

月次報告書と基準価格を用いた投資信託商品の要因分析

大西 恒彰[†] 馬 強[†]

[†] 京都大学情報学研究科社会情報学専攻

〒 606-8501 京都市左京区吉田本町

E-mail: [†]onishi@db.soc.i.kyoto-u.ac.jp, ^{††}qiang@i.kyoto-u.ac.jp

あらまし NISA 制度が導入され、貯蓄から投資への誘導が進んでいる。中長期的な商品である投資信託は NISA に適しているが、似たような商品が多く、それぞれの商品に影響を及ぼす要因を理解するのが難しい。そこで、本稿では、投資信託の理解支援のため、月次報告書と基準価格の変動を用い、投資信託商品の要因分析手法を提案する。提案手法では、まず、CRF により月次報告書から要因事象、結果事象、時間情報を抽出する。同時に基準価格や日経平均価格の変動幅を利用して要因の影響の強度を計算する。さらに、時間、地域と商品の三つの側面から要因の及ぼす範囲を分析し、要因の影響の広度を計る。提案手法を、過去 1 年分の月次報告書を用いて評価をする。

キーワード 投資信託 理解支援 因果関係 要因分析

1. はじめに

貯蓄から投資へというスローガンがある。国は直接投資を増やす策として「証券優遇税制（株の儲けや配当などの税率を引き下げる）」を長期にわたって続け、2014 年 1 月に「NISA（少額投資非課税制度）」を始めた^(注1)。しかし、株や債券への投資は、「多くの銘柄の中から選ぶのは大変」、「株式を買った企業が倒産してしまったらどうしよう・・・」などと不安を持つ人が多く、現状リスクマネーの流入は限定的である。こうした現状で、投資信託の需要は飛躍的に高まりつつある^(注2)。しかし、2014 年の投資信託に関するアンケート調査報告書^(注3)によると投資信託を購入しない理由として、「投資信託がよくわからないので不安」が 63.8% と高く、保有未経験者にとって投資信託は不安と考えられていることもわかる。これは、投資信託といっても、「国内外の様々な債券に投資して、銀行預金よりも良い利回りで比較的风险の低い商品」や「新興国株式に投資して 1 年間で 2 倍以上に値上がりした実績をもつハイリターンな商品」など、運用方法の違った投資信託が沢山あり、選択が難しく、値動きの理由が分かりづらいことにある。

本研究は上記の問題を解決すべく投資信託の理解支援のために要因分析を行う。我々は経済事象で頻出する「急落」や「前半」などテキストデータを数字データを用いて定量化する。さらに、事象の殆どが商品の価格の上昇・下落といった結果事象の要因事象であることから経済事象の要因分析を行う。要因分析では基準価格や日経平均価格の変動幅から要因の影響の強度を計算する。同時に、時間、地域と商品の三つの側面から要因の及ぼす範囲を分析し、要因の影響の広度を計る。

本研究の主な貢献は以下にまとめる。

- テキストと数字データを併用して経済事象を分析する。経済事象の記述における「急落」や「前半」など抽象的な表現を数字データと対応付けて、経済事象の定量分析に用いる。
- 金融商品の報告書の特徴を利用して、経済事象の因果関係の分析を行う。月次報告書においては、事象がたくさん記述されているが、その殆どはその商品の価格の上昇・下落といった結果事象の要因事象であることを利用するとともに、手がかり表現を用いた因果関係抽出手法を併用して、経済事象の要因分析を行う。
- 要因の影響について強度と広度の二つの側面から分析を行う。強度はその事象による金融商品への影響の強さであり、価格の変動幅に基づいて計算される。広度はその事象の影響を及ぼす範囲を、時間・空間と商品数から分析する。

本稿の構成は次の通りである。2 節で関連研究について述べ、3 節で要因分析の手法について説明する。4 節では実験を行い、5 節でまとめと今後の課題について述べている。

2. 関連研究

2.1 数字データを用いた因果関係分析

数字データに基づいて抽出される因果関係はグレンジャー因果性 [1] と呼ばれ、経済のドメインにおいて有用とされる。小野ら [8] は、文書に明示的に書かれていない因果関係をイベントの時系列分析により獲得をする手法を提案している。提案手法では、バーストの時系列データ間に統計的に因果性があると判断されたイベント対を因果関係知識として抽出する。バーストとは、あるデータが急激に増加する現象のことをさす。統計的な因果性の判断にはグレンジャー因果性検定を利用している。上記の研究はテキストデータからは得られない因果関係も取得できている。我々は、因果関係抽出ではなく要因分析に数字データを利用している。

2.2 テキストデータを用いた因果関係分析

坂地ら [2] は、企業が Web ページに掲載する決算短信文章

(注1): 日本電子計算株式会社 更なる個人の投資拡大へ 新たな税制改正とその背景 <https://www.jip.co.jp/report/detail.php?report=00175>

