DEPARTMENT OF MATHEMATICAL INFORMATICS
THE UNIVERSITY OF TOKYO

特許の質の予測モデリング
～機械学習とテキストマイニングによるアプローチ～

鹿島 久嗣, 柴田 直樹, 坂田 一郎, 渡部 俊也
（東京大学）
比戸 将平, 坪井 祐太, 田島 玲
（IBM東京基礎研究所）
上野 剛史
（日本IBM）

背景: 企業にとって特許の価値は極めて重要であるが、社会全体の利益という観点からは必ずしもそうではない場合がある

- ビジネスにおいて自社のもつ特許の価値を見極めることは重要:
  1. 技術的価値 (パイオニア特許か、改善による特許か)
  2. 法的価値 (特許が認められるか)
  3. 経済的価値 (キャッシュフローを生み出すか)
- 特許の価値をモデル化し、評価を行おうという試みが数多く存在する
- しかし、特定の企業からの観点からみた価値は、社会的なコストを増大し、イノベーションを阻害する可能性がある
  - 広く曖昧な請求項で中身の無い特許は、将来の訴訟を生み出す
  - パテントトローリングによるイノベーションの阻害

背景: 特許の質とは特許の「社会全体にとっての好ましさ」を表す概念だが、その定量化はまだ未知

- 特許の質とは（特許の価値とは異なり）特許の公共的／協調的側面を考慮した概念
- 特許の質は、発明者や利用者、特許庁など、特許を取り巻く社会全体からみた好ましさを測る指標として定義される
- 特許庁による審査水準の向上などの行政による施策に加えて、特許の質の定量的な指標を共有できれば、質の高い特許を生み出すコミュニティの促進につながると期待される
- 「特許の質」をどのように定量化するべきか？

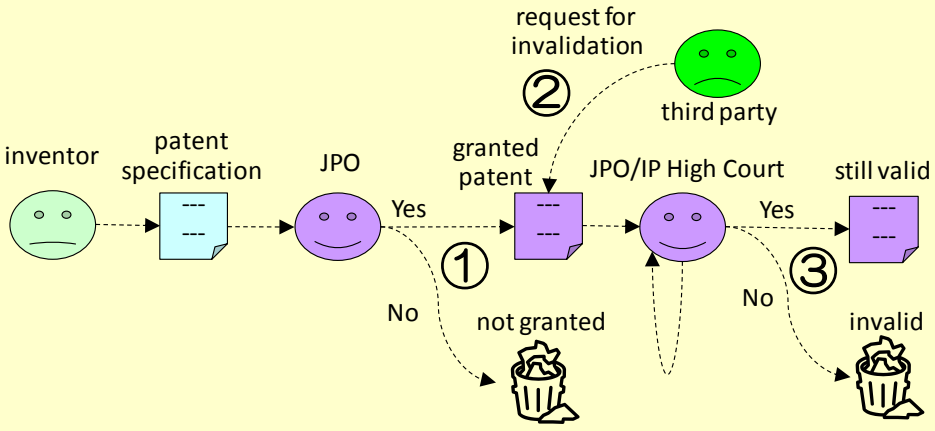
これまでの研究: 永田らは、特許の質＝法的有効性 と捉え、そのモデル化を試みた

- 永田らは、法的な有効性を特許の質の一指標と捉え、これをモデル化する試みを行った
  - 訴訟に対して頑健である特許は、適切な記述／権利主張／審査がなされている可能性が高く、ひいては社会的コストの低減につながる
- 裁判所による特許の有効無効の結論を、いくつかの説明変数から説明する回帰モデルを構築した
  - データは、知財高裁によって判決がなされた710件
    - 2000年1月1日から2006年12月31日までの期間に東京高等裁判所で判決がなされた審決等取消訴訟のうち、特許異議申立てと特許無効審判が原審となっている案件

Nagata, K, M Shima, N Ono, T Kuboyama and T Watanabe (2008). Empirical Analysis of Japan Patent Quality, In *Proc. 17th international conference on Management of Technology, the International Association for Management of Technology (IAMOT).*

