1F4-2

オススメ論文検索システム: OSUSUME

OSUSUME: Recommender System for Scholarly Papers

内山清子[*1] 高須敦宏[*1] 相澤彰子[*1] 難波英嗣[*2] 宮尾祐介[*1]
Kiyoko Uchiyama Atsuhiro Takasu Akiko Aizawa Hidetsugu Nanba Yusuke Miyao

[*1]国立情報学研究所 National Institutte of Informatics
[*2]広島市立大学 Hiroshima City University

We propose basic technologies and algorithms for a recommender system of scholarly papers based on multi-facets. The system is designed for beginners such as undergraduate students who have not decided on their research topics or researchers transitioning to work in a new field. Academic search is quite variable depending on users' preference. In order to reflect users' preference in recommendation, the system integrates multi-facets. This paper focuses on basic methods for two facets; internationality and essentiality. We introduce a cross-lingual keyword recommendation method as an international facet, which is built on an extended latent Dirichlet allocation model, for extracting latent features from parallel corpora. For an essential facet, we discuss a method for extracting basic and important terms and identifying their levels of essentiality based on occurrence and concatenation frequency.

1. はじめに

本研究では，論文を推薦するために複数の観点に基づく推薦手法について概観し，二つの観点における学習モデルや基礎的な手法を記述し，その有効性を議論する．本研究におけるシステムの利用者は，これから新しい学術分野の研究を進めようとする時に基礎知識を養うための必要な論文を探す学生などの初心者，新しい分野の概要を知りたい他の分野の研究者，特定の手法について実験，実装するために類似した過去の事例を参照したい専門家や企業のエンジニアなどを想定している．これらの利用者に対して，論文に含まれる単語やその周辺の共起情報を利用した多様な観点を設定したシステムを構築した[8]．システムフローは利用者がキーワードや名前を入力してログインする方法により，キーワードや過去に執筆した論文に関連した論文リストが表示される．そのリストの中からシードとなる論文を一つ選択し，そのシード論文に関連した論文を以下の 8 つのレコメンダ毎に推薦する．

- 国際論文レコメンダ：日本語の論文から海外ジャーナルの論文を推薦
- 類似論文レコメンダ：類似した論文を推薦
- 速報論文レコメンダ：最新の論文を推薦
- 発掘論文レコメンダ：異なる分野における類似した論文を推薦
- 対象論文レコメンダ：同じ対象を持つ論文を推薦
- 手法論文レコメンダ：論文と同じ手法を持つ論文を推薦
- 入門論文レコメンダ：関連した基礎技術を説明している論文を推薦
- 基礎論文レコメンダ：関連分野における引用回数の多い論文を推薦

この推薦システムを使った評価実験の結果，推薦論文の欠如（レコメンダが推薦論文を表示しない状態），推薦論文の質が良くないという意見があった．そこで各観点を再整理し，導入している手法の精度向上を検討した．特に，論文を推薦する際，テキストの類似度は文書に含まれる単語（その単語をここではキーワードと呼ぶ.）をベースにして計算を行うため，キーワードの抽出が重要な役割を果たす．そこで本研究では，8 つの観点の中から，国際論文レコメンダと入門論文レコメンダに焦点を当てて，論文の特徴抽出やキーワード抽出に関連する手法について述べる．この二つのレコメンダは評価実験においてユーザのお気に入りに選ばれている．その理由として初心者にとって英語論文を探すのが難しいことと，新しい分野において初期段階で読むべき論文がわからないなどレコメンダによる推薦の需要が多いと予測される．

国際論文レコメンダでは，日本語の論文をシードに選択しても，英語の論文を推薦可能であるが，推薦に利用しているのは，日本語と英語の抄録，タイトルである．しかし，対訳情報が揃っている論文が少ないため，論文の推薦精度が落ちてしまう．そのため，日英対訳情報がない論文に対しても適切な国際論文を推薦する仕組みが重要になってくる．本研究では，日本語と英語の対訳抄録をあらかじめ学習させておき，日本語抄録しかない論文から重要となる英語のキーワードを対訳辞書なしで抽出し，そのキーワードを基に論文を推薦する手法に用いられる学習モデルについて記述する．