(注2): NISA に適した金融商品は? 投信市場のトレンドに変化 <http://www.morningstar.co.jp/nisa/product/fund20.html>

(注3): 投資信託に関するアンケート調査報告書-2014 年(平成 26 年) <https://www.toushin.or.jp/statistics/report/research2014/>

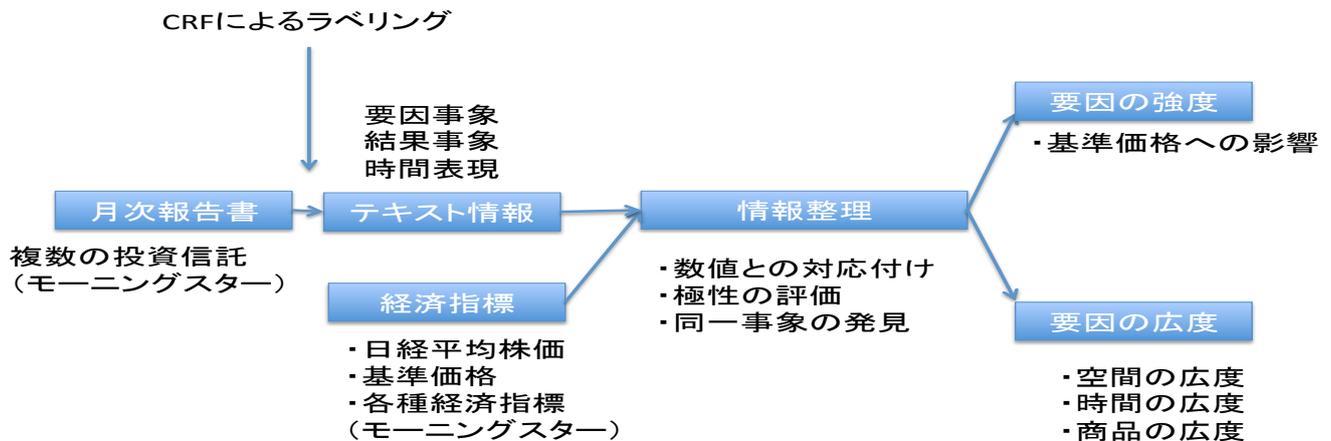


図 1 要因分析の手順

に着目し、決算短文文章 から原因・結果を抽出する。彼らは、素性に言語オントロジーや構文情報を用いた機械学習手法により、決算短文文章から原因・結果を含む文を抽出する。さらに、抽出した原因・結果を含む文から原因・結果表現を抽出している。原因・結果表現抽出には、構文情報を用いた Pattern、更に原因・結果表現を抽出するために用いる手がかりとなる表現をブートストラップ法を用いている。

粟納ら [3] は投資信託商品の数があまりにも多く、また投資信託商品はそれ自身が複雑で理解が困難であることに着目し、月次報告書のような非構造化テキストから事象や因果関係を抽出し、理解支援のために視覚化する手法を提案する。さらに提案手法を投資信託商品の理解支援のアプリケーションに適用しユーザ評価を行っている。

青野ら [7] は、手がかりとなる表現を基に、Web 文書から要因を検索、抽出し、さらに因果関係ネットワークを構築する。彼らは要因検索システムとして、因果関係の要因とその結果をそれぞれ始点ノードと終点ノードに配置することで因果関係ネットワークを構築する。因果関係ネットワークにより、事象の類似性や発生順を推測したり、ネットワークを縮約したりできる可能性を示した。

我々は上記の研究と同じくテキストデータから因果関係を抽出する。また坂地、粟納らと同様に金融商品の報告書の特徴を利用して、経済事象の因果関係の分析を行う。我々は要因の影響についてテキストデータと基準価格より強度と広度の二つの側面から分析を行うことにより詳細な情報を提示する点が既存研究と異なる。

2.3 極性評価

酒井ら [5] は投資家にとって、企業の業績に関する情報を収集することは重要であると考え、業績を決算短文文章より抽出

する手法を提案している。さらに、酒井らの手法では手がかり表現を用いて業績要因に対して業績に対する極性を付与している。例えば、業績要因「半導体製造装置の受注が好調でした。」に対しては「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。」に対しては「ネガティブ」の極性を付与する。

東山ら [6] は述語の選択選好性に着目した名詞評価極性の獲得より事態の評価極性を判別している。彼らは事態を表す「名詞+ 格助詞+述語」の 3 つ組について Web コーパス上で頻出する表現を収集し、名詞の評価極性の重要性を調べ、約 4 割の評価極性を持つ組が名詞の評価極性に依存し、名詞が事態の評価極性判定には重大な役割を持つことを確認した。例えば「を好き」はポジティブな名詞と共起しやすく、「が嫌い」はネガティブな名詞と共起しやすいといった評価極性についての述語の選択選好性を利用することにより、5 億文の大規模な Web コーパスから「名詞」と「格助詞+述語」の共起関係を抽出し、機械学習手法によって、名詞の評価極性の知識を自動獲得する手法を提案している。