参考：日本における特許システムのフローチャート（簡略版）

- 永田らは③のモデル化を行った



参考：「特許の質スコア」の回帰モデル

- 特許は、説明変数の集合として表される
- 各説明変数の「特許の質スコア」への貢献度合いがパラメータ
- 有効な特許ほどスコア $f(x)$ の値が高くなるようにする

patent quality score

features of a patent

$$f(x) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

a patent

model parameters

参考：永田らの用いた説明変数（抜粋）

Action	Parameters	Definition
Applicant	Domestic_P	Number of Domestic Priorities
	Paris_P	Number of Priorities under the Paris Convention
	App_JPR	Number of Japan Patent References disclosed in a patent application by applicants
	App_FPR	Number of Foreign Patent References disclosed in a patent application by applicants
	Inventors	Number of Inventors
	Applicants	Number of Applicants
Agent	Claims	Number of Claims
	Claims_I	Number of Independent Claims
	Claims_D	Number of Dependent Claims
	Claims_C	Number of kinds of Claims Categories (e.g. "method claim" "product claim" "system claim")
	Words_AC	Number of words in All Claims (0.1 times)
	Words_TC	Number of words in Claim 1 (Top Claim)(0.1 times)
	Words_DE	Number of words in the item Detailed Explanation of invention (0.01 times)
	Words_DEEBA	Number of words in the rest except Background Art from "Words_DE" (0.01 times)
	Effects	Number of words describing "effect" in "Words_DEEBA" (e.g. "may/can" "could" "superior" "useful" "advantageous")
	Arguments	Number of filing arguments
	Amendments	Number of filing amendments
	Request_OI	Number of requests for oral Interview with examiner
Examiner	Exa_JPR	Number of Japan Patent References cited by the examiner
	Exa_FPR	Number of Foreign Patent References cited by the examiner
	Exa_NPR	Number of Non Patent References cited by the examiner
	Exa_AJPR	Number of Japan Patent References added by the examiner
	Exa_AFPR	Number of Foreign Patent References added by the examiner
	Rejection_W	Number of rejection notices for lack of written description
	Rejection_N	Number of rejection notices for lack of novelty/inventive step
	Rejection_O	Number of rejection notices for other reasons

優先権主張制度の利用

"効果語"の利用

外国文献引用数

本研究の目的: 永田らの研究を、予測という面から捉えなおす

- 永田らの研究では、どの説明変数が効くか、という記述的なモデル化に主眼
- 一方、指標として用いるためには、ある程度の予測精度が必要
  - 出願／維持すべき特許の選定などにも有効
- 予測精度の向上を主眼においた「予測的モデル化」を行う

本研究の成果: 機械学習／テキストマイニングの手法を用いて審判結果の予測精度を向上した

- 機械学習における予測精度向上のテクニックを用いて、予測精度を向上する
  - サポートベクトルマシン（SVM）：予測精度の高いモデル
  - クラス比重みづけ：偏りのあるデータへの対処法
  - L1-正則化：高次元データへの対処法
- テキストマイニングによって、特許文書の文書表現から予測に有効な特徴をモデルに取り入れる
- さらに、これらを組み合わせることで、一層の予測精度向上を試みる

予測精度評価の方法：交差検定と、2つの評価指標（AUCとBEP）

- 交差検定によって未知の特許に対する予測精度を擬似的に検証する
  - 全体の 80% のデータをモデル化に用いる
  - 残りの 20% を（審判の結果を隠して）予測精度の評価に用いる
- 2つの予測精度の評価指標（AUCとBEP）を用いる
  - AUC (Area Under the ROC Curve)
    - 予測の順序付けの正しさを評価する
    - 金融分野で倒産判別モデルの評価指標として定められているAR値(Accuracy Rate)と等価
  - BEP(Break-Even Point)
    - 最適な閾値で判断したときの予測正解率
    - テキスト分類精度の評価に用いられる指標

検証１：
多くの説明変数を使った方が予測精度は上がるだろうか？

- 永田らは、60個の説明変数から、有望な24個を選びモデルを構築していたが、全て使うことで予測精度は上がるだろうか？
- 機械学習においては、使えそうな説明変数は全て放りこんで、正則化という枠組みで解決するのが一般的
  - 正則化：モデルパラメータ$(w_1, w_2, \ldots, w_d)$ が極端な（絶対値の大きい）値を取らないようにバイアスをかける

