# Eディスカバリ効率化のための機械学習アルゴリズムの提案

三戸 智浩† 佐々木 良一‡

†東京電機大学
120-8551 東京都足立区千住旭町 5 番
mito@isl.im.dendai.ac.jp, sasaki@im.dendai.ac.jp

**あらまし** 近年，米国に進出した日本企業の間では”E ディスカバリ”が企業経営における大きなリスクであると認知され始めている．これは訴訟当事者同士が訴訟に関連する全ての電子データを開示する米国の民事訴訟の制度である．E ディスカバリが要求されると，社内に存在するすべての電子データから訴訟に関係する電子データを裁判所に提出しなくてはならない．しかし，企業内に存在する電子データ量は膨大であり，訴訟に関係するデータを漏れなく抽出するのは弁護士への大きな負担となる．そこで本研究では機械学習と自然言語処理の技術を利用し，E ディスカバリを効率化するためのアルゴリズムを提案する．

# Proposal of a machine learning algorithm for effective E-discovery

Tomohiro Mito† Ryoichi Sasaki‡

†Tokyo Denki University.
Senju-Asahi-cho, Adachi-ku,Tokyo, 120-8551 JAPAN
mito@isl.im.dendai.ac.jp, sasaki@im.dendai.ac.jp

**Abstract** Japanese companies that expanded one’ business to the United States, ”E-discovery” is beginning to be recognized as high risk in corporate management. The E-Discovery is required in a civil litigation system of the United States. When the E-discovery is required, the pertinent company must be submitted electronic data related to litigation to court. However, electronic data amount stored in the company is extremely large. Therefore, the procedure to extract data related to litigation is a major burdenon to the lawyer. Accordingly, we propose an algorithm for efficient E-discovery using a technology of machine learning and natural language processing.

## 1 はじめに

米国では民事訴訟の際，原告と被告が開示する証拠書類は電子化されている．この電子化された書類を裁判で開示することを E ディスカバリという．

E ディスカバリでは原告と被告は，内部情報も含めて，訴訟に関連した証拠の全面的な開示を相手に要求できる．証拠を持っているのに隠しているとみなされると、制裁の対象になる．

実際の例として，2007 年にパソコン用メモリの特許侵害で訴えられていた東芝米国法人は E ディスカバリを要求されたが，関連するソフトウエアのソースコードを故意に隠したことで，米国の地方裁判所から弁論時間の大幅削減といった，敗訴に直結しかねない厳しい制裁命令を受けることとなった．

こうした背景からEディスカバリの重要性が高まってきている．

しかし，開示される証拠書類は膨大であり，弁護士が裁判で必要となる証拠書類を全て抽出するのは大きな負担となる．逆に抽出作業が不十分だとその後の証拠調べに手間を要して弁護士費用がかさんでしまう．さらに，必要以上に情報を出してしまうと機密情報を訴訟相手に漏らすことにもなりかねない．そのため，証拠書類の分類の効率化が裁判の判決を左右している．

また，海外に進出した日本企業も戦略的な特許訴訟などにより，Eディスカバリの対象となることが多くなってきている．これに多くの日本企業が苦慮している．

その理由として，日本ではEディスカバリを効率化させる手法が普及していないこと，また存在するEディスカバリ効率化のためのツールは，言語の誤認などが原因で日本語に適切な対応していないことが挙げられる．

以上のことから，電子化された日本語の証拠書類を効率よく分類するための機械学習アルゴリズムを提案する．

# 2　提案手法

## 2.1　提案手法の概要

米国ではEディスカバリを効率化させるための手法として, Predictive Cordingの利用が主流となっている．これは収集した大量のデータ（ドキュメント・電子メールなど）のうち，コンピュータが特定した一部のサンプルデータを人間がレビューし，重要単語への重み付けを行い，その結果に基づきコンピュータが残りのドキュメントをコード付けするというものである．

Predictive Cordingの流れは以下の通りである．

1. 全文書のうちの一部をコンピュータでランダムに抽出する
2. 限定されたサンプルのドキュメントを人間がレビューし，コード付を行う
3. 2のコード付の結果を予測コード付の技術を用いて全文書に適用する

実際に人間レビューアのレビューよりも，Predictive Cordingを用いたレビューの方が，精度が高い結果が得られた場合も確認されている [1].