我々はこれらの研究と異なり、極性評価を基準価格や日経平均株価などの関連する数字データを用いることで行っている。

3. 要因分析

3.1 概要

この節では、我々の要因分析手法の全体構成を説明する。図 1 に我々の要因分析手順を示した。まず、投資信託月次報告書、基準価格等はモーニングスターから取得する^(注4)。まず CRF[9]

(注4): モーニングスターはモーニングスター株式会社により運営される株式・投資信託・ETF の総合金融情報サイトである。投資信託であれば、国内外の投資信託 4403 本の情報を利用できる。さらにモーニングスターではリスクリターン、

表 1 抽出される具体例

該当月	時間情報	要因事象	結果事象
6月	前半	黒田日銀総裁の国会答弁を材料に円安の進行に不服感が出た、ギリシャ支援協議の先行き不透明感が増してきた	株価は下落しました
	後半	ギリシャが新たに示した財政改革案をユーロ圏側が評価した	株価は上昇する局面もありました
	月末	ギリシャがデフォルト（債務不履行）に陥る可能性が高まった	株価は下落しました
7月	前半	ギリシャのデフォルトとユーロ離脱への懸念が高まった、中国株が急落した	株価は一時大きく下落しました
	中盤	中国株が上昇に転じたギリシャ支援協議が合意に達したこと	株価は反発しました
	後半	一部の米国企業の決算下振れ、中国需要の減退を主因とする業績下方修正などが嫌気され	やや弱含みで推移しました

より月次報告書のテキスト情報から要因事象、結果事象、時間情報を抽出する。次にテキストデータと数字表現の対応付けを行うと同時に要因事象と結果事象の極性評価を行う。さらに文章類似度を用いて同一事象の結合を行うことを考えている。その後、基準価格や日経平均価格の変動幅を利用して要因の影響の強度を計算する。同時に、時間、地域と商品の三つの側面から要因の及ぼす範囲を分析し、要因の影響の広度を計る。

3.2 要因抽出

我々は、投資信託商品の月次報告書の現状報告に関する部分「市況概況や運用経過」と「今後の運用方針」を対象に、要因事象の抽出を行う。月次報告書は構造化されていないため、記述の中のどこが要因事象であるかがわからない。これらのテキストの構文解析を Cabocha[10] を用いて、構文木によって生成されたチャンクを対象に CRF[6] によるラベリングを行う。

ラベリングでは文章を、要因事象、結果事象、時間表現、接続詞、句読点に分類する。ただし、結果事象は株価及び基準価格が上昇したか、下落したかの価格変動の事象であり、要因事象は結果事象の材料となっている事象である。

同時に形態素解析より地域タグを取得する。表 1 にこのラベリングで取得された要因事象、結果事象、時間表現の具体例を示す。例えば「後半は、ギリシャが新たに示した財政改革案をユーロ圏側が評価したことで、支援協議の合意期待が高まり、株価は上昇する局面もありました。」という文章であれば、「ギリシャが新たに示した財政改革案をユーロ圏側が評価したことで、支援協議の合意期待が高まり」が要因事象にあたり、「株価は上昇する局面もありました」が結果事象にあたる。時間表現は「後半」があたる。また地域タグとして「ギリシャ、ユーロ」が取り出される

3.2.1 同一事象の発見

経済事象は同一の事象の表現揺れがみられる。投資信託であれば、同月の別会社から出されている月次報告書においてよく見られる現象である。例えば QE3、量的金融緩和、量的金融緩和政策が同じ経済事象を指す、我々は LDA により投資信託の文章をトピックごとに整理し、トピックごとに出現する単語より文章を分類することを目指し、同一経済事象を指す用語を発見する。

3.3 要因分析

3.3.1 テキストと数字表現の対応付け

「株価は大幅に上昇しました」のような結果事象は具体的にいくらの上昇があったのか、テキスト情報だけでは判断できない。また「前半」のような時間情報も同様に、月の前半が具体的にどの部分に該当するのか判別できない。ここでは、時間情報と実際の基準価格の変動を利用することでこの問題を解決する。以下にその具体例を示す。

ある投資信託における 2015 年 10 月の文章から取り出された時間表現、結果事象が（前半、大幅に下落しました）、（下旬、回復に転じました）、（月末、大幅に上昇しました）であったとすると、

(1) 結果事象の極性評価の手がかり的な形態素列（以降、手がかり表現と定義）より結果事象をポジティブがネガティブが判定する。表 2 に用いた手がかり表現の一例を載せている。この結果（前半、ネガティブ）、（下旬、ポジティブ）、（月末、ポジティブ）となる。

