Report

J-REIT の ESG 開示情報をテキストマイニング（3）
＜TF-IDF 分析による重要取り組みの識別と、その活用例＞

2020 年 3 月 5 日

株式会社三井住友トラスト基礎研究所
REIT 投資顧問部　研究員　小西勝也

J-REIT 各銘柄の ESG 取り組み状況を評価する上で、各取り組みの重要度を考慮することは不可欠だと考えている。一般的に開示情報内での出現頻度の高い単語（取り組み）の重要度が高いことに異論はないであろう。しかし、それ以外の取り組みの中にも重要度の高いものがあると考えている。例えば、現状実施する銘柄が少ない取り組みでも、今後、J-REIT 全体でその重要性が共通認識となる可能性が高い取り組み等は評価上の重要度が高く、重視すべきである。そこで、本稿前半では、テキストマイニング手法（TF-IDF 分析）を用いて、各単語の文書内での重要度を数値化し、各銘柄の開示資料内で重要度が高い（各銘柄が重要と考える）取り組み（単語）の抽出を行う。その上で、銘柄間比較することで、それら評価上重要な取り組みの識別を試みている。

また、このように各単語を数値化し、文書をその集合と捉えて数量化することができれば、より高度な分析を行うことが可能となる。そこで、本稿後半ではその一例としてクラスター分析を用いた各銘柄の開示文書の類似度に基づく分類を行っている。

分析の結果、定性的にみても違和感のない重要度が高い ESG 取り組みや、実施する ESG 取り組みの傾向が近い銘柄群を概ね把握することができた。本稿では各単語（各取り組み）の重要度に注目し分析を行ったが、前稿まで[1]で紹介したテキストマイニング手法や、それ以外の手法とも組み合わせ様々な角度から分析を重ねることで J-REIT の ESG 取り組み状況のより深い理解が可能だと考えている。

1　はじめに

前稿までは、J-REIT の ESG 取り組みの全体像を把握し、各セクターの ESG 開示内容の特徴について分析した。これらの分析は、単語の出現頻度や出現割合等を基に分析を行っており、各単語（各取り組み）の重要性については考慮していない。しかし、各銘柄が注力する取り組みのうち、どの取り組みが評価すべき重要な取り組みなのかという情報は各銘柄の ESG 評価を行う上で不可欠だと考えている。多くの銘柄の文書に出現し、かつ、その文書内でも繰り返し出現する単語（取り組み）の重要度が高いことに異論はないであろう。しかし、それ以外の取り組み（単語）、例えば、現状実施する銘柄が少数でも、今後、J-REIT 全体でもその重要性が共通認識となる可能性が高い取り組み等についても評価上の重要度は高いと考えている。そこで本稿では、文書内の各単語の重要度を数値化するテキストマイニング手法（TF-IDF 分析）を用いて、各銘柄の開示資料内で重要度が高い単語を抽出し、抽出された重要語を銘柄間で比較し、その傾向を把握することで、それら J-REIT 全体において評価すべき重要度の高い取り組みを識別することを目的に分析を行った。なお、既にほとんどの銘柄が実施している普遍的な ESG 取り組みも分析上の重要度は高くでるが、定性的判断、銘柄間比較などの対象には含めていない。

また、TF-IDF 分析による重要度の数値化は本来更なる分析を行う上での前処理として使われることが多い。そこで、本稿では算出した TF-IDF 値を用いた分析の一例として各単語と TF-IDF 値の文書ごとの集合を用い、クラスタ

1　前々稿（2019 年 10 月 16 日掲載「J-REIT の ESG 開示情報をテキストマイニング＜共起ネットワークによる全体像の可視化＞」）、及び前稿（2019 年 12 月 19 日掲載「J-REIT の ESG 開示情報をテキストマイニング（2）＜対応分析による ESG 開示内容のセクター間比較＞」）

一分析による銘柄別文書の類似度に基づく分類の結果を示した。本分析では、重要度の高い単語を重視した文書分類を行うため、より文書の特徴を正確に捉えた分類が可能となると考えられる。そして、その結果を観察することでどのような銘柄間で実施する ESG 取り組みの傾向が近いのか把握することが可能だと考え、検証を行った。

2　分析手法

2.1　分析の流れ

本稿では、まず始めに TF-IDF 分析を用いて各銘柄の開示文書から重要語の抽出を行う。その上で、銘柄ごとに算出した TF-IDF 値により数量化した文書ベクトルを用いて、クラスター分析による各銘柄の開示文書の分類を行う。