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

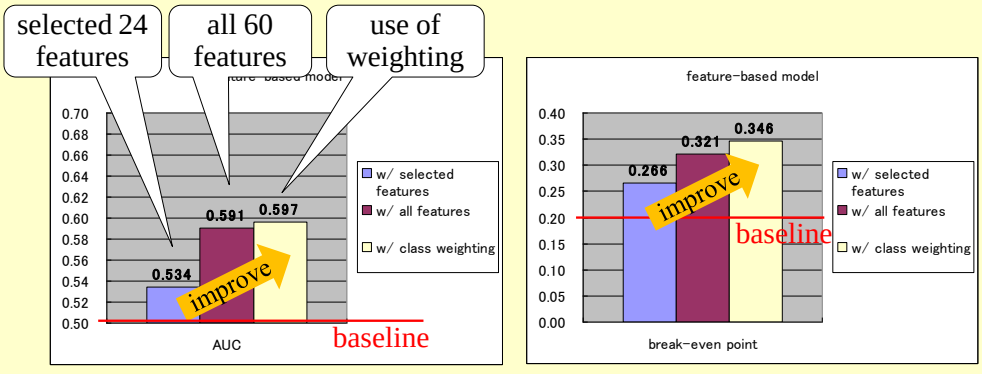
    - 具体的には、$\|w\|_2^2 := w_1^2 + w_2^2 + \ldots + w_d^2$ を小さくするようにする

検証２：
データの偏りに対処することで予測精度は上げられるだろうか？

- データ中、有効であると判断された特許は全体の20%にすぎない
  - 無効なケースは有効なケースの4倍ある
- この偏りの情報を推定に有効に用いることができるだろうか？
- 直感的には、少ない方（有効である特許）を重視するようにモデルを推定するのがよさそう
  - 有効なケースを4倍の重みをつけてモデル推定を行う
  - これは「クラス比重みづけ」と呼ばれ、偏りのあるデータの場合に予測精度を改善することが知られている

データを用いた検証結果１＆２：
機械学習のテクニックを用いることで予測精度は改善する

- 予測精度に定評のあるサポートベクトルマシンを用いた予測で
  - 全て(60個)の説明変数を用いた方が予測精度は良い
  - クラス比重みづけによって予測精度はさらに向上



AUC　　BEP

検証３：
テキスト情報を用いることで予測精度は上げられるだろうか？

- 特許には明細書などの文書情報があるので、これらを網羅的に用いることで予測精度が上げられだろうか？
- 特許文書から網羅的に説明変数を生成
  - 日本語の文書は「分かち書き」されていないため、テキストマイニングで用いられる形態素解析を行って単語を取りだす

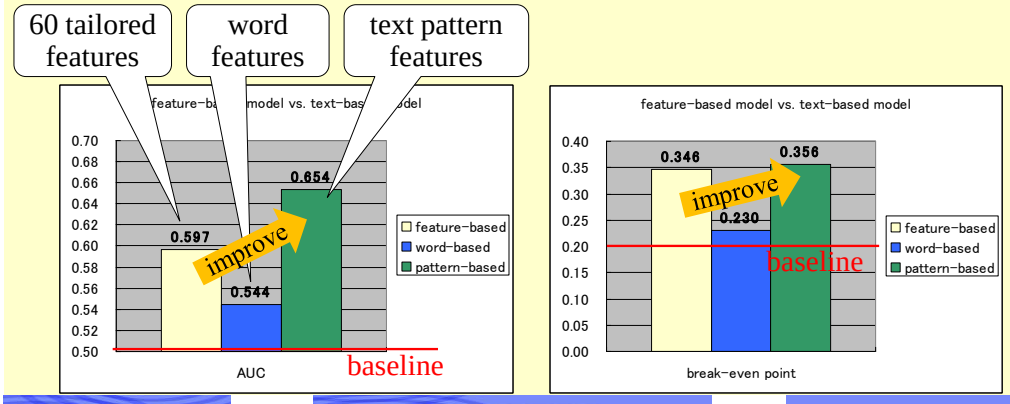
醸造に要する時間を大幅に短縮することができる
we can significantly shorten the time for brewing

醸造	に	要する	時間	を	大幅	に	短縮	する	こと	が	できる
[noun]	[particle]	[verb]	[noun]	[particle]	[noun]	[particle]	[noun]	[verb]	[noun]	[particle]	[verb]

  - データ全体に20回以上現れる単語(2,400語)および2-3語の連続した表現(13,000表現)を抽出
  - それぞれが特許文書中に現れるなら1、そうでないなら0として説明変数を定義
- 「データの数＜説明変数の数」の問題を「L1正則化」によって解決
  - L1正則化はモデルで用いられる変数を大幅に削減する効果がある
  - $|w|_1 := |w_1| + |w_2| + \ldots + |w_d|$ を小さくするようにする