(2) 次に、時間表現の始点日、終点日を定める。始点日とは「前半」のような時間表現に対応する初日であり、終点日とは時間表現の終日を指す。まず基準価格の前日比の極性の転換が結果事象の極性の転換と一致している日、転換日を探す。複数見つかった場合、転換日の基準価格と始点日の基準価格の差が最も大きくなる日を終点日とする。図 2 であれば 2014 年 10 月 17 日が前半日の終点日に該当する。これにより求められた始点日から終点日が時間表現と対応する期間だとみなす。この場合「前半」は 10 月 1 日から 10 月 17 日になる。

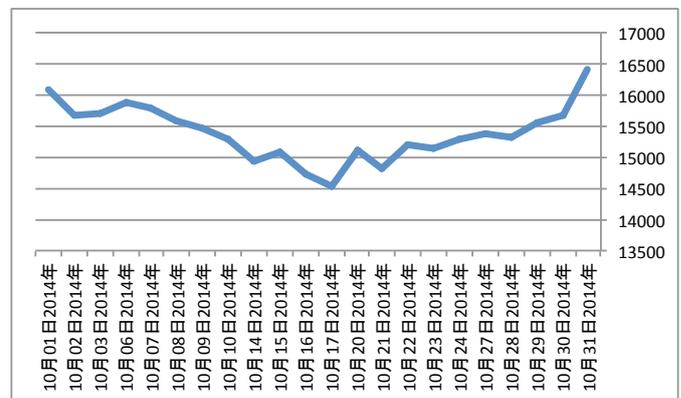


図 2 価格の変動データ

表 2 極性付与に用いた手がかり表現の例

手がかり表現	極性
下落しました	ネガティブ
幾分調整しました	ネガティブ
弱含みで推移しました	ネガティブ
大幅に上昇しました	ポジティブ
回復に転じました	ポジティブ

(3) 中盤から月末にかけては結果事象がポジティブで続く。結果事象の極性が(中盤, ポジティブ)から(後半, ネガティブ)のように反転しない場合は等分する。10月18日から10月31日の営業日は10日であり営業日を二分割し、「中盤」は10月18日から10月24日、「月末」は10月25日から10月31日となる。

(4) この結果{前半(1日から17日), 大幅に下落しました(-9.6%)}(中盤(18日から24日), 回復に転じました(+5.2%)}(月末(25日から31日), 大幅に上昇しました(+7.3%))のような結果事象と時間表現のペアが得られる。

3.3.2 極性

要因の極性はまず、高村ら[4]が公開している単語感情極性対応表を用いて評価をする。しかしながら、経済事象に対応している対応表ではないため、不十分なことがある。結果事象の極性が取得できている場合は、そちらを優先する。例えば、「月末には、日銀による予想外の追加緩和が打ち出されたこともあり、株価は大幅に上昇しました。」という文章であれば、「株価は大幅に上昇しました。」はポジティブに分類される結果事象であるため、この場合、要因である「日銀による予想外の追加緩和が打ち出された」はポジティブと判断されるべきである。単語感情極性対応表を用いて極性の評価を行うとネガティブとされるため、結果事象より反転させてポジティブ要因とする。

また文章において逆接の接続詞が確認できる場合、その前後で要因の極性を反転させる。例えば「支援条件に関するギリシャとユーロ圏の対立などが懸念されましたが米国の1月雇用統計が賃金の上昇を伴う良好な内容だったため株価は上昇した」のような文章が該当する。この場合逆接の「が」により極性の反転をさせる。またこの文章では前者の要因が後者の要因より弱い要因であることが結果事象「株価は上昇した」がポジティブであることから分かる。

3.3.3 要因の強度

要因の強度とは要因の基準価格への影響の度合いのことである。要因の期間における基準価格の変動基準価格の傾きで定義する。

これを求めることにより、どういった要因がそれぞれの投資信託の基準価格に影響の強かった要因であるかを定量化することができる。

$SF(A, i)$ は要因集合 A の商品 i に対する強度であるとする。要因集合 A を観測した始点日 d_s , 終点日 d_e , 商品 i の d 日における基準価格 $p(i, d)$, また始点日から終点日における営業日数 $D(s, e)$ を用いて式1で求めることができる。表3の例であれば、6/10(基準価格)から6/1(基準価格)を引算し、営業日数

表 3 要因集合 A が検出される日程

日時	基準価格	前日比
6/1 (始点日)	20569.87	+0.03
6/2	20543.19	-0.13
...
6/10 (終点日)	20046.36	-0.25

の8日で除算することになる。

$$SF(A, i) = \frac{p(i, d_e) - p(i, d_s)}{D(s, e)} \quad (1)$$

基準価格の上昇もしくは下落が複数の要因により形成されている場合、例えば $A = \{a_1, a_2, a_3\}$ の時、 a_1 の傾きを計算する際は式2のように要因集合 A に含まれる要因数 $|A|$ で除算を行う。

$$SF(a_1, i) = \frac{SF(A, i)}{|A|} \quad (2)$$

異なる月で確認される要因であれば要因強度を指数移動平均の考えを使い、調整する。指数移動平均では重みは指数関数的に減少するので、最近のデータを重視するとともに古いデータを完全には切り捨てないことになる。重みの減少度合いは平滑化係数と呼ばれる0と1との間の値をとる定数式4で決定される。指数移動平均とは、個々のデータに異なる重みをつけて平均を計算するものである。

