平成 29 年度

学士学位論文

ラフ集合とクラスタリングを用いた就職活動向け企業推薦システム

Company recommendation for job seekers using rough set and clustering

1180352　　中井祐貴

指導教員　　吉田真一

2017 年 2 月 28 日

高知工科大学 情報学群

要 旨

ラフ集合とクラスタリングを用いた就職活動向け企業推薦システム

中井祐貴

興味のある企業を見つけるためには企業調査によってマッチングを図る．手がかりとして事業内容や所在地が挙げられるが，企業を網羅することは困難である．また中小企業では大手企業程，十分な情報を得られる可能性が多いとはいえず，情報が不十分である恐れがある．そこで，本研究では就職活動者向けにラフ集合とクラスタリングを用いた企業推薦システムの提案を行う．テキストマイニングを使ったクラスタリングによって企業理念から各社が大切にしている事を指標として得られ，情報が不十分な企業でも活用可能となる．事業内容にもクラスタリングを行うことで，行っている事業毎に細分化され，探索しやすいシステムとし実装する．評価実験では，ユーザに興味ある事業内容を 4 つ入力させ，各類似性の高い企業提示による関連企業の平均適合率を調べた．その結果，適合率は高いもので 70%と入力に関連ある企業が提示可能となったが，他 3 つは 30%，20%，40%と低い適合率であった．興味ある企業の適合率は高い被験者で 30%と全体としても低い適合率となった．

キーワード　ラフ集合，クラスタリング，テキストマイニング

Abstract

Company recommendation for job seekers using rough set and clustering

In order to find a company that a job seeker is are interested in and desire to join, we suvery company information. Although business contents and location could be found on the web, it is difficult to cover companies. In small and medium-sized enterprises, it is also difficult to obtain sufficient information, and the information may be inadequate. Therefore, in this research, we propose a company recommendation system for job hunters using rough set and clustering. Through clustering using text mining, companies values are extracted from their corporate philosophy, and insufficient information can be used. By clustering of business contents, companies are segmented by their business In experiment, subjects input 4 business contents, precision of resolting companies to user interests are shown. The highest precision is 70% for on interest, while the other three is 30%, 20%, and 40% respectively.

key words rough set, clustering, text mining

目次

第 1 章　序論 **1**

第 2 章　関連研究・技術 **3**

2.1　テキストマイニングに関する研究 3

2.2　テキストマイニング 3

2.2.1　形態素解析 3

2.2.2　TF-IDF 4

2.2.3　コサイン類似度 6

2.2.4　ウォード法 6

2.3　ラフ集合による関する研究 6

第 3 章　提案手法・システム **10**

3.1　企業検索システム構造 10

3.2　システム構築前の調査 11

3.3　提案手法 11

3.4　用いたデータセットについて 11

3.5　マイニングによる指標生成 12

3.6　ラフ集合による指標生成 16

第 4 章　評価実験 **18**

4.1　実験内容 18

4.2　実験結果 18

第 5 章　考察 **21**

5.1　関連と興味ある企業の関係 21

5.2 テキストマイニングの再検討 21
5.3 興味ある企業の指標検討 22
5.4 企業推薦について 22
5.5 ラフ集合の再検討 23

第 6 章 まとめ **24**

謝辞 **25**

参考文献 **27**

図目次

2.1 形態素解析の例 4
2.2 TF-IDF 例 5
2.3 決定表の例 7
2.4 識別行列の例 8

3.1 企業検索システム 10
3.2 一部，事業内容によるクラスタリング 14
3.3 一部，企業理念によるクラスタリング 15
3.4 チェビシェフ距離によるクラスタリング結果 15

4.1 関連ある企業に関する平均適合率 20
4.2 興味ある企業に関する適合率 20

表目次

2.1 TF-IDF の例に用いる文書 5
2.2 下近似、上近似分けた結果 8
2.3 「好き」決定クラスに対する行列 9
2.4 「どちらでもない」決定クラスに対する行列 9
2.5 決定ルール 9

3.1 データの実例 12
3.2 理念を 7 クラス 13
3.3 「技術」とタグ付けした基準とした導出表 13
3.4 業績の向上，識別精度 16
3.5 業績の低下，識別精度 16
3.6 業績の安定，識別精度 16

第 1 章

序論

就職支援することにおいて，就職活動者と企業のマッチングが最も重要なことである. 就職活動者は自身の希望を考慮し，一致する企業を探索する．ここで探索で手がかりとなる要素は勤務地や事業内容，企業理念といったものである．しかし，インターネット上に様々な企業の情報が掲載されているものの，その量は膨大であり，就職活動者一人で把握できる範囲が限られる．また東証一部上場しているような大手企業は業績を決算期ごとに公開しているが，中小企業では十分な情報が公開されていないことがある. 具体的には，会社の規模を知るための従業員数や売上や営業利益のような業績，福利厚生等，全ての情報を開示している企業は多いとはいえない．そんな実情がある中，近年，AI マッチング等の報道があるように，知能技術を用いたマッチングシステムが開発されてきている. 例えば，福岡県糸島市では移住希望者を対象に，人口知能を介したシステムの実験を実施している [1]．日立製作所や山口フィナンシャルグループ等では、新規の取引先企業を見つけるための支援としてビジネスマッチングの実証実験を行っている [2]．AI の発達によって人による経験や知識から助言するのではなく，ログデータを解析することでマッチングのサポートができるようなシステムが多く現れ始めている．

そこで，本研究では不十分な情報でも就職活動者と企業マッチングを図るための企業推薦システムを提案・実装する．本研究の対象者は高知工科大学の情報学群に所属する学生とし，入社したいと思える会社と出会えることを目的とする．また，欠損値やテキスト情報等の数値化しにくい情報を扱うため，ラフ集合や階層的クラスタリング等の知能技術，データマイニング技術を用いる．実装した機能として検索機能をメインとして実装を行い，評価実験よりシステムに用いた手法の有効性を検証する．

第 2 章では本研究に用いた関連技術。研究を記す．第 3 章では，実際の企業情報を用いたシステムについて構築の流れを記す．第 4 章では，被験者に対して行った評価実験について記す．第 5 章では，本研究全体をまとめる．