入門論文レコメンダでは，初心者でも読みやすい論文や，解説論文などを推薦するために，特定分野において最初に学習すべき重要で基礎的な用語を設定している．基礎的な用語の度合いがわかっていれば，最初に理解しなければならない用語を解説する論文を読むことにより，学習効率が上がり，また初学者だけでなく学習者の理解の度合いに従って難易度順に論文の推薦も可能になると考えられる．本研究では，基礎性の高い語の抽出とランキングについて従来手法で実験を行い，問題点を分析する．

本論文の構成は，2 章において，学術論文の推薦システムについて述べ，3 章では，国際論文レコメンダと入門論文レコメンダの基礎になる手法を記述し，4 章でまとめと今後の課題を述べる．

連絡先: 内山清子，〒 101-8430 東京都千代田区一ツ橋 2-11-2, kiyoko@nii.ac.jp

2. 関連研究

推薦システムはユーザが興味を持ちそうなアイテムを予測する技術を応用しており，これまで協調フィルタリング，コンテンツベースフィルタリングのアプローチがとられてきた．協調フィルタリング手法は好みが似ているユーザが購入や閲覧の履歴パターンといった暗黙的な情報や，アイテムに対して直接的に評価値を与える明示的な情報を利用している．一方，コンテンツベース手法では，ユーザが好むアイテムの類似性といった属性を利用する．論文の推薦を目的とした場合，ユーザ（研究者）に対してアイテム（論文）を推薦するタスクとなる．

学術論文推薦システムの場合，研究者の閲覧・検索履歴のログを取ることが難しいことから，分野調査のために読んだ論文や，参考文献として掲載した論文，自分が過去に執筆した論文などが情報として使われる．論文推薦システムに利用可能なデータとして CiteULike[*1]がある．これは研究者が自分が参考にした文献を登録したり，キーワードなどのタグを付与して，共有することができるサービスである．CiteULike のデータを利用して協調フィルタリング手法で論文を推薦する手法 [1]，著者が過去に執筆した論文情報に基づいて類似した論文を推薦する手法 [6]，ハイブリッド手法（協調フィルタリングとコンテントベースフィルタリングを組み合わせる手法）を用いて修士や博士課程の学生が論文執筆のために読むべき論文リスト（reading list）を自動生成する方法 [3] などが提案されている．

本研究では，初心者を対象とした推薦論文の精度向上のために必要な論文特徴の抽出方法やキーワード抽出手法に焦点を置くこととした．これらの手法に基づいて，論文を推薦するアルゴリズムについては今後の課題と位置付ける．

3. レコメンダにおける手法

3.1 国際論文レコメンダ

3.1.1 概要

本節では，多言語の論文を推薦するための論文の類似度の計算法について述べる．多言語文書の処理には，機械翻訳や対訳辞書などを使って言語間の変換をした後に処理する方法と複数言語で記述された文書を同一の特徴空間にマップして処理する方法が考えられる．学術文献は新しい概念や技術を扱うことが多く，新しい用語が使われる頻度も通常の文書より多いことが予想される．このような文書に対しては作成コストの高い対訳辞書等を用いる手法よりも統計的な特性を用いて同一特徴空間にマップするほうが適していると考え，本稿では後者のアプローチによる日英 2 言語で書かれた論文を推薦するための論文の特徴の抽出法を提案する．

提案手法は，Latent Dirichlet Allocation (LDA) と同様に生成モデルに基づいた特徴を抽出する．まず，隠れトピックの集合 T を仮定し，各論文 $\boldsymbol{x}$ は，このトピック集合上の多項分布 $\boldsymbol{\theta}_{\boldsymbol{x}}$ として特徴づけられるとする．LDA は，文書を語の集合とみなし，文書ごとに割り当てられた多項分布 $\boldsymbol{\theta}$ に従ってトピック t を生成し，このトピックより多項分布 $\boldsymbol{\phi}_t$ に従って語を生成する．

論文は，タイトル，キーワード，概要，論文本文など，さまざまなテキストから構成される．また，日本語論文の場合には，タイトル，キーワード，概要については，日本語と英語の両方で記述されることも多い．以下では，タイトルやキーワードを**フィールド**と呼ぶ．本稿では，各テキストを単語の集合とみなし，論文を以下のタプルで表す．

$$\boldsymbol{x} := (\boldsymbol{x}_{j1}, \boldsymbol{x}_{j2}, \cdots, \boldsymbol{x}_{jn}, \boldsymbol{x}_{e1}, \boldsymbol{x}_{e2}, \cdots, \boldsymbol{x}_{en})$$