2.2　TF-IDF 分析

TF-IDF 分析とは文書中の単語の重要度を数値化する分析手法であり、主に情報検索やトピック別文書分類などに活用される。TF-IDF の TF は「Term Frequency」、ある単語が、ある文書内で出現する頻度を表す数値である。そして、IDF は「Inverse Document Frequency」、直訳で逆文書頻度となり、その単語の全文書内での希少性（全ての文書のうち、出現する文書の数が多い単語は希少性が低く、逆に出現する文書の数が少ない単語は希少性が高い）を表す数値である。この 2 値を掛け合わせたものが TF-IDF 値であり、文書内で出現頻度が高い（TF 値が高い）ほど、また出現する文書が少ない（IDF 値が高い）ほど、その単語の TF-IDF 値は高くなる。

なお、TF 値、IDF 値、TF-IDF 値、各々の算出に用いる数式は以下の通りである。

TF(Term Frequency)

$$\mathrm{tf}(\mathrm{t},\mathrm{d}) = \frac{\text{単語 t の文書 d での出現回数}}{\text{文書 d の全ての単語の出現回数の総和}} = \frac{n_{t,d}}{\sum_{s\in d} n_{s,d}}$$

IDF(Inverse Document Frequency)

$$\mathrm{idf}(\mathrm{t}) = \log\frac{\text{総文書数}}{\text{単語 t が出現する文書数}} + 1 = \log\frac{N}{df(t)} + 1$$

注 ）最後に 1 を足すのは単語 t が全文書に出現する場合に、IDF 値が 0 になることを防ぐため。

TF-IDF

$$\mathrm{tfidf}(\mathrm{t},\mathrm{d}) = \mathrm{tf}(\mathrm{t},\mathrm{d}) \cdot \mathrm{idf}(\mathrm{t})$$

2.3　クラスター分析（クラスタリング）

クラスターとは、「集団」や「群れ」を意味し、特徴の類似するものが多く集まっている様を表す用語である。クラスター分析は、異なる特徴を持ったサンプルが混在する集団から、互いに似た特徴を持ったサンプルを抜き出し、併合していくことで、クラスターを形成する手法である。本稿では、銘柄単位でのクラスター分析を行った。クラスター分析は計算方法の違いにより、階層的手法と非階層的手法に区分することができるが、本稿では階層的手法を採用している。階層的手法は全サンプル間の類似度を計算した上で、似たもの同士を併合していく方法であり、本稿ではクラスター内の各値からその質量中心までの距離（偏差平方和）を最小化する「ウォード法」を用いる。なお、サンプル間の距離の計算方法には一般的な「ユークリッド距離[2]」を採用している。

クラスター分析は、オープンソースで広く使われているテキストマイニングツール、KH Coder（立命館大学　樋口准教授作）により行う。

2.4　使用するテキストデータ

分析には、以下のテキストデータを収集し、使用した。

1. 投資法人の HP 及び直近決算説明会資料の ESG 情報記載ページ
2. 投資法人が HP で開示する ESG レポート
3. 資産運用会社 HP の ESG 情報記載ページ

注 ） 上記開示資料は全て 2019 年 11 月末時点取得

形態素解析（文を最小単語単位に分割、品詞判定をする処理）はオープンソースの形態素解析エンジン「MeCab」を用いて行った。その上で、分析精度を向上させるために前処理として、テキストの表記（半角・全角、記号等）を一定の規則に基づいて統一する処理や、J-REIT や ESG 関連用語の登録、抽出する語の品詞選択（名詞を使用）、表記揺れの吸収（例：「社員」と「従業員」を同一語と見做すなど）等を行っている。なお、一般的で J-REIT の ESG 取り組みと関連性が低い単語については本分析では分析対象から除外している。

[2] ユークリッド距離では、K 次元の特徴ベクトル、$X_i = (x_{i1}, x_{i2}, \cdots x_{iK})$と$X_j = (x_{j1}, x_{j2}, \cdots x_{jK})$の距離$d_{ij}$は次式で表される

$$d_{ij} = \sqrt{\sum_{k=1}^{K}(x_{ik} - x_{jk})^2}$$

3. 分析結果

3.1 TF-IDF 分析による重要語の抽出結果

前述した TF-IDF 分析により、各銘柄の ESG 開示情報から抽出した重要語が以下の通りである。なお、表中の単語の背景色は 3.2 で後述する①～③の特徴の分類を示しており、①を青色、②をオレンジ色、③を緑色で表している。以下で例示する銘柄はセクターや時価総額規模がばらつくように、選択している。