第 2 章

関連研究・技術

本研究ではデータマイニング技術としてラフ集合とテキストマイニングを用いる．各手法について解説する前に関連研究について述べる．

2.1　テキストマイニングに関する研究

向井らは，Twitter のリツイート情報からユーザのプロファイリングから商品の推薦が行っている [3]．推薦のタイミングに着目し，つぶやき数の急激な増加に合わせて推薦することでユーザに関連ある商品のマッチングを行っている．ユーザのリツイート情報を TF-IDF 値でそれぞれクラスタリングすることで各ユーザに合わせた推薦を可能としている．次に向井らが行なっている形態素解析について説明する．

2.2　テキストマイニング

2.2.1　形態素解析

形態素解析とはテキストを形態素ごと，1 単語ずつに分割する手法である [5]．日本語のように区切りが存在しない文章において有効である．例えば「おすすめの SF 小説を読む」という文章に対して形態素解析を行うと図 2.1 のようなフローで単語毎に区切り，品詞毎に抽出することが可能になる．形態素の分割は辞書に単語を登録することで修正することが可能である．本研究では形態素解析器として MeCab を使用した．辞書は mecab-ipadic-NEologd と呼ばれる Web 上の言語を辞書に追加されている辞書を用いた [6]．未知の単語・形態素に

対応し，解析が行えるようにするために活用する．

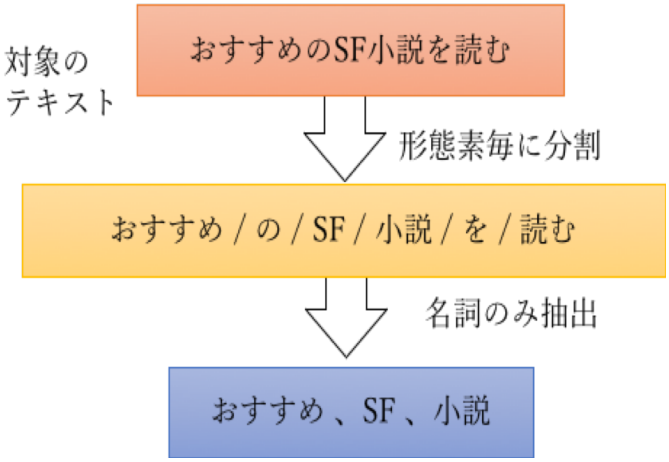

図 2.1　形態素解析の例

2.2.2　TF-IDF

TF-IDF とは文書中の単語を特徴ベクトルとして得るための手法であり，各文書の単語に重み付けを行う一般的な手法である [7]．TF(term frequency) は各文書中の単語の出現頻度を表し，よく出る単語ほど重要であるものとして扱う．IDF(inverse document frequency) は全文書数からある単語を含む文書数で割った値に対して対数をとった値である．式 2.1 は TF を，式 2.2 は IDF を求める式である．d はある文書中の単語頻度，N_i はある文書 i の総単語数，A は解析対象となる全文書，B_d はある単語 d が含まれる文書数である．この 2 つを組み合わせることで各文書がどんな特徴を持っているのか判明する．例として表 2.1 のような文書に対して TF-IDF を行うことにする．

$$tf = \frac{d}{N_i} \tag{2.1}$$

$$idf = \log \frac{A}{B_d} + 1 \tag{2.2}$$

まず各文書中にある単語の頻度を求める．例えば文書 1 の「うどん」の値を知りたい時は $\frac{2}{4}$ より $\frac{1}{2}$ という値が得られる．次に IDF を求める．「うどん」を対象に値を求めると $\log \frac{2}{2} + 1$ より 0.823 と得られる．最後に 2 つの値を掛け合わせると $\frac{1}{2} \times 0.823$ より 0.4119

表 2.1　TF-IDF の例に用いる文書

文書 1	カレー、うどん、うどん、カツ丼
文書 2	スパゲッティ、うどん、親子丼、カツ丼

という値が得られ，これが文書 1 の「うどん」という単語の重要度を指し示す．図 2.2 にうどん以外の単語に対する重み付けの値を示す．赤文字は 2 つの文書共に現れた単語である．ここで文書 1 の「うどん」が文書 2 の同単語よりも重要度が高い理由として文書 1 内で複数回現れた事で，文書 1 を特徴付ける要素のためである．

文書・単語	TF	IDF	TF-IDF
文書1 カレー	$\frac{1}{4}$	$\mathrm{Log}(\frac{2}{1})+1$	0.32
文書1 うどん	$\frac{1}{2}$	$\mathrm{Log}(\frac{2}{2})+1$	0.5
文書1 カツ丼	$\frac{1}{4}$	$\mathrm{Log}(\frac{2}{2})+1$	0.25
文書2 スパッゲッティ	$\frac{1}{4}$	$\mathrm{Log}(\frac{2}{1})+1$	0.32
文書2 うどん	$\frac{1}{4}$	$\mathrm{Log}(\frac{2}{2})+1$	0.25
文書2 親子丼	$\frac{1}{4}$	$\mathrm{Log}(\frac{2}{1})+1$	0.32
文書2 カツ丼	$\frac{1}{4}$	$\mathrm{Log}(\frac{2}{2})+1$	0.25

図 2.2　TF-IDF 例

2.2.3　コサイン類似度

コサイン類似度は TF-IDF と共に文書の類似性を検討する事によく使われる手法である．語句ベクトル，文書ベクトルについて相関係数や内積などを用いて類似性を検討する手法でもある [8]．a_i と b_i はベクトル a と b の要素であり，ここでベクトルとは文書，要素は文書中の単語ベクトルである．式 2.3 から文書と語句から類似度を算出することができる．また文書の長さに依存しないように比較するため、文章の長さに関わらず、類似性を算出可能である。

$$\cos(a, b) = \frac{\sum a_i b_i}{\|a\| \cdot \|b\|} \tag{2.3}$$

2.2.4　ウォード法

ウォード法は 2 つのクラスタ距離をクラスタ内変動の増加分で適宜，距離の小さなクラスタから融合していく方法である [9].

$$D(A \cdot B) = \sum_{x \in A,B} d(x, \mu_{AB})^2 - \left(\sum_{x \in A} d(x, \mu_A)^2 + \sum_{x \in B} d(x, \mu_B)^2 \right) \tag{2.4}$$

2 つのクラスタ A，B 距離をクラスタに属するデータの平均ベクトルを用いて式 2.4 によってクラスタ内の変動を算出する．これの増加分が小さい値，すなわち類似度が高いものであれば，クラスの融合を行われ，階層型クラスタリングとして樹形図で提示される．樹形図はデンドログラムとも呼ばれ，本研究ではコサイン類似度をベクトルの値として取り扱い，距離の算出によってクラスタリングを行う．