式3は要因 a の加重平均を求める計算式であり、 $SF(M)$ は要因 a の M 月に計算された強度であり $EMA(M)$ は M 月の指数移動平均である。

$$EMA(M) = EMA(M-1) + \alpha \times (SF(M) - EMA(M-1)) \quad (3)$$

$$\alpha = \frac{2}{N+1} \quad (4)$$

3.4 要因の広度

ここからは要因の時間、地域と商品の三つの側面から要因の及ぼす範囲を分析し、要因の影響の広度を考える。

3.4.1 地域の広度

要因事象から得られる地域タグを用いて、その事象のカバーする地域の広さを分析する。カバーする地域の広さを分析することで特定地域に影響を与える要因か、世界全体に与える要因かを分析する。投資信託から得られる地域タグには以下のようなものから取得する。

- 「大統領が選出できず議会在解散したギリシャの政治情勢」といった要因事象に国名、もしくは地名が含まれる。
- 「日本の市況概況」のように投資信託の文章タイトルで地域が明記されている。

これらの文章を用いて要因は最低でも一つの地域タグを保有することとする。

要因の地域の広度 $LF(a, i)$ はタグを用いてそれに対応する面積もしくは、GDPの指標を用いる。面積データ及びGDPデータはThe United Nations Statistics Division^(注5)より取得

(注5): <http://unstats.un.org/unsd/aboutus.htm>

表 4 収集した投資信託 (2015 年 1 月-12 月)

投資信託商品名	販売会社	カテゴリー
ダイワ好配当日本株投信 (季節点描)	大和証券投資信託委託	国内中型ブレンド
DIAM 新興市場日本株ファンド	DIAM アセットマネジメント	国内小型グローブ
J オープン (店頭小型株)	三菱 UFJ 国際投信	国内小型グローブ
MHAM 新興成長株オープン	みずほ投信投資顧問	国内小型グローブ
ひふみ投信	レオスキャピタルワークス	国内小型グローブ
フィデリティ日本株アクティブファンド	フィデリティ投信	国内中型ブレンド
フィデリティ・欧州株・ファンド	フィデリティ投信	国際株式・欧州
フィデリティ・US ハイ・イールド・ファンド	フィデリティ投信	国際債券・ハイイールド債
ダイワ・グローバル REIT・オープン (毎月分配型)	大和証券投資信託委託	国際 REIT・グローバル・除く日本
ラサール・グローバル REIT ファンド (毎月分配型)	日興アセットマネジメント	国際 REIT・グローバル・含む日本
三菱 UFJ グローバル・ボンド・オープン (毎月決算型)	三菱 UFJ 国際投信	国際債券・オセアニア

する要因 a に付加された地域タグに該当する面積を $area(a)$, GDP を $gdp(a)$ とする. このとき世界全体の陸地面積 (WS とする) および世界全体の GDP (WG とする) を元に地域の広度 $LF(a, i)$ は式 5 で定める.

$$LF(a, i) = \begin{cases} \frac{area(a)}{WS} & (\text{面積ベース}) \\ \frac{gdp(a)}{WG} & (GDP \text{ ベース}) \end{cases} \quad (5)$$

3.4.2 時間の広度

投資信託から取得される要因には「パリで同時テロ」というような短期的に観測されるもの、「ギリシャの不透明感が強まり」といった長期的に観測されるもの、「米国雇用統計」といった周期的に観測されるものが存在する. 一時的に現れる要因よりも長期的に確認される要因がその投資信託により影響を与える要因であると言える. そのため, 要因に時間表現を用いて要因の影響を及ぼす時間範囲を計算し, 要因の時間の広度を分析する. 要因による影響期間は投資信託より取得される「前半」のような時間情報を用いる.

要因の時間の広度 $TF(a, i)$ は式 6 で定義するものとする. 商品 i における要因事象 A の検出回数を N とし, $D_d(s, e)$ は検出 d 回目における要因の始点日 s から終点日 e の日数を示す.

$$TF(a, i) = \sum_{d=1}^N D_d(s, e) \quad (6)$$

3.4.3 商品の広度

一般によく見られ多数の投資信託に影響を与えている要因か, それとも, ある投資信託特有の要因であるかにより要因の商品における広度を特定する. 表記ゆれ事象は統合されているとして, 要因事象 a がある投資信託商品の報告書において出現するかしないかを元に商品の広度 $DF(a, i)$ を計算する. $|U|$ は総投資信託数, $|\{I : I \ni A_i\}|$ は要因 A_i を含む投資信託数である. このとき要因の商品の広度 $DF(a, i)$ は式 7 で定義するものとする.

$$DF(a, I) = \frac{|\{d : d \ni t_i\}|}{|U|} \quad (7)$$

4. 実験

4.1 実験概要

本節では, 月次報告書から実際に要因抽出を行い分析を行う.

以下では, その手順についてまとめる.

- データソースを取得する. 投資信託月次報告書はモーニングスターから取得する. このサイトより月次報告書を 1 年分取得する. 同時に該当する期間における基準価格と日経平均株価を取得する. 我々は 2015 年 1 月から 2015 年 12 月の月次報告書を用い実験を行う.

- 要因事象, 結果事象, 時間表現の抽出をする. これは投資信託商品の月次報告書の経済市況を述べた文章を対象に抽出を行う. 取得された文章に対し Cabocha を用いて係り受けを取得し, 得られた素性情報と共に, CRF による要因事象, 結果事象等のラベリング付けを行う. 正解ラベルは人手による作業で与え, 学習データは一年分の投資信託より人手で作成した.

- 基準価格の変動を元に要因の強度を分析し, 地域と時間と商品に関して要因の広度を分析する. 2015 年 12 月の月次報告書から頻出する要因事象群を与え, 被験者 3 名に投資信託ごとに評価をしてもらう. その結果に基づいて提案手法の nDCG の値を算出した.

4.2 データソース

実験にむけて投資信託の月次報告書及び基準価格の収集を行う. 2015 年 1 月から 12 月における投資信託を対象にモーニングスター及びその販売会社で公開されていた投資信託を取得した. 以下表 4 は実際に取得した投資信託である. 過去 1 年分の月次報告書が公開されているものの中から無作為に選んだ. 選ばれた投資信託における市況報告及び運用経過は 200 字のものから 1000 文字以上の投資信託まであり, 国内株式, 国内債券, 国際株式, 国際債券と多種に及んでいるため検出される要因を事象も多種であった.

4.3 ラベリング

月次報告書より要因事象, 結果事象を時間表現と共に抽出するためのラベリングの評価を行う. ラベリングでは月次報告書文章に対し, Cabocha を用いてチャンクを取得し, 素性情報, 文章上の位置情報をもとに, チャンクに対し要因事象, 結果事象, 時間表現, 句読点, 接続詞のラベリング付けを行う. ここではダイワ好配当日本株投信 (季節点描) の 2014 年の月次報告書を学習データとし, 2015 年のダイワ好配当日本株投信及び J オープン及びダイワ・グローバル REIT においてラベリングの精度を確認した. CRF によるラベリングは表 6 のようになっ

表 5 要因の強度実験において用いた投資信託

投資信託名	文章量	文章の特徴
DIAM 新興市場日本株ファンド (A)	中程度	要因数は中程度時間表現は伴う
J オープン (B)	簡潔	抽出される要因が少ないが時間表現を伴う
MHAM 新興成長株オープン (C)	簡潔	抽出される要因は中程度が時間表現を伴わないものが多い
ダイワ・グローバルREIT (D)	多め	抽出される要因は多いが時間表現が伴わないものが多い
ひふみ投信 (E)	多め	抽出される要因は多く時間表現を伴う

表 6 CRF によるラベリング

投資信託名	ダイワ好配当	J オープン	ダイワ・グローバルREIT
要因事象	0.929	0.959	0.557
結果事象	0.949	0.969	0.666
時間表現	0.885	1	0.765

た。要因結果、結果事象、時間表現のラベリングをした際に人手による作業と一致したチャンク数の割合を出している。結果から学習する投資信託と予測する投資信託が一致しているダイワ好配当は正確にラベリングができていることが分かる。また文章の記述が簡潔である J オープンのような月次報告書は良い精度のラベリングが行えた。しかし、記述量が多く、要員数が多いダイワグローバル REIT のような月次報告書では、要因事象と結果事象を取り間違ったり、要因に対応する時間表現が複数とられ、どの時間表現に対応するかわからないといったケースが存在した。

4.4 事象の結合

月次報告書からの抽出された文章は様々なトピックから構成されている。文章中の各単語がそれぞれ背景にトピックを持つことを利用し、LDA を用い、文章中に存在する事象のトピックを取得する。表 4 より取得された文章を LDA を用いてトピック分類した結果は表 7 のようになった。LDA はトピック数を 20 にして行っている。実験結果を見ると 2015 年に頻出した要因を表す単語の抽出をすることができているため、この単語を使い要因の分類と結合を行うことができると考えられる。しかし、言及の少ない「ウクライナの地政学リスク」のような事象をうまく取り扱うことができなかつたため、今回の実験では全て手作業で事象の結合を行った。

4.5 要因分析

表 4 に挙げられた投資信託を用いて要因分析の提案手法を評価するための実験を行う。表 4 の月次報告書に記載される 1 年間の要因数の合計は人手で収集したところ 879 件であった。この要因に対し、人手により同一の要因であると判断した場合は事象の結合を行い実験を行う。実験では、2015 年 12 月の月次