【図表1】

	銘柄A	銘柄B	銘柄C	銘柄D	銘柄E	銘柄F
セクター	オフィス特化型	住宅特化型	商業特化型	物流特化型	ホテル特化型	複合総合
時価総額	5,000億円超	1,000～5,000億円	5,000億円超	5,000億円超	1,000～5,000億円	5,000億円超
1	従業員	従業員	サステナビリティ	グリーンボンド	評価	イニシアティブ
2	環境	仕事	社会	企業倫理	チームワーク	グリーンファイナンス
3	会議	マテリアリティ	環境	責任	ゲスト	マテリアリティ
4	テナント	サステナビリティ	サステナビリティ戦略	社会	サステナビリティ	スタンダード
5	社会	採用	グリーンボンド	環境	従業員	リスク管理体制
6	人材	キャリア	評価	貢献	優待	枠組み
7	ガバナンス	制度	署名	活動	ストロー	GRI
8	評価	面接	SDGs	ガバナンス	制度	エンゲージメント
9	派遣	就職	汚染	ガバナンス体制	環境	国際
10	制度	研修	賛同	従業員	地域	ガバナンス
11	専門	業務	環境憲章	テナント	チャレンジ	サステナビリティ
12	環境配慮	社会	UNEP	発電	環境パフォーマンス	報告
13	CO2・温室効果ガス	選考	テナント	グローバル	他者	環境
14	認証	新卒	イニシアティブ	コミッティー	ガバナンス	署名
15	専有部	貢献	ガバナンス	エネルギー	社会	MSCI
16	プログラム	評価	指数	地域	配慮	社会
17	研修	学生	従業員	FCPA	改装	外部認証
18	役員	外部認証	MSCI	出力	プラスチック	執行役員
19	コンプライアンス	環境配慮	地域	断熱	モチベーション	ステークホルダー
20	排出	ガバナンス	土壌	効率	失敗	指数

	銘柄G	銘柄H	銘柄I	銘柄J	銘柄K	銘柄L	銘柄M
セクター	オフィス特化型	住宅特化型	商業特化型	物流特化型	ホテル特化型	複合総合	ヘルスケア特化型
時価総額	1,000億円未満	1,000億円未満	1,000億円未満	1,000億円未満	1,000億円未満	1,000億円未満	1,000億円未満
1	認証	サステナビリティ	認知症	サステナビリティ	ガバナンス体制	ロードショー	オペレーター
2	共有	認証	商店街	環境	e-learning	運用報酬	EGAO link
3	社会	環境	巡回バス	社会	女性	評価	ICT
4	報酬	社会配慮	お買い物タクシー	テナント	グループ	GRESB	介護
5	エコキッズ	DBJ グリーンビル認証	ガバナンス体制	働き方改革	一体化	セイムボート	ナースコール
6	内部統制	審議	コンセプト	貢献	講演	弁当	スマートフォン
7	環境	ブレーカー	ビジョン	従業員	プレゼンテーション	足湯	ソーシャルローン
8	キャンドルナイト	制度	連動	ガバナンス	地域	社会	病院
9	探検	コンプライアンスオフィサー	報酬体系	認証	外部人材	投資主利益	老人
10	運用報酬	台風	試験運用	評価	従業員	Green Star	効率
11	方針	災害	こども	制度	協賛	周辺	労働
12	理念	評価	アイシティ	記念財団	厚生労働大臣	専有	センサー
13	投資主価値	電力	コンタクトレンズ	野球	グロービス	オペレーター	サービス
14	子供	研修	アイバンク	ハンディキャップ	平常	誘致	短縮
15	体験	支援	投資主利益	労働	社内体制	BELS	睡眠
16	社会配慮	取締役	回収	尊重	セイムボート	宿泊	巡視
17	評価	地震	サポーター	宣言	外部委員	天井	学園
18	承認	従業員	CS	人権	訪日	認証	カルテ
19	環境問題	スキル	開催	駐車場	投資委員	観光	IC
20	開催	社会	証明	安全性	旅行	開放	投資主価値

（出所）投資法人開示資料をもとに三井住友トラスト基礎研究所が作成

注）セクター分類に関しては、当社 HP に掲載する SMTRI J-REIT Index®のセクター区分に倣い分類している。各セクターの構成銘柄名等、詳細は当社 HP<https:/www.smtri.jp/market/jreit_index/>を参照のこと。