2.3　ラフ集合による関する研究

大東は，企業の信用度評価として企業が倒産するか，存続するかどうかをラフ集合にて解析し，学習結果から予測を行っている [4]．現金の流れと経済記事といった定量的データと定性的データから倒産企業の判別を行っている．株価の指標として用い，実際に倒産すると識別された企業は用意されているデータから半年から 1 年の間下落が確認されており，ラフ

集合の学習結果から生成された指標は企業評価支援に有効であることが示唆されている．次に大東が行なっているラフ集合について説明する．

■ラフ集合　ラフ集合は企業の所在地，従業員数，売上等の属性情報から，将来の業績上昇下降等の予測したい情報 (決定属性と呼ぶ) を導くルールを求める手法である [10]．利点は識別が困難な対象に対し，少ない手がかりでも識別可能になる知識を獲得できることである．例として，対象とするモノに対する「好き」又は「どちらでもない」を識別するためのルール，知識を求めることにする．図 2.3 のような決定表から求める．これは対象集合と形や色といった条件属性集合,「好き」「どちらでもない」のような決定属性集合から構成される，決定属性集合の属性値を決定クラスとして利用し，クラス分けを近似で行い，ルールの生成を行う．

対象集合	条件属性集合			決定属性集合
対象	形	色	イメージ	好み
A	有機的	白黒	スポーティー	好き
B	曲線的	色彩	パーソナル	どちらでもない
C	曲線的	白黒	パーソナル	どちらでもない
D	曲線的	白黒	パーソナル	好き
E	有機的	色彩	スポーティー	好き

図 2.3　決定表の例

まず，決定表を用いて識別行列を生成する．決定クラスが異なる同士を比較し，属性値が異なる箇所があればその条件属性を，一致すれば空集合 (ϕ) とする．また決定クラスが同じなものは比較しないでアスタリスク (*) と置く．実際に識別行列にしたものが図 2.4 である．

ここから決定クラスの識別に必要な属性の特定をする．図 2.4 から識別に必要な属性が列挙されているので，それを論理式の形で識別ルールとして取り出す．例の場合 ”(形 $\vee$ イ

	A	B	C	D	E
A	*				
B	形,色, イメージ	*			
C	形, イメージ	*	*		
D	*	色	null	*	
E	*	形, イメージ	形,色, イメージ	*	*

図 2.4　識別行列の例

メージ)∧ 色 ” が識別に必要最低限なルールである．

識別行列より生成された識別ルールを用いて，分類してみると {{A}，{B}，{C，D}，{E}} となった．この分類された結果から，各決定クラス毎に下近似と上近似の 2 つの集合を生成する．下近似とは確実に決定クラスに属する対象の集合，上近似は決定クラスに属する ” かもしれない ” 対象の集合である．ルールを用いて分けた結果が表 2.2 である．

表 2.2　下近似、上近似分けた結果

	下近似	上近似
どちらでもない決定クラス	B	B,C,D
好き決定クラス	A,E	A,C,D,E

ここで下近似に分類された対象の属性値に注目すると同じ要素を持つ対象がいないことがわかる．同じ要素を持たない対象は，付与されている決定クラスとして識別できることから下近似に分類される．上近似にしか現れない対象は，同じ要素を持つ対象が複数存在するため，決定クラスの識別が困難である．識別困難ではあるが，属する可能性がある集合として取り扱えるため上近似として A と E が該当する．識別ルールを用いることで，知識として「好き」「どちらでもない」であるための必要最低限な要素を知ることができる．

最後に，求められた下近似と決定クラスに属するデータを比較，差を取る事で，決定クラ

スに属するデータの識別を可能とする知識，決定ルールを獲得することができる．具体的には各決定クラスの下近似と他の決定クラスの要素と比較し，異なる属性値を列挙することでルールを特定する．表 2.3，表 2.4 から決定ルールを導出するための条件部を知る．この条件部を論理式で求めることで，表 2.5 のように条件部から構成された決定ルールを導出することが可能になる．

表 2.3　「好き」決定クラスに対する行列

	B	C
A	有機的，白黒	有機的，スポーティー
E	有機的，スポーティー	有機的，色彩

表 2.4　「どちらでもない」決定クラスに対する行列

	A	D	E
B	曲線的，色彩	色彩	曲線的，パーソナル

表 2.5　決定ルール

決定クラス「好き」ルール	有機的	スポーティー
決定クラス「どちらでもない」ルール	曲線的かつ色彩	色彩かつパーソナル

第 3 章

提案手法・システム

本章ではシステムの構築の流れを説明を行う．

3.1　企業検索システム構造

本研究で構築したシステムは図 3.1 のような構造となっている．ユーザーは業種や理念，所在地，業績予測の一覧から選択し情報の送信を行う．また事業内容，企業理念のテキスト情報も送信可能である．サーバで送信された情報を基に適合する企業をサーバを介してブラウザへの提示を行う．事業内容と企業理念の情報が送信された場合はサーバ上で R によるクラスタリングを実施する．クラスタリング結果から入力されたテキスト情報に近い企業の提示を行うことで，事業内容又は企業理念の近い企業の発見を行えるように実装する．

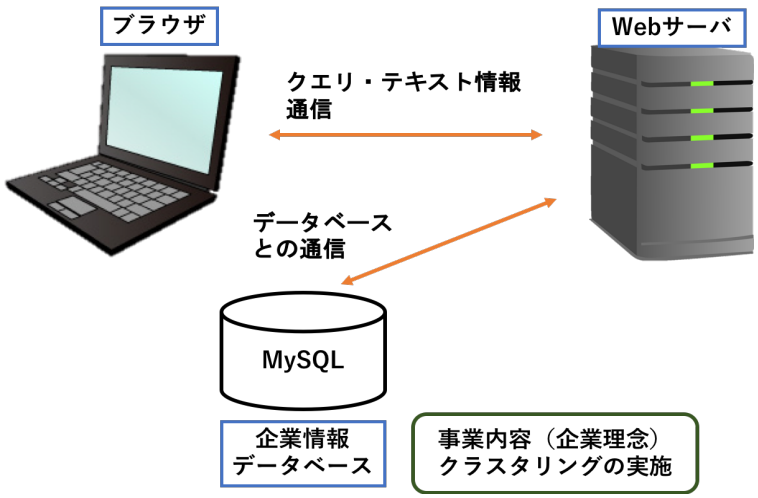

図 3.1　企業検索システム

3.2　システム構築前の調査

システムを構築する前に就職活動者が何を決め手としていたかを調査するために，ヒヤリングを実施した．企業を探す上で重要視した点は事業内容，勤務地，福利厚生，社内の雰囲気であった．また社員の声や企業理念，福利厚生といった会社の環境も重要視されていた．ヒヤリング結果から，企業理念と事業内容からクラスタリングを行うことにした．各クラスの理念はキーワードとして提示することで大切にしている考え方が何かを示し，データセットに用いることにした．事業内容は記載されている内容が近い企業をグルーピングすることで探索の容易化を図るため，検索する際のシステムとして実装する．