データを用いた検証結果３：テキスト情報には予測力がある

- 2種類のテキスト情報を用いたモデルを比較
  - 単語（2400語）を用いたモデル
  - 2-3語の表現(13,000表現)を用いたモデル
- 2-3語の表現を用いて永田らの説明変数を上回る予測精度を達成



15 AUC BEP NIVERSITY OF TOKYO

データを用いた検証結果３：
特許の質につながるテキスト表現が発見される

- モデル中で高いスコアをもつ説明変数を調べると
  - 請求項の範囲を明確化／制限する表現
  - 効果を表す表現 (永田らの示唆を支持)

  が自動的に発見された

	interpretations	patterns (in Japanese)	meanings of the patterns
clarifying or limiting coverage of claims	parameters	度合い[noun]-を[particle]	degree of …
		確率[noun]-の[particle]	probability of …
		の[particle]-設定[noun]	setting of …
	extension of existing patents	(実施)形態[noun]-による[particle], で[particle]-用い[verb]-て[particle]	executed in the condition of …
		に[particle]-置き換え[verb]	substitute … with …
		薄型[noun]-化[noun]	reduce the thickness of …
	effect representations	を[particle]-良く[adjective]	well
		正しい[adjective]	correct
		可撓性[noun]	flexibility
		利点[noun], 利点[noun]-を[particle]	advantage
		調整[noun]-可能[noun]	adjustable

16 THE UNIVERSITY OF TOKYO

検証４：もとの説明変数とテキスト情報を組み合わせることで予測精度は向上するだろうか？

- もともとのモデルとテキストベースのモデルを組み合わせる
- 2種類の組み合わせかたを考える
  - 協調モデル：2つのモデルのスコアを足し合わせる

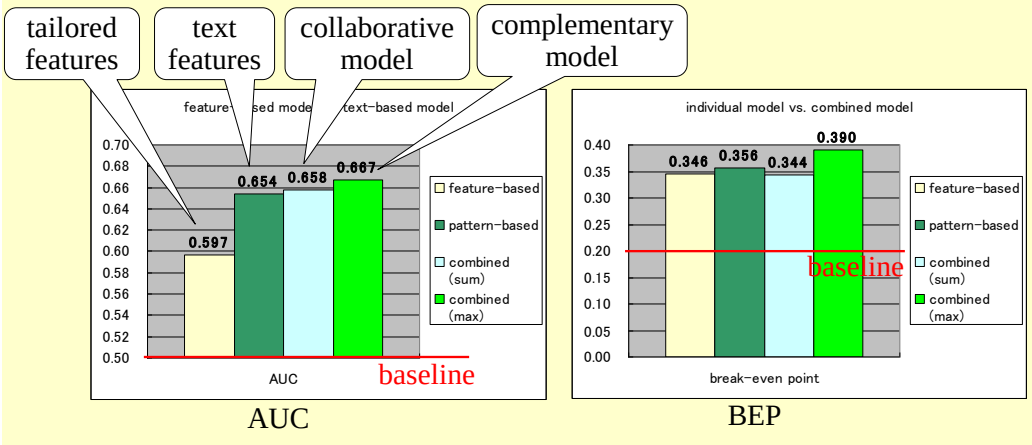
$$f^{tailored}(x) + f^{text}(x)$$

  - 相補モデル：2つのモデルのスコアの大きい方をとる

$$\max\{\ f^{tailored}(x)\ ,\ f^{text}(x)\ \}$$

データを用いた検証結果３：テキスト情報は相補的に働き予測精度を向上させる

- もともとのモデルとテキストベースのモデルは相補的に働く
  - 相補モデル（大きいほうを取る）のが予測精度がよい



AUC　　BEP

結論：機械学習とテキストマイニングによる「特許の質」の予測モデル化を行った

- 特定企業のための「特許の価値」ではなく、社会的な好ましさである「特許の質」を、予測という観点からモデル化を試みた
- 機械学習の手法を用いて予測精度が向上できることを確認した
- 特許文書から網羅的に抽出したテキスト情報には予測力があり、人手で構築した説明変数と相補的に働くことを示した
- 今後の課題としては：
  - より大規模なデータを用いたより高精度なモデル化
  - 他の指標に注目した特許の質のモデル化