ここで，$\boldsymbol{x}_{jf}$ は，日本語で表された f 番目のフィールドを，また，$\boldsymbol{x}_{ef}$ は，英語で表された f 番目のフィールドを表す．本稿で提案する方法は，LDA と同様に各論文をトピック上の多項分布によって特徴づけるが，テキストの種類や記述言語ごとに異なった多項分布を用いてトピックから語を生成するところが LDA とは異なる．

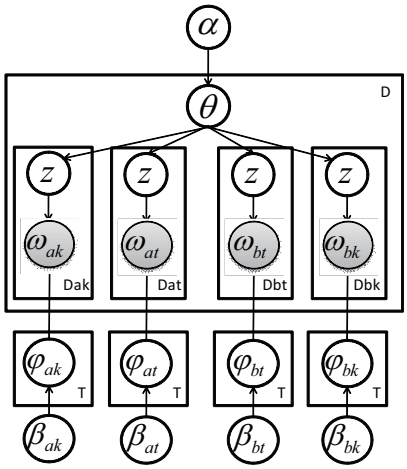

図 1: 提案手法で用いるグラフィカルモデル：D は訓練データに含まれる論文数を，D_{et}, D_{jt}, D_{ek}, D_{jk} は，それぞれ各言語 (a, b) と各テキストタイプ (t, k) に含まれ単語数を表す．T は，トピック数を表す．

モデルの学習フェーズでは，2 言語で書かれたタイトルや概要を用いて異なる言語の単語とトピックとの対応づけを行う．一方，論文推薦のフェーズでは，単一言語で論文がかかれていても，単語とトピックの関係を用いて，論文の特徴をトピック集合の多項分布として特徴づけ，論文間の類似性を確率分布の類似度に基づいて求める．

3.1.2 モデル

以下に提案手法で用いるモデルを示す．$Multi(\boldsymbol{\theta})$ および $Dir(\boldsymbol{\alpha})$ は，それぞれ，パラメタ $\boldsymbol{\theta}$ および $\boldsymbol{\alpha}$ の多項分布とディリクレイ分布を表すものとする．

トピックの集合 T, トピック生成確率の事前ディリクレイ分布のパラメタ $\boldsymbol{\alpha}$, 単語生成確率の事前ディリクレイ分布のパラメタ $\boldsymbol{\beta}_{lit}$ $(l \in \{j, e\}, 1 \leq f \leq n)$, に対して，以下の手順で論文の集合 $\boldsymbol{D}$ を生成する

1. 各言語 $l \in \{j, e\}$, 各フィールド f, 各トピック $t \in T$ に対し単語生成確率を生成: $\boldsymbol{\phi}_{lft} \sim Dir(\boldsymbol{\beta}_{lft})$
2. 各論文 $\boldsymbol{x} \in \boldsymbol{D}$ について
   (a) $\boldsymbol{\theta}_{\boldsymbol{x}} \sim Dir(\boldsymbol{\alpha})$
   (b) 各言語 l，各フィールド f の各単語 x_{lfj} について
      i. トピックを生成: $t \sim Multi(\boldsymbol{\theta}_{\boldsymbol{x}})$,
      ii. 単語を生成: $x_{lfj} \sim Multi(\boldsymbol{\phi}_{lft})$

図 1 は，2 つのテキストタイプ $\{k, t\}$ に対するグラフィカルモデルを表している．

*1 http://mecab.sourceforge.net/

3.1.3 パラメタ推定

本稿では，ギブスサンプリングを用いてパラメタを推定する．ギブスサンプリングでは，各語にトピックを一定の確率分布に従って繰り返し割り当てる．紙面の制約のため導出は省略するが，提案モデルにおいて，論文 $\boldsymbol{x}$ の言語 l, フィールド f の j 番目の単語に対するトピック割り当ての確率分布は以下の式で与えられる．

$$\Pr(Z_{lfj} = \hat{t}) \propto (\alpha_{\hat{t}} + N_{\boldsymbol{x}\hat{t}}^{\boldsymbol{Y}^{-lfj)}}) \frac{(\beta_{fx_{lfj}} + N_{\hat{t}x_{lfj}}^{\boldsymbol{Y}^{-lfj}})}{\sum_{w \in W_{lf}} (\beta_{lfw} + N_{\hat{t}w}^{\boldsymbol{Y}^{-lfj}})} \quad (1)$$