3.3　提案手法

本研究では，ラフ集合の解析結果から業績予測指標を導出し利用するが，中小企業ではデータが十分に得られない可能性がある．そこで，テキストマイニングによるクラスタリング結果を用い，併用することで十分に得られなかった企業の指標を増やし，業績予測指標の導出が行えるようにする．業績予測指標である決定属性は大東の研究では 2 値としていたが，企業は業績が上がる、下がるだけでなく安定している企業もあるため，3 値として「上がる」「下がる」「安定」とすることで業績が安定している企業の発見も行えるようにする．

3.4　用いたデータセットについて

システムに登録するデータセットとして 537 社分の 2016 年度の決算情報を用いる．学習データを 203 社とし，残り 334 社は一部データの記載のない欠損値のある属性集合で構成されている．データセットは高知工科大学の情報学群を卒業した学生が入社した企業を基本とし，ルール精度の向上のために東証一部上場企業を含め学習を実施する．属性の要素としては売上，営業利益の増減とその差分，連結又は単独従業員数，社員 1 人分の生産性，業種，所在地方，理念，業績とした．決算情報として連結決算で公開されている企業は連結での結果を使用する．これは単独の場合，子会社や関連会社との取引による粉飾によって業績

が悪いにも関わらず，良いと判断される可能性があるため，連結決算の情報を利用した [11]. 連結決算の開示がない企業は関連会社が存在しないということで単独決算の情報を用いる. 営業利益は主となる事業でどれだけ儲けることができたか，売上はサービスや事業全ての儲けである．データセットは表 3.1 のような構成である．空白は企業のサイトで情報が開示されていなかったため，欠損値として取り扱う．純利益は本研究において業績として取り扱い，2016 年度の純利益と過去 2 年分の平均純利益の差を比較する．学習データはあらかじめ比較の結果から「上がる」「下がる」「安定」を付与することでラフ集合による解析から業績予測ルールを得る.「安定」が付与される条件として，純利益の差額が 2016 年度の純利益において ±10%の割合で収まっているならば，差が微々たるものであるため「安定している」と見なした.

表 3.1　データの実例

企業	売上 (十億円)	営業利益	営業差分 (十億円)	従業員数 (連結)	従業員数 (単独)	生産性	業種	所在地	理念キーワード	純利益
A	20	増加	0.2		2000	10000000	ソフト受託開発	関東	技術	安定
B	8	減少	0.01	700	700	11428571.43	ソフト受託開発	関東	技術	上がる
C	20				900	22222222.22	ソフト受託開発	関東	お客様・サービス	上がる可能性がある
D	3				400	7500000	ソフト受託開発	関東	社会貢献	下がる可能性がある

3.5　マイニングによる指標生成

まずデータセットを用意するために企業理念をクラスタリングを行う．類似性の近い企業についてクラスタリングを行うために，TF-IDF で文書を解析し，コサイン距離によって類似度を算出し，ウォード法によって階層型クラスタリングを実施した．閾値は企業理念の場合，少ないクラス数で提示することで大まかに考え方を把握するため，距離が 2 の地点とし 7 クラスで表現した．各クラス名は表 3.2 にとなっている．これはクラス分け後，各クラスの単語の頻出度の多い名詞のものを代表語とし与えている，クラスタリング結果はデータ

セットに反映させ，ラフ集合の解析にも用いる．技術に関するタグ付けの基準となったものは表 3.3 で上位 6 単語を提示し，名詞のみに着目した．社会やお客様といったものはどのクラスにも上位 3 単語に存在するため，組みあわせて 1 単語とできる場合は，1 単語とした．

表 3.2　理念を 7 クラス

お客様・サービス	お客様主義	会社	技術	社会貢献	創造	未来

表 3.3　「技術」とタグ付けした基準とした導出表

上位 6 単語
技術
社会
貢献
お客様
創造
価値

事業内容も企業理念と同様のクラスタリング手法を用いた．事業内容は細分化する必要があったため，距離を 1.5 の地点でクラス分けを行うようにし，25 クラス程度で示す．事業内容によるマイニングは検索機能として実装し，入力された内容に近い類似性の高い企業を提示する．よって事業内容にはタグ付けは行わなかった．図 3.2，図 3.3 は実際のクラスタリング結果である．横軸が各社の企業理念又は事業内容である．コサイン類似度の他に類似性を比較するために距離行列を求める、チェビシェフ距離を活用を考えていたが、図 3.4 のような形となった．クラスタリングの信用度を考慮すると、クラス分けが出来ていない恐れがあったため活用せずに、コサイン類似度を利用する．