しかしPredictive Cordingでは，単語へ重み付けを行い，文書のスコアを算出し，そのスコアを元に証拠文書かどうかを判定しており，機械学習アルゴリズムなどで厳密に分類しているとは考えられにくい．

また，文書抽出を行う機械学習アルゴリズムにおいても，word2vecを利用した手法 [2] や大規模コーパスを利用した文書分類が高い性能を示しているが [3]，これらは大量のラベル付き文書が既に存在するという前提がある．しかし，実際のEディスカバリの現場では，ラベル付けを行うのは弁護士で，企業が開示する膨大なデータ全てを手作業でラベル付けしていくのは現実的ではない．

そこで本研究は，文書へのラベル付けと証拠文書の分類に機械学習アルゴリズムを用いることで，弁護士への負担低減と証拠文書の検出率を向上させることを目的とする．

## 2.2　システムの動作の流れ

提案手法は以下の2つのフェーズで行う．

- フェーズ1：抽出された証拠データおよびデータプールに対して特徴選択と特徴変換，スムージングを行う．
- フェーズ2：特徴変換されたデータ群に対して教師あり学習とActive Learningを行う．

フェーズ1を図1に，フェーズ2を図2に示す

## 2.3　システムの利用の流れ

(1) 教師データの作成（人による分別）

調査対象の文書群から，弁護士がサンプルドキュメントをランダムにピックアップし，それが調査対象として重要なドキュメントであるか否かを判定する．

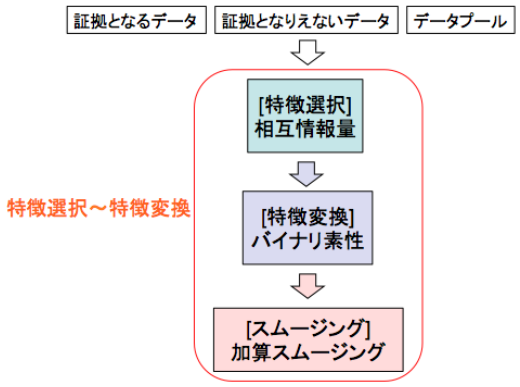


図 1: フェーズ 1：特徴選択，特徴変換

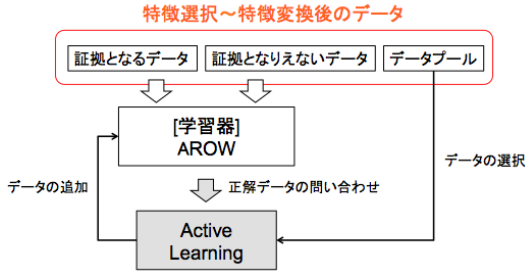


図 2: フェーズ 2：学習，Active Learning

また，調査対象の文書群の一部をデータプールとして利用する．

(2) ラベル付き文書の作成

(1) にてピックアップされた文書に付与されたラベルに基づき，データプールの残り文書全てにラベルを付与する．

(3) 教師データをもとに分類器の作成

ラベリングされたデータを対象に機械学習アルゴリズムに学習させ，分類器を生成する．

(4) 証拠文書の抽出

分類器が学習を元に、未査定の文書群から証拠となる重要なドキュメントを自動的に抽出する．

## 2.4 各種アルゴリズムの解説と採用理由

(1) 相互情報量

単語の重要度を算出し，素性の特徴選択に利用する．

相互情報量とは，2 つの確率変数の相互依存の尺度を表す量とのことである．テキスト分類の特徴選択で用いる場合は，ある単語 t の出現を表す確率変数 U とあるカテゴリの出現を表す確率変数 C を用いて相互情報量を定義する．U は 1 または 0 の値をとり、U=1 のとき単語 t が出現する事象、U=0 のとき単語 t が出現しないという事象を表す．C も 1 または 0 の値をとり，C=1 のときカテゴリが c である，C=0 のときカテゴリが c でないという事象を表す．相互情報量の定義は以下の通りである．

$$I(U;C) = \sum\sum P(e_t, e_c) \log 2 \frac{P(e_t, e_c)}{P(e_t)P(e_c)}$$

t は term，c は category の略で具体的な単語やカテゴリが入る．同時分布や周辺分布が出てくるが，クロス表を用いることで容易に導出できる．

本研究では「重要文書」と「文書中の単語」の関連と,「特定の単語の有無」と「重要文書か否か」を全て考慮した上でクロス表を生成する．クロス表を 1 に示す．

| | 特定の単語がある | 特定の単語がない |
|---|---|---|
| 証拠書類である | N11 | N10 |
| 証拠書類でない | N01 | N00 |

表 1: クロス表

このようなクロス表が求まれば、すべての同時確率と周辺確率が計算できる．相互情報量は 0 以上の値をとり，値が大きいほどカテゴリの特徴を表すような単語と見なすことができる．

相互情報量は，重要語句の組み合わせを網羅的に考慮できるため，証拠文書の検出率の向上が期待できる．

(2) 加算スムージング

テストデータを，あるクラスに属するかどうか判定するとき，そのクラスでは一度も出現したことの無い単語が現れた場合，確率が 0 になってしまう．これをゼロ頻度問題という．これを解決するのがスムージングであり [4]，加算スムージングはその中でも最も単純なもので，出