ここで，式中の記号は以下のように定義される．

- $N_{\boldsymbol{x}\hat{t}}^{\boldsymbol{Y}^{-lfj}}$: 論文 $\boldsymbol{x}$ 中の着目している単語 x_{lfj} 以外の単語に対して トピック $\hat{t}$ が割り当てられている回数
- $N_{\hat{t}w}^{\boldsymbol{Y}^{-lfj}}$: 着目している単語 x_{lfj} を除き，単語 w にトピック $\hat{t}$ が割り当てられている回数

ギブスサンプリングの結果，各論文のトピック生成確率は以下の式で与えられ，

$$\theta_{\boldsymbol{x}t} \propto (\alpha_t + N_{\boldsymbol{x}t}^{\boldsymbol{Y}}) \quad (2)$$

単語生成確率は以下の式で与えられる．

$$\phi_{lftw} = \frac{\beta_{lfw} + N_{tw}^{\boldsymbol{Y}}}{\sum_{v \in W_{lf}} (\beta_{lfv} + N_{tv}^{\boldsymbol{Y}})} \quad (3)$$

3.1.4 論文の特徴抽出

単一の言語で記述された論文が与えられた場合，(3) 式の単語の生成確率を固定してギブスサンプリングを行い，(2) 式で表されるトピック生成確率を求める．この分布を当該論文の特徴ベクトルとして用いる．トピックと単語の関係は，(3) 式に示される単語生成確率として求められているため，この確率分布より論文に含まれる単語から対応するトピックの分布を求める．論文間の類似度には，トピック分布の KL-divergence に加え，分布を特徴ベクトルと見なしてコサイン尺度やユークリッド距離をもちいることも考えられる．文献 [7] では，このモデルを論文に付与されるキーワードを推定する問題に適用し，日本語の概要から英語のキーワードを付与したり，逆に英語の概要から日本語のキーワードを付与する性能を調べたところ，同一言語間での推定とほぼ同じ精度で 2 言語間での推定が行えることが実験的に確かめられた．

3.2 入門論文レコメンダ

3.2.1 概要

本節では，分野基礎性が高い用語を抽出し，その度合いを識別する方法について述べる．入門論文レコメンダは，これからその分野を学ぼうとする初学者（学部学生から他の分野の研究者など）に，分野の概要が分かるような解説論文や，理解しておくべき基礎的技術を説明している論文などを推薦することを目指したレコメンダである．入門論文を推薦するために，初学者が理解しなければならない基礎的な用語を抽出することが重要となる．その用語の理解がなければ論文を読み進めることができない重要で基礎的な用語のことを分野基礎性が高い用語と定義する．分野基礎性の高い用語は優先度によってレベル分けをする必要がある．用語基礎性が高い順にレベル分けされていれば，効率的に学習を進められる．

入門レコメンダに必要な要素技術として，(1) 分野基礎性が高い語の抽出方法，(2) 分野基礎性の度合いを識別する方法，の二種類が必要となる．分野基礎性が高い語の抽出は従来行われてきた専門用語抽出の手法が有効であるかを確認する．度合いを識別する方法は，先行研究において対象となる学習者を設定し，それに対応する用語のレベル分けを提案したが [9]，ここでは (1) の抽出方法において，上位に順位づけられている語が分野基礎性の度合いが高いものであるかを調べた．

論文において分野基礎性が高い語は，経年推移性，網羅性，語構成性から考えることができる．まず，経年推移性の観点において，分野基礎性が高い語は，論文中の出現パターンが比較的安定して，長い期間出現しつづけると考えられる．たとえば，自然言語処理分野で基礎性が高い用語の「形態素解析」は，その分野の論文が出版された時から出現し，初期の段階では手法に関する研究が多かったが，現在はすでにツールとして広く用いられているため，継続して論文中に記述される．