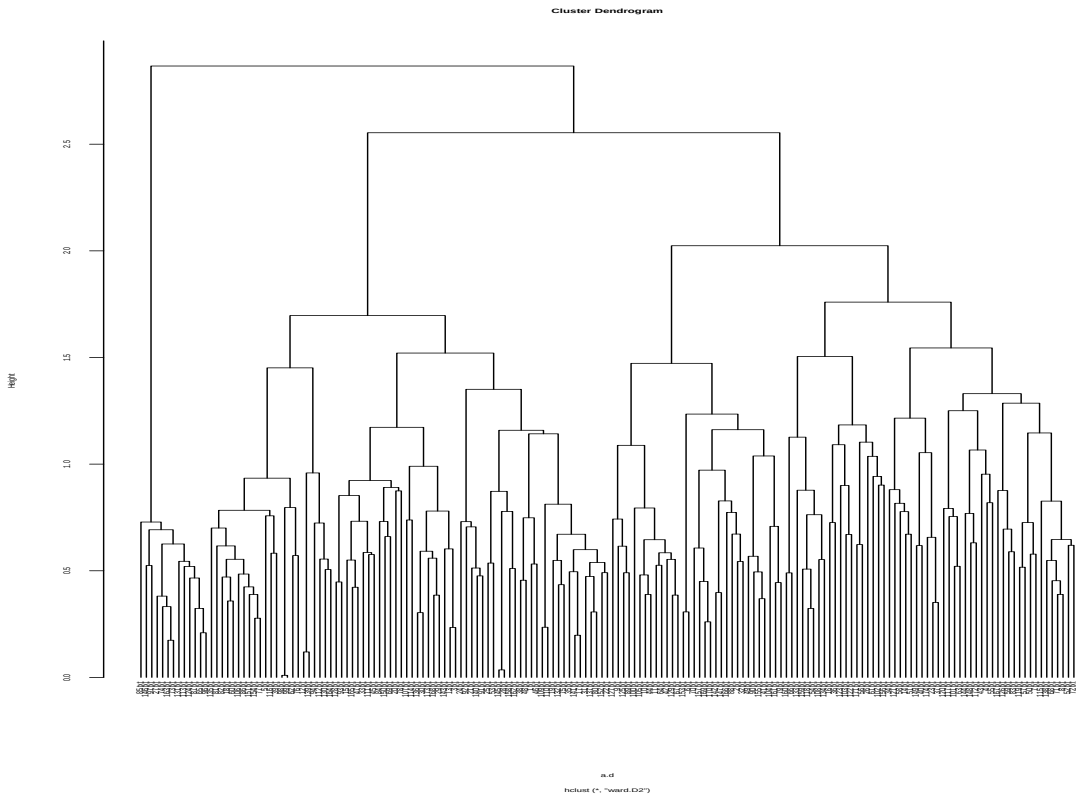

図 3.2　一部，事業内容によるクラスタリング

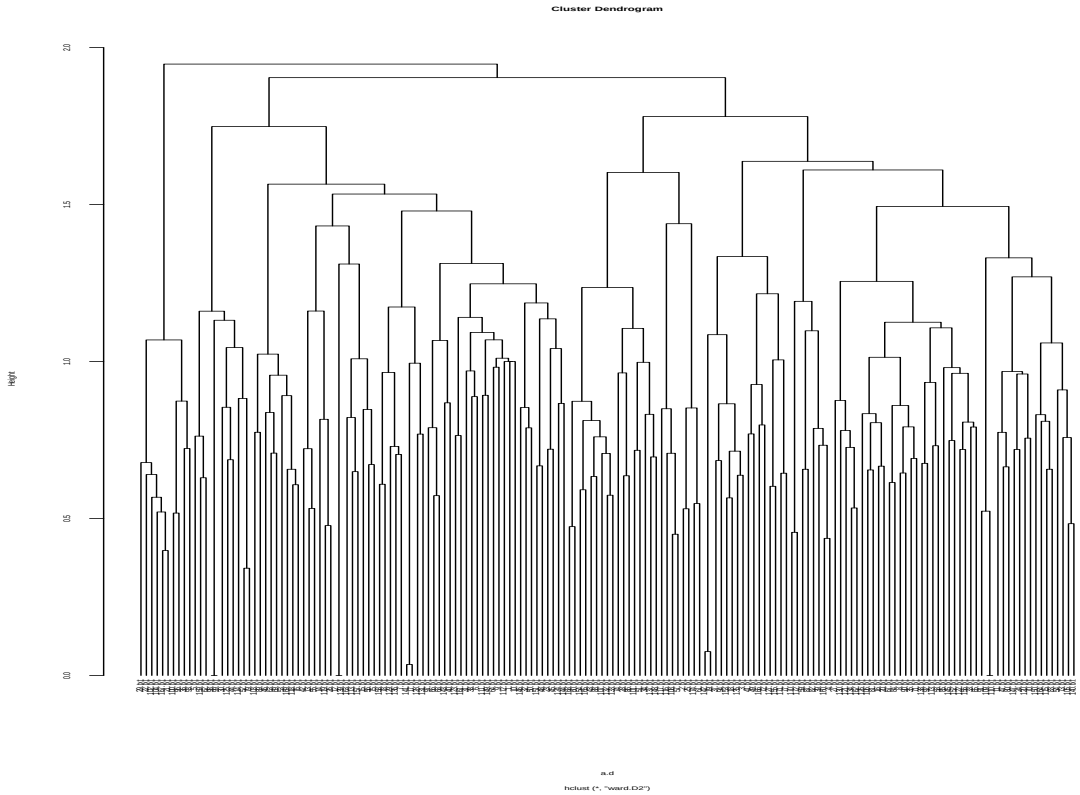

図 3.3　一部，企業理念によるクラスタリング

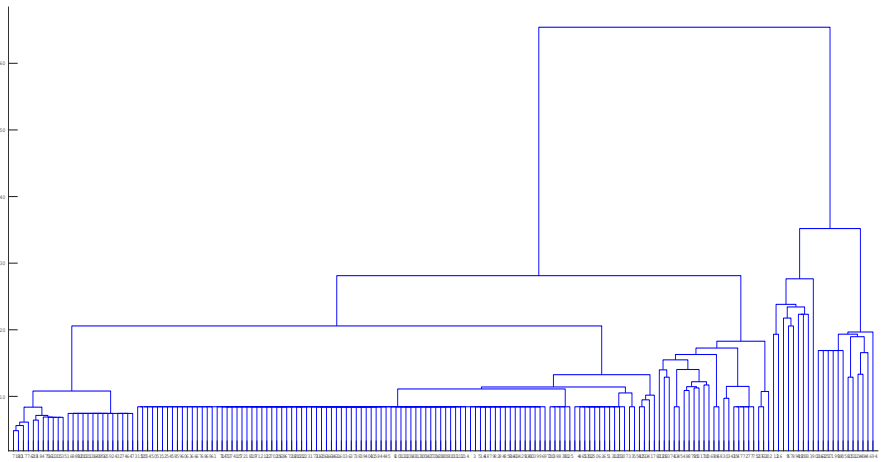

図 3.4　チェビシェフ距離によるクラスタリング結果

3.6　ラフ集合による指標生成

ラフ集合による解析結果で業績予測の指標を生成する．学習データに対して，各決定クラスを導出するためのルールを得る．ルールを使用する前に全ての条件を使ったものと条件が複数回使用されているものだけを使ったもので学習データに対する精度検証を行った．表 3.4，表 3.5，表 3.6 に精度を示す．