報告書から頻出する要因事象群を与え、被験者 3 名に投資信託ごとに評価をしてもらう。被験者は 5, 4, 3, 2, 1 の 5 段階の評価を用いて、要因の基準価格へ及ぼす影響と要因の出現期間、要因の範囲に加えて投資信託の月次報告書において広く確認される要因かどうかを判断してもらった。その結果に基づいたランキングを理想のランキングとし、提案手法のランキングにおいて nDCG の値を算出している。

4.6 要因の強度

要因の強度としてある要因が基準価格に与える影響を定量化する。ユーザの評価に基づいたランキングと提案手法によるランキングを比較する事で提案手法を評価する。

実験では、表 4 より以下の表 5 の特徴をもつ投資信託を用いた。

横軸 i が評価するランキングの長さ、縦軸が $nDCG@i$ の値を表し、 $nDCG@i$ ($i = 5, 7, 10$) の各値をプロットし、グラフ化したものが図 3 である。以下、説明では、DIAM 新興市場日本株ファンドは A のように A,B,C,D,E で置き換えて説明する。対応は表 5 に示している。結果をみると、 i が大きくなるにつれて nDCG の値が下がっている。これはランキングが長くなればなるほど予測することが難しくなるからであると考えられ

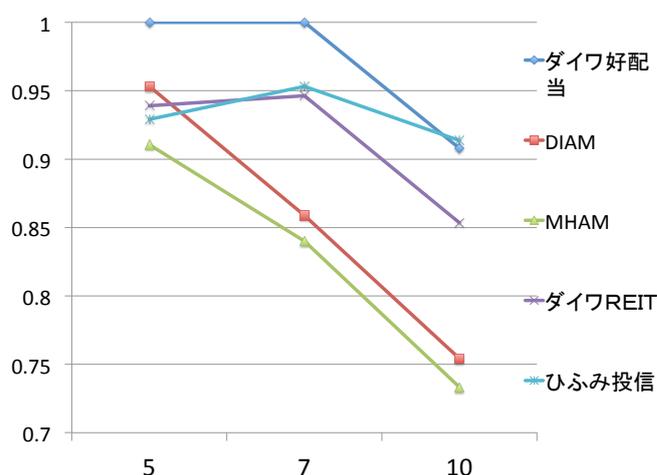


図 3 要因の強度

表 7 月次報告書より取得されたトピック

トピック 1	トピック 2	トピック 3	トピック 4	トピック 5
米国	金融緩和	懸念	企業	安
利上げ	ecb	ギリシャ	良好	上昇
雇用時計	欧州中央銀行	景気	ギリシャ	下落
連邦公開市場委員会	期待	中国	決算	円
懸念	緩和	支援	予想	国債利回り

表 8 要因の強度の nDCG

	nDCG@5	nDCG@7	nDCG@10
DIAM 新興市場日本株ファンド (A)	1	1	0.907
J オープン (B)	0.953	0.858	0.754
MHAM 新興成長株オープン (C)	0.910	0.840	0.733
ダイワ・グローバルREIT (D)	0.938	0.946	0.853
ひふみ投信 (E)	0.929	0.953	0.913

る。A と C もしくは D, E を比較すると A, E の方が nDCG の結果が良い。これは C, D の投資信託において要因事象に時間表現が伴わず、基準価格に影響を及ぼしたと思われる期間を正確に特定できないからであると考えられる。また文章量で A, B, E を比較した時に i が小さいときは文章量と要因数が少ない方が良く、 i が大きくなるにつれて文章量と要員数が多い月次報告書が良いことが確認できる。実験でうまくいった強度の評価は、「月後半に入ると、原油やハイイールド債の下落一服を受けて漸く地合いが改善しました。」のような文章で月後半という時間表現から基準価格の変動期間を取得することができたときであった。しかし「ECB 理事会で決定された緩和策は事前予想よりも消極的な内容であったことから失望を誘い、世界的に株売りが進みました。」のような文章では、期間を特定できず、月全体の基準価格で要因の強度を評価することになり nDCG の値が下がった。また基準価格に大きな変動があるが、要因事象を特定できないケースも存在した。我々の仮定では、基準価格に影響があったと考えられる要因は月次報告書に記載されるものだとしているため、要員の抽出数が極端に少ないような投資信託では他の投資信託に見受けられる要因で補完するといった操作が必要であることがわかった。

4.7 要因の広度

要因の時間、地域と商品の三つの側面が反映できているかをユーザの評価を用いたランキングと提案手法によるランキングを比較する事で評価する。地域の尺度は GDP ベースとした。実験では表 4 の投資信託全てを対象に行っている。

横軸 i が評価するランキングの長さ、縦軸が $nDCG@i$ の値を表し、 $nDCG@i$ ($i = 5, 7, 10$) の各値をプロットし、グラフ化したものが図 4 である。nDCG の結果を見ると時間、商品、空間の順になっていることから、要因の期間の尺度は良く