網羅性の観点においては，特定の論文に多く出現するような用語ではなく，平均的に論文中で用いられる用語や，事典の場合には，複数の異なる章にまたがって出現するような用語が分野基礎性が高い用語と考えられる．

語構成性とは，多くの複合語を生成する性質のことで，新しい複合語の基になる用語は基礎性が高いと考えられる．前述の「形態素解析」という用語では「日本語形態素解析」「形態素解析システム」など前後に多くの用語が接続して複合語を生成したり，文章中に「形態素解析中」や「形態素解析失敗」など接尾辞や文章を短縮したような表現の臨時一語を生成する傾向にある．専門用語の前後に連接する頻度や異なり語数に基づいた用語抽出手法も提案されており [4]，分野基礎性を測る上で語構成性は重要な観点である．本節では，この語構成性に絞って用語の抽出を行う．

3.2.2 属性の設定

用語抽出を行う際，論理構造における出現パターンを利用した方法 [5] が有効であったことから，本節でも論理構造を考慮した．論理構造とは「タイトル」「著者」「抄録」などの書誌情報や，本文中の「関連研究」「実験」「考察」「まとめ」など論文の詳細な構造のことを指す．本文テキストの入手が困難であるため，今回はタイトル，抄録，著者キーワードに絞った．

対象複合語 (CN) について，先行研究 [4] にも使用されている語構成性に関する属性を 7 つ設定し，頻度は論理構造として「タイトル (*title*)」「抄録 (*abst*)」「著者キーワード (*kw*)」における出現回数と，これらすべて「合計 (*total*)」の出現回数に分けて算出した．

FL：CN の左側連接頻度
FR：CN の右側連接頻度
f(CN)：CN の単独出現頻度
n(CN)：コーパスにおける CN の出現頻度
t(CN)：CN を含むより長い複合名詞の出現頻度
c(CN)：CN を含むより長い複合名詞の異なり語数
length(CN)：CN の構成単名詞数

3.2.3 分野基礎性が高い語の抽出

実験に使用したデータは，情報処理学会自然言語処理研究会 14 年分 1993 年から 2006 年までの 1421 論文の書誌情報（タイトル，抄録，キーワード）を対象として分野基礎性が高い語の抽出を行った．デジタル言語処理学事典 [10] の索引語と論文の著者キーワードの中から，自然言語処理の専門家が特に分野基礎性が高いと判断した 500 語を 4 段階に分類したものを正解データとして利用した．対象となる複合名詞 CN は，

MeCab[*2]で形態素解析をし，名詞が連続している単語列を抽出し，各属性の頻度を計算した．評価方法として，分野基礎性が高い語の抽出精度については，各属性の単独頻度と，これらの属性値を利用した用語スコア付けの FLR 法, C-Value 法 [2]，MC-Value 法 [4] を用いて上位 100 語 (@100)，300 語 (@300)，500 語 (@500) の順位を求め，それぞれ F 値を算出して評価を行った．C-Value は次式で定義される．

$$C\text{-}Value(CN) = (length(CN) - 1) \times (n(CN) - \frac{t(CN)}{c(CN)}) \quad (4)$$

表 1 に単独属性と用語スコア付けで最も精度が良かった単独出現頻度 *f(CN)* と C-Value 法の値を示す．全体的に精度が低かった理由として，従来の専門用語抽出と異なり，正解セットを基礎性が高い用語に絞ったため，用語として適切であっても正解と見なさなかったことが考えられる．

また，C-Value と MC-Value は語構成の特徴を利用したスコア付けであったが，二字漢語の場合，多くの専門用語や複合語の語基となっているため頻度が高く，語構成要素数を考慮してもスコアが高くなってしまう傾向にある．C-Value は，*length(CN)* から 1 を引くことにより，1 つの構成要素からなる用語を上位にランキングすることを抑えることができ，本研究の分野基礎性において MC-Value よりも概ね有効であった．今後は分野基礎性が高い用語に対しては，各属性値に文字数や語構成数を総合的に組み合わせて拡張する必要がある．

次に，論理構造別の精度では，特に目立った差がなかった．著者キーワードが有効であることは確認できたが，全ての論文にキーワードが付与されているわけではないため，書誌情報全体の頻度を扱うことが妥当だと考える．将来的には，書誌情報以外の「はじめに」「関連研究」「実験」「考察」「おわりに」等の論理構造毎の出現傾向も分析する予定である．

表 1: 抽出された完全一致用語における F-値

	@100	@300	@500
kw_f(CN)	0.1906	0.3269	**0.3562**
kw_C-Value	0.1352	0.2085	0.2047
abst_f(CN)	0.1282	0.2059	0.2497
abst_C-Value	0.1490	0.2342	0.2313
title_f(CN)	0.1248	0.2625	**0.2886**
title_C-Value	0.1456	0.2085	0.2416
total_f(CN)	0.1456	0.2342	0.2805
total_C-Value	0.1560	0.2703	**0.2948**

3.2.4 分野基礎性の度合いの識別

分野基礎性の度合いを識別する方法については検討段階のため，今回は正しく抽出できた用語について分野基礎性が高い語の抽出方法を利用して，4 段階の分野基礎レベル順にランキングされているかを調べた．表 2 は *total*（タイトル，著者キーワード，抄録の合計頻度）に基づいて，専門家によって 4 段階（基礎性が最も高いものが 1）に分類された各レベルにおける用語数（正解用語数），C-Value と単独出現頻度 (*f(CN)*) によってランキングされた用語の各レベルにおける平均順位を示している．結果として，単独出現頻度 *f(CN)* の方が，分野基礎性が高い語が上位に出現しており，分野基礎レベル順に用語がランキングされていることがわかる．C-Value は抽出精度は *f(CN)* よりも高かったが，基礎性が高い語が上位に位置していなかった．今後は分野基礎性が高い語を中心としたネットワーク構造等で関連語を表示できる仕組みを検討していく．

*2 http://mecab.sourceforge.net/

表 2: 分野基礎レベルにおける平均順位

分野基礎性 (正解用語数)	C-Value	*f(CN)*
1 (30)	113	63
2 (244)	218	221
3 (172)	277	283
4 (30)	316	314

4. おわりに

本研究では，論文を推薦するための 2 つのレコメンダにおいて重要となる学習モデルや基礎的な手法を提案し，詳細を記述した．国際論文レコメンダでは，複数言語で記述された文書を同一の特徴空間にマップして，日英 2 言語で書かれた論文を推薦するための特徴抽出法を提案した．入門論文レコメンダでは，分野基礎性が高い語を従来手法で抽出し，そのランキングを利用して分野基礎性の度合いを識別する方法を分析した．今後はこれらの手法を論文の推薦に適用し，評価をすることが課題である．

参考文献

[1] T. Bogers and A. van den Bosch. Recommending scientific articles using citeulike. In *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 287–290, 2008.

[2] K. Frantzi and S. Ananiadou. Extracting nested collocations. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96)*, pp. 41–46, 1996.

[3] J. A. Stemper J. T. butler J. A. Konstan M. D. Ekstrand, P. Kannnan and J. T. Riedl. Automatically building research reading lists. In *Proceedings of the 4th ACM conference on Recommender systems*, pp. 159–166, 2010.

[4] 中川裕志, 森辰則, 湯本紘彰. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, pp. 27–45, 2003.

[5] T. H. Nguyen and M. T. Luong. Wingnus: Keyphrase extraction utilizing document logical structure. In *Proceedings of SemEval-2010 Task5: Keyphrase Extraction from Scientific Articles*, 2010.

[6] K. Sugiyama and M. Y. Kan. Scholarly paper recommendation via user's recent research interests. In *Proceedings of the 10th annual joint conference on Digital Libraries*, pp. 29–38, 2010.

[7] A. Takasu. Cross-lingual keyword recommendation using latent topics. In *Proceedings of International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2010)*, pp. 52–56, 2010.

[8] K. Uchiyama, H. Nanba, A. Aizawa, and T. Sagara. Osusume: Cross-lingual recommender system for research papers. In *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation*, pp. 39–42, 2011.

[9] 内山清子．専門用語の分野基礎性を判定する基準に関する一考察．情報処理学会自然言語処理研究会, Vol. NL-199(15), pp. 1–6, 2010.

[10] 自然言語処理学会（編）．デジタル言語処理学事典. 共立出版株式会社, 2010.