現回数に一定の値を加える方法である．式は以下のとおり．

$$P(w) = \frac{n(w) + k}{C + kV}$$

ここで，$n(w)$ は文書中の単語の出現回数，$C$ は文書中の全単語の出現回数，$k$ は加算する定数である．

(3) Adaptive Regularization of Weight Vectors(AROW)

文書を特徴変換した後に用いる学習器である．AROW はオンライン学習を行う線形分類器である．線形分類器とは，入力データから線形の超平面を求める分類器を線形分類器という．また，各訓練事例が 1 つ与えられるたびにパラメータを更新する方法をオンライン学習という．

AROW は Confidence Weighted Online Learning(CW) という手法が前身である．CW では D 次元ベクトル $x$ を入力として受け取り，重みベクトル $w$ をかけた値 $w \cdot x$ の符号で分類を行う．このとき，CW では重みベクトル $w$ にガウス分布 $N(\mu, \sum)$ を導入する．$\mu$ は分散の平均を，$\sum$ は共分散を表しており，訓練例を受け取るごとに、重みベクトルの平均・共分散を更新する．分散の大きい重みパラメータは大きく更新し，分散の小さいパラメータは小さく更新するアルゴリズムを実現している [5].

CW は，与えられたデータについて正しい分類を行う確率が $\eta$ よりも高くなるようにする，更新を行う際は以前の多変量ガウス分布に最も近いガウス分布を選択する，という 2 つの条件をもとに更新する．これを以下の式に基づき最適化を行う．

$$(\mu_{i+1}, \sigma_{i+1}) = \min D_{KL}(N(\mu, \sigma) \parallel N(\mu_i, \sigma_i))$$
$$s.t. Pr_{w \sim N(\mu, \sigma)} [y_i (w \cdot x_i) \geq 0] \geq \eta$$

この式は KKT-condition で解析的に解くことが出来る．

しかし CW には欠点があり，必ず正しい分類を行う際に $Pr_{w \sim N(\mu, \sum)} [y_i (w \cdot x_i) \geq 0] \geq \eta$ となるように更新するという非常に aggressive な分類を行うため，誤ったラベルが付けられていたときにそのラベルについても正しい分類を行うような学習を行ってしまい，ラベルノイズを含むデータに対しては精度が極端に落ちてしまう．そこで，この欠点を克服した AROW が考案された．

AROW は既存のオンライン学習の手法の良い点を取り入れたものであり，Non-mistake の場合でも Update，更新回数の少ない重みベクトルをより大きく更新，線形分離不可能なデータに対しても精度があまり落ちない，という特徴を持つ [6].

AROW では，データが与えられるたびに次の目的関数を最小化する．

$$C(\mu, \sum) = D_{KL}(N(\mu, \sum) \parallel N(\mu_{t-1}, \sum_{t-1}))$$
$$+\lambda_1 l_{h^2}(y_t, \mu \cdot x_t) + \lambda_2 x^{\mathrm{T}} \sum x_t$$
$$l_{h^2}(y_t, \mu \cdot x_t) = (\max 0, 1 - y_t(\mu \cdot x_t))^2$$

ここで，$l_{h2}$ は損失関数で，$\lambda_1$，$\lambda_2$ は正のパラメータである．この目的関数は次の 3 項からできている．

- 第 1 項：KL Divergence を最小化することで分布が急激に変わることを防ぐ役割を持っている．
- 第 2 項：新しいパラメーターが現在のデータに対して損失関数を小さくできることを表す．
- 第 3 項：データが増えることで confidence が一般的には上がっていくことを表す．

第 2 項と第 3 項が CW の弱点を補うために付け加えられた項である．

上記により，AROW はラベルノイズに頑健でありつつ高い分類精度を発揮することが可能となった．また，ハイパパラメータは一つしかなく，その変動にも鈍感であるため，使い勝手の良さから本研究では教師あり線形分類器に AROW を採用した．

また，文書分類では Naive Bayes や Term Weighted Complement Naive Bayes[7] などが主流であるが，これらはバッチ学習であり，オンライン学習を行うことはできない．AROW はオンライ

ン学習であるため，後段の Active Learning にデータの問い合わせとデータの学習を逐次的に行うことができる．

(4) Active Learning

弁護士が膨大な文書群からピックアップした少数の教師データをもとに仮識別器を生成し，データプールの文書に対してラベル付けを行う．

Active Learning は，教師データの作成のコストを抑えながらも，分類器の性能向上を測る手法である [8]．基本的な手法は，膨大なラベルなしデータ群から，このデータの正解がわかれば性能が向上すると思われるデータを選択する．そのデータを Oracle と呼ばