- シソーラスの導入
- 枝重みの考慮
- 3部グラフ(〇〇が△△を××した。)
- 分散処理への対応