表 3.4　業績の向上，識別精度

	全ての条件	C.I. 値を優先
正解	0.686	0.961
誤認識	0.314	0.039

表 3.5　業績の低下，識別精度

	全ての条件	C.I. 値を優先
正解	0.250	0.799
誤認識	0.750	0.201

表 3.6　業績の安定，識別精度

	全ての条件	C.I. 値を優先
正解	0.456	0.775
誤認識	0.544	0.225

精度結果より全ての条件を使用するよりも，C.I. 値より複数回利用している条件のみを決定属性の識別に利用した方が精度が良いため，複数回利用されている条件を使用する．業績予測の属性をルールから導出するために優先度を決めた．ルールの優先度として，”業績が上がる＞業績が下がる＞業績が安定する ”とし．各ルールの決定クラスと一致するか否かを判定する．優先度は各ルールの精度より決定した．欠損データに対しては業績が下がるルー

ルを適用すると，全ての会社が「下がる」と認識してしまうので”業績が上がる＞業績が安定する＞業績が下がる”の優先順にした．ラフ集合とクラスタリングによって得られた指標をデータセットに付与し，企業探索できるようにした．

データセットを利用し，企業推薦システムとして構築を行い，探索機能に特化した．企業本社の所在地方，業種，理念キーワード，業績で企業一覧を提示する機能と事業内容又は企業理念を入力してもらい，類似性の高い企業を提示する機能の 2 つを実装した．ラフ集合で得られたルールも提示した．

第 4 章

評価実験

本章では本研究で構築したシステムについて，評価実験の内容と結果について述べる．

4.1 実験内容

データセットより得られた解析結果を用いてラフ集合による業績予測指標とクラスタリングによる類似性の高い企業の提示が有効であるか評価する．被験者 7 名に対し，事業内容に関するワード入力してもらい，提示した会社に興味を持ったか，又クラスタリングがうまくされているか検証する．入力する事業内容は制限し,「システム」「ネットワーク」「アプリケーション」「IT コンサルティング」に限定した．

興味のある企業の提示の検証として，被験者 7 名のうち 3 名に検索結果から提示された企業の内，興味のあった企業数から適合率の算出をおこなった．またクラスタリングの検証として，提示企業が入力内容に関係ある企業かどうか検証した．指標の有効性について検証するために，実験後に興味ある企業は何を決め手としたヒヤリングを実施した．

4.2 実験結果

クラスタリングによる提示は関連あるとした会社の平均適合率は「システム」以外は 50%未満とクラスタリングによる提示は良くない．システムは 70%と高い数値を示しているが，ヒヤリングから「システム」に関連ありそうな会社が別の入力で提示されていたという意見があり，うまくグルーピングされているとはいえない． 興味を持った会社の適合率は高いものでも 30%であった．被験者の一番興味ある事業は E と F は「システム」，G は

「ネットワーク」に興味がり，E に対する企業推薦が良い結果が得られた．各適合率の算出は式 4.1 と式 4.2 で行った．

$$興味ある企業に関する適合率 = \frac{興味あり企業数}{提示企業数} \tag{4.1}$$

$$関連ある企業に関する適合率 = \frac{関連あり企業数}{提示企業数} \tag{4.2}$$

実験後のヒヤリングから興味を持った企業の決め手として事業内容に注目していたことがわかった．また業績を気にしない被験者が多かった．しかし，理念を決め手とした被験者がおり，クラスタリングによる理念をキーワードとして提示することの有効性があった．システムの評価として，おおまかに分野が分かれている点や比較しやすい点が良い点として挙げられていた．

4.2　実験結果

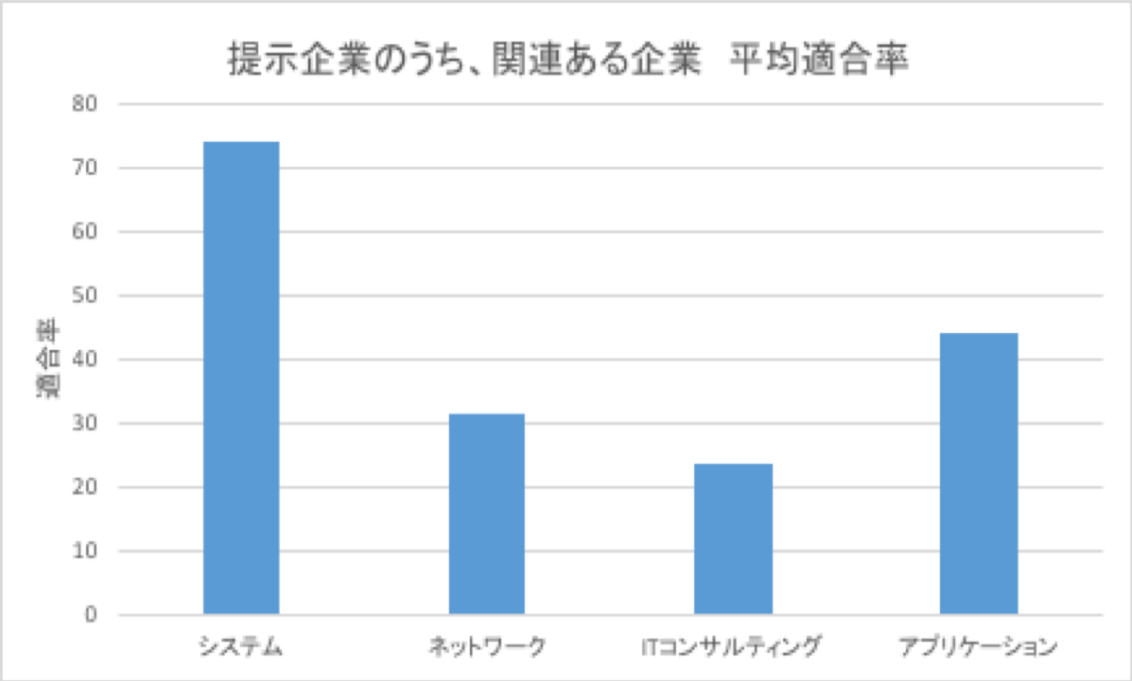

図 4.1　関連ある企業に関する平均適合率

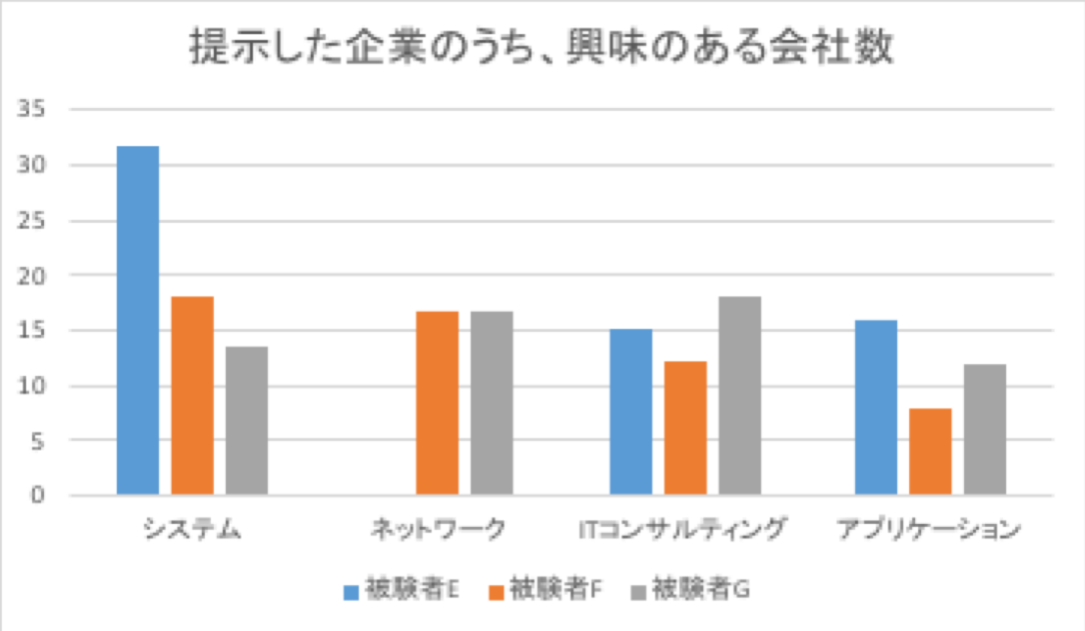

図 4.2　興味ある企業に関する適合率

第 5 章

考察

5.1　関連と興味ある企業の関係

関連ある企業数と興味ある会社には関係性があり，関連ある企業数が多いほど興味ある会社数が増加している．これはクラスタリング結果より，類似性の高い企業のクラス分けに影響があると思われる．事業内容ごとに細かく分けることで IT というくくりではなく，公共向けのシステム開発やネットワーク保守といった視点での探索を可能することでシステムの利用者が「やりたい事」で検索しやすいように行った．しかし，クラスタリングに使用した企業の事業内容は会社によってはシステム開発だけでなく，ネットワークやプラットフォーム等，他にもサービスしている事が多くあった．そのため，他にサービスを行っていれば，システム開発を主に提供している会社だとしてもクラスタリングがうまくできなかった恐れがある．よって，事業内容の種類を考慮する必要がある．クラスタリング後に主としている事業と副事業を明確にすることで事業内容で検索しても，多様に提供できる会社として判断することが可能になると考える．

5.2　テキストマイニングの再検討

本研究でクラスタリング結果を良くするため，TF-IDF とは違う手法の検討が考えられる．本研究では文書の特徴を捉えるために使用したが，共通する単語も重要視するポイントであると思われる．事業を数種類やっているかどうか，捉えるためには文書間で複数使用されている単語に注目することで，クラスタリング結果が改善されることが見込まれる．また

共通する言葉があるという事は，その情報を掲載している企業は似ている可能性があるためである．

5.3　興味ある企業の指標検討

興味のある会社数が全体的に 30%程度と低い結果から「やりたい事」に近い会社が少なかった事が考えられる．これは対象者を高知工科大学の情報学群に属する学生が，入りたい会社を見つけるために卒業生が入社している会社を中心にデータを集めたため，知りたい企業が登録企業でも少なかったことが考えられる．また興味ある会社があれば，会社説明会や会社選考に進むはずなので，学生にアンケートを取り，多い人でどれだけ会社説明会や選考に進んだか調査する必要がある．会社説明会や選考に参加した数の平均を取れば，興味を示した数が多い又は少ないの判定に利用できる．

5.4　企業推薦について

興味のある企業の評価に協力してもらった被験者に対する企業推薦の考察として，被験者の一番興味がある事業の適合率が高いわけではない理由として、単純に興味を持てなかったことが考えられる．それは事業内容を重視した結果であり，事業内容に偏りがあった恐れがある．そこでまず，データセットを増やすことが検討される．卒業生の入社の有無に関わらず，IT 企業を網羅することで興味ある企業の適合率の増加が見込めると考える．次に事業内容に関して，具体例や導入例を含めた上でテキストマイニングを行うことで事業での細分化だけではなく，サービスをどこへ提供しているのかを明確にすることが出来，探索の手がかりとなり，システム利用者が興味ある企業の発見を多くすることができるのではないかと考える．

5.5　ラフ集合の再検討

ラフ集合によって導出された業績予測指標はヒヤリング結果からあまり手がかりとはならなかった．手がかりとしていた被験者は１人だけであったが，興味ある企業の手がかりとして事業内容に注視していた．この事より今後の業績によって入りたい，興味のある会社が働ける環境がなくなる恐れを考慮するよりも，事業内容が企業を探したい人にとって興味引くものかどうかを重視しているといえる．そこでラフ集合を何か予測する指標を導出するためではなく，選ぶ傾向と被験者の性格と考え方を導出するように見直しが必要である．会社にも人によって性格や考え方が異なることで社内の雰囲気に「合う」 又は「合わない」があると思われる．そこで事前に性格診断から性格や考え方を知り，ラフ集合での解析で考慮することで会社選択の手がかりに「こんな人に向いている」という指標より，個人個人に合いそうな企業の推薦を行えることが可能になると考える．