3.2 各銘柄で抽出された重要語について

抽出された重要語を観察すると「サスティナビリティ・環境・社会」等の ESG 取り組みにおける基本的な単語が出現頻度（TF 値）の絶対的な高さから抽出されていることがわかる。しかし、本分析上は、これらの単語が示す既にほとんどの銘柄が実施する普遍的な ESG 取り組み（単語）については重視していない。そこで、それ以外の単語に注目し分析したところ以下 3 つの特徴が確認出来た。

① 時価総額が大きい銘柄で、従業員への取り組み（研修や福利厚生等）を示す単語が上位に多く見られた。

J-REIT 全体でみた場合に時価総額が小さく、上場年数も短い投資法人では従業員への取り組みに関する開示はまだ十分に行われていないことが多い。その一方、規模が大きい銘柄では有給休暇の取得率や研修受講率等の詳細な開示が行われることも少なくない。そこで、そのような銘柄では単語の希少性（IDF 値）、出現頻度（TF 値）の双方の高さから、重要語として抽出されたとみられる。また、定性的に考えても数十人程度の人員で運用を行う J-REIT の投資法人にとって運用の継続性や質の維持という観点から、従業員への取り組みは重要と言って相違ないだろう。

② 時価総額が特に大きい銘柄で、国際的な組織やガイドライン等や、グリーンファイナンス（環境債等）に関する単語が抽出された。

投資法人の規模が大きくなれば、その社会的な責任も当然増大する。国際的なガイドラインの遵守や、グリーンファイナンスの実行は J-REIT 全体の国際的な評価に繋がる重要な項目であり、そのように社会的影響力の大きい投資法人で重視すべき取り組みである。しかし、それ以外の投資法人において取り組む優先度はそこまで高くないといえる。そこで、これら取り組みを実施し開示する銘柄は自然と規模等で上位の銘柄に限られてくる。そのため、希少性（IDF 値）が高く、実施する銘柄では出現頻度（TF 値）も高くなることから重要語として抽出されるのである。また、本取り組みは本来的には全ての投資法人にとって重要な取り組みであり、今後は相対的に規模の小さい投資法人に広まる可能性も十分に考えられることから、評価上の重要度は高いと考えている。

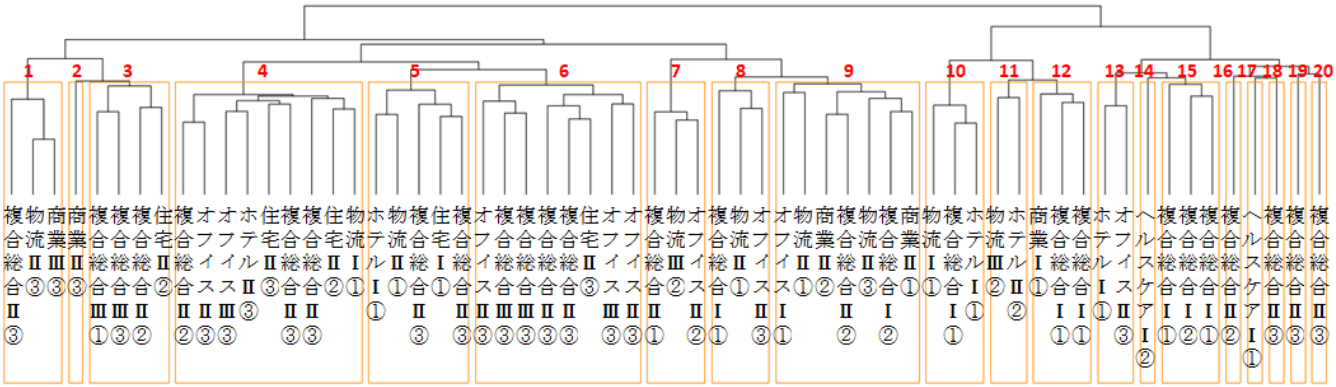
③ 時価総額が小さい銘柄において、その銘柄独自の取り組みを表す単語が上位に見られた。

規模が小さな銘柄で独自の取り組みが上位に入るのは、希少性（IDF 値）の高さでついた差を補うだけの出現頻度（TF 値）が高い単語が存在しないからであり、規模が大きい銘柄と比較して独自取り組みについての記載が多いわけではない。しかし、規模が小さい銘柄では独自の取り組みを行っていない銘柄も多く、これら銘柄間で ESG 取り組みへの注力度を比較する上で、この独自の取り組みを実施し、開示しているか否かは比較基準となり重要といえるだろう。

3.3 クラスター分析の結果

前述したTF-IDF値を用いたクラスター分析により、各銘柄のESG開示情報を分類した結果は図表2のデンドログラム（併合過程を表す樹形図）により、確認することができる。なお、具体的な銘柄名は記載せず、各サンプルにはセクター区分及び時価総額、上場年数のクラスを示す番号を付けている（各番号の意味は図表2右下部を参照）。

【図表2】



[各番号の説明]

Ⅰ	時価総額　1,000億円未満	①	上場年数　5年未満
Ⅱ	時価総額　1,000～5,000億円	②	上場年数　5～10年
Ⅲ	時価総額　5,000億円超	③	上場年数　10年以上

（出所）投資法人開示資料をもとに三井住友トラスト基礎研究所が作成

【図表2の説明】

各銘柄から伸びる直線が他の銘柄から伸びる直線と結合する位置が低ければ低いほどその銘柄間の類似度が高いことを表している。つまり、ESG取り組みにおいて実施する重要な取り組みが似通っており、ESG取り組みに対する姿勢や方向性が近い銘柄とみなすことができる。また、図中のオレンジ色の線の四角はクラスターの併合水準を参考にクラスター数を20と設定し、クラスタリングを行った結果である（赤字の数字はクラスターに左から順に番号を付けたもの）。図表2に含まれる銘柄数は、19年11月末時点上場63銘柄からESG関連情報が得られなかった1銘柄を除く62銘柄である。なお、その銘柄特有の語（独自の造語や略語等）や独自性の極めて高い取り組みを示す単語を分類の基準となる単語から除外するため使用する単語の最低出現回数を50回以上と設定している。

3.4 銘柄分類の結果に対する考察

各銘柄から伸びる直線の結合位置が低い銘柄を見ると同一スポンサーの銘柄間での類似度が高い結果となった。運用会社HPやESGレポート等、同一スポンサーの投資法人で開示資料が全く同一になるものはクラスター分析では分析対象に含めていないが、やはり同一スポンサーの投資法人では運用する物件の属性に関わらず実施するESG取り組みの傾向が近いとみられる。

それ以外の銘柄では、時価総額規模や上場年数が近い銘柄間での類似度が相対的に高い結果となった。これは、規模や上場年数が同程度の投資法人であれば法人としての社会的責任や、人的・コスト面でESG取り組みに費やすことのできる余力も概ね同程度になること、また、各投資法人がESG取り組みを実施し、開示をする際に、自社の競合となる銘柄の開示状況を参考にしていることが理由として考えられる。

また、運用する物件の属性が近い銘柄間でも類似度が高くなると考え、確認したところ、オフィス物件や商業施設を運用する銘柄（複合・総合銘柄を含む。投資比率は問わない。）では、その傾向が認められた。例えば、クラスター6と15がオフィス物件を運用する銘柄を中心とした銘柄群、クラスター1～3が商業施設を運用する銘柄が

中心の銘柄群である。一方、それ以外の銘柄（住宅・物流・ホテル・ヘルスケアを運用する銘柄）では、その傾向は確認できなかった。これは、住宅や物流ではテナント属性や立地等の特性から実施可能な取り組みが限定されるため、同一セクターの銘柄間で共通して実施される取り組みの数もオフィスや商業と比較して少ないことが（詳細は前稿を参照）、ホテルやヘルスケアでは実施する取り組みに各銘柄の独自性が強いものが多いことが要因として考えられる。

4.　最後に

本稿前半の TF-IDF 分析による重要語抽出の結果、定性的にみても違和感のない J-REIT 全体において重要な ESG 取り組みを識別することができた。本分析を用いることで、異なる時点の様々な情報媒体において、同一の基準で重要度を判断することが可能であり、各取り組みの重要度の変化を捉えるにあたっても役立つと思われる。また、本稿後半のクラスター分析では、実施する ESG 取り組みの傾向が近い銘柄群を概ね把握することできた。このように単語や文書を数量化することができれば、例えば各銘柄の ESG 開示状況と投資口価格との関係性の分析など、行える分析の幅は格段に広がるであろう。以上のことから、各単語の重要度を定量化する本手法は当分野の理解を深める上で有用な手法だと考えている。そして、今後もテキストマイニングに限らず、多様な分析手法と組み合わせ、様々な角度からの分析を行うことで J-REIT の ESG 取り組みへの理解をより深めることが可能となるであろう。

【参考文献】

- 樋口耕一 2004「テキスト型データの計量的分析 —2 つのアプローチの峻別と統合—」『理論と方法』(数理社会学会) 19(1): 101-115
- 樋口耕一　KH Coder Index ページ　＜https://khcoder.net＞

【お問い合わせ】REIT 投資顧問部
https://www.smtri.jp/contact/form-reit/index.php