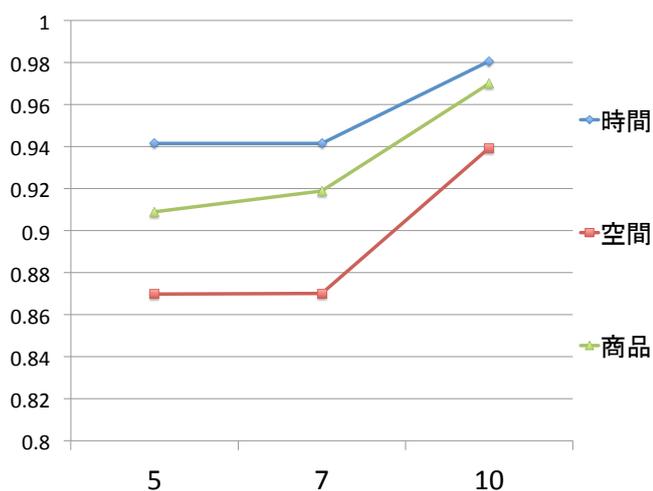


図 4 要因の広度

表 9 広度の nDCG の結果

	nDCG@5	nDCG@7	nDCG@10
時間	0.941	0.9415	0.980
空間	0.941	0.870	0.918
商品	0.980	0.939	0.969

評価できているが、要因がどれだけ広く投資信託商品に確認されるかや要因の空間的影響の評価がうまくいっていない。 i が小さいときに nDCG の値が悪いことからランキング上位の要因がうまく評価できていないことが分かる。時間の nDCG が良い理由としてももとのデータソースが月次報告書ということもあり、最初から月ごとに要因が整理されていることにある。また、商品の nDCG の結果が悪い原因としては、人手により事象を結合してしまった結果、事象の抽象度が上がり、すべての投資信託で確認できる事象になったことが考えられる。また空間の nDCG が悪い原因としては、GDP ベースでは、要因が「米国の雇用統計」と「米国の利上げ」において付加される地域タグが米国になってしまい、差別化できないことにある。これは他の共起する要因の地域タグを用いるなどして改善することが考えられる。

5. 終わりに

本論文では、テキスト情報を基準価格といった数値情報で定量化することができた。さらに月次報告書に書かれた要因事象を元に基準価格や日経平均価格の変動幅を利用して要因の影響の強度を計算した。また事象に対し地域タグを考慮することで要因の空間的影響度、同時期の月次報告書で同じ事象を見つけること要因の商品的広度、要因に付随する時間表現をもとに要因期間とし、空間、商品、時間の尺度で要因の影響の広度を定量化できた。実験結果によると要因事象が時間表現とともに取れる文章であれば、要因の強度を正確に捉えることができていたが、広度に関して商品や空間の実験結果がうまくできていなかった。また投資信託に出現する要因の粒度が統一されておらず事象の結合がうまく行っていないことが判明した。この問題に対し、事象の粒度を捉え適切な事象の粒度で結合を行う必要がある。また、要因事象の周期性を考察し基準価格変動の周期性を特定する予定である。同時にこれら要因分析を視覚的に提示するアプリケーションを考える。

5.1 謝辞

本研究の一部は、科研費（課題番号 25700033）による。

文献

- [1] Granger, C. W.: Investigating causal relations by econometric models and cross-spectral methods, *Econometrica: Journal of the Econometric Society*, pp. 424-438, 1969.
- [2] Sakaji, H., Sekine, S, and Masuyama, S.: Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns, in 7th International Conference on Practical Aspects of Knowledge Management, pp. 111-122, 2008.
- [3] Yuki Awano, Qiang Ma, Masatoshi Yoshikawa: Causal Analysis for Supporting Users' Understanding of Investment Trusts, *iiWAS 2014*:pp, 524-528, 2014.
- [4] Hiroya Takamura, Takashi Inui, Manabu Okumura : Extracting Semantic Orientations of Words using Spin Model, In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp 133-140, 2005.
- [5] Sakai, H. and Masuyama, S.: Assigning Polarity to Causal Information in Financial Articles on Business Performance of Companies, *IEICE Trans. Information and Systems*, Vol. E92-D, No. 12, pp. 2341-2350, 2009.

- [6] 東山昌彦, 乾健太郎, 松本裕: 述語の選択選好性に着目した名詞評価極性の獲得, 言語処理学会第14回年次大会論文集, pp.584-587, 2008.
- [7] 青野荘志, 太田学: 要因検索による因果関係ネットワークの構築と因果知識の獲得, DEIM Forum2010, B9-1, 2010.
- [8] 小野 博紀, 内海 彰: イベントの時系列分析による因果関係知識の獲得, 人工知能学会論文誌, 30-1B, 2015.
- [9] Lafferty, J. D., McCallum, A. and Pereira, F. C. N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., pp. 282-289, 2001.
- [10] Taku Kudo, Y. M.: Japanese Dependency Analysis using Cascaded Chunking, CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002, pp. 63-69 2002.