第 6 章

まとめ

本研究はラフ集合とクラスタリングによる企業推薦システムとして，高知工科大学の情報学群の学生を対象とし構築した．評価実験によって業績予測指標の有効性とクラスタリングによる類似性の高い企業の提示，並びに理念に対しても実施しキーワードとして提示することの有効性を検証した．結果は興味を持った企業の適合率は低く，クラスタリング結果が良くないことがわかった。また関連ある企業数による適合率は「システム」は 70%と高い適合率であるものもあったが，興味を持った企業と同様，その他の項目は低い適合率であった．そのことから興味と関連ある企業には関係性があり，関連ある企業の提示が多いほど，興味をもってもらえる企業が増やすことができることが判明した，ラフ集合で導出した業績予測指標は手がかりは興味ある企業を選択する際に手がかりとはなっておらず，有効性はなかった．そのことからラフ集合をデータとしてではなく，傾向の解析に用いることで利用者毎の推薦が可能になると考えられる．また適合率の結果からクラスタリングに共通の単語も考慮することで類似性の高い企業を見つけることが可能になるのではないかと考える．より個人に合わせたマッチングによる企業推薦を行うことで，個人に合う会社を見つけやすくなるのではないかと考える．各適合率の増加のためにはデータセットを増やし，IT 企業を網羅するだけではなく，事業内容とサービスをどこへ提供しているのか明確化することで，事業内容での細分化だけではなく，提供先の細分化も考慮し興味ある企業の発見がよりできると考える．

謝辞

本研究を進めるにあたり，ご指導して頂きました高知工科大学　情報学群　吉田真一准教授には大変お世話になりました．吉田先生には卒業研究の梗概，締め切り 30 分前まで研究室内で進捗の悪かった私のサポートを 1 日中して頂いた事や活用した手法の解説やご指導をして頂き，研究では常に助けてくださりました．研究室活動ではよくイベントや飲み会の席でお話させていただくことが多く，タメになることが多くあり，活動を通して様々な経験をさせていただきました．今後社会に出ても，交流があると思いますので，これからもよろしくお願い致します．深く感謝致します．

また，本研究の副査を引き受けて頂いた高知工科大学　情報学群　妻鳥貴彦准教授ならびに高知工科大学　情報学群　横山和俊教授には，深く感謝致します．卒業研究発表当日，質問して頂き，自分の研究の改良点を知ることが出来，よりよい研究にすることができました．吉田研究室の皆様には配属から今までお世話になりました．今の私があるのは吉田研究室で勉強だけでなく，色々な貴重な体験から成長することが出来た結果だと思います．所属当初はほとんどの方から心配されるほど泊まり込みで作業していたり，入り口で座っているのに入りづらい空気を出してしまい，色々ご迷惑をお掛けしました．修士 1 年の方々には日頃からお世話になり，些細な質問から真面目な質問まで忙しいのに答えてくれてよかったです．佐々木氏は，初めて研究室訪問での仰られた言葉がなければ，今の自分がいないと言えるほど，きっかけをくれたと思います．笹谷氏は，3 年時は真面目な質問ばかりでしたが，4 年時からは一緒に出かけたり，遊ぶことが多くなり勉強面と共に非常にお世話になりました領内氏は，いつも気にかけてもらい，3 年時に色々抱えてた時期に話かけてもらい，さらに長時間、話を聞いてもらった際は気持ちが落ち着き，本当に助かりました．中山氏は，些細な相談から研究面までお世話になり，また夜遅くまで作業している所をよく見かけ，何事にも根気よく取り組む姿を学んだ気がします．

同期の 4 年生の方々とは，ソフトウェア工学から仲が深まり，喋れるようになったと思い

ます．私が人見知り気味なので，なかなかこちらから喋ることが少ないのに，皆しゃべりかけてくれてた事は本当に嬉しかったです．今川氏は，配属前の実験からお世話になり，同期の中でまとめ役として動いていた印象が強いです．誰よりも真面目で卒業研究で特許に繋がる研究に成功したのも納得します．松崎氏は，4 年時に席替え後からよく話す仲になりましたが，いつも「出来ない」と悲観しながらも作業の異常な速さと要領の良さには見習いたいと思ってました．鎌倉氏は，研究室での静かさから一転，サークル活動での姿とのギャップのインパクトが強いです．喋る時は喋り，作業する時は作業すると気持ちの切り替えがうまく出来ていて羨ましく感じます．馬場氏は，配属時からお世話になり，お互いに足りない要素を補い，OC ポスターやソフトウェア工学をしたことを思い出します．プライベートでも同期の中で一番仲良くさせてもらい，楽しい思い出から学ぶことも多くあり，出会えてよかったと思います．山中氏は，夜遅くまで作業する際に休憩時間によく色んな話をして，特に研究に関する話を多くしたことが思い出としてあります．OC や研究における解析も苦労しながらやっている姿からしっかりしている印象が強いです．山口氏は，朝早く研究室に来て，夜遅くまで作業している姿をよく見て無理してないか、心配でした．ちゃんとスケジュールを立てて行動し，後輩の面倒も見ていて，視野が広いと思います．

また，同研究室の 3 年生皆様にはオープンキャンパスやイベントなどで活躍してもらい，また被験者実験にも協力して頂きました．来年度は，卒業研究や就職活動が大変だと思いますが，頑張ってください．

最後に，大学生活を金銭面および精神面で支えて頂いた家族に深く感謝致します．また，大学生活で出会った全ての皆様に感謝致します．

参考文献

[1] 蓬田正志，“記者有情:AI マッチング/福岡-毎日新聞，” https://mainichi.jp/articles/20171209/ddl/k40/070/528000c(参照 2018/2/12).

[2] 金澤雅子，“AI で「ビジネスマッチングサービス」を高度化–山口 FG と山口銀行、日立が実証実験へ，”http://www.itmedia.co.jp/enterprise/articles/1801/16/news060.html(参照 2018/2/21).

[3] 向井友宏，黒澤義明，目良和也，竹澤寿幸，“マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案，” 言語処理学会 第 17 回年次大会発表論文集，2011.

[4] 大東真，“企業の信用度評価への定性的情報の適用，” 平成 20 年度高知工科大学学士学位論文，2009.

[5] 石田基広，“R によるテキストマイニング入門，” 森北出版株式会社，2008.

[6] 佐藤敏紀，橋本泰一，奥村学，“単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討，” 言語処理学会 第 23 回年次大会，2017.

[7] 新納浩幸，“R で学ぶクラスタ解析，” 株式会社オーム社，2007.

[8] 豊田秀樹，“データマイニング入門 -R で学ぶ最新データ解析-，” 東京図書株式会社，2008.

[9] 平井 友三，“はじめてのパターン認識，” 森北出版株式会社，2012.

[10] 森典彦，田中英夫，井上勝雄，“データからの知識獲得と推論 ラフ集合と感性，” 海文堂出版株式会社，2004.

[11] 吉木伸彦，田中英夫，井上勝雄，“会計知識ゼロでも読める連結決算の読み方 使い方，” 東洋経済新報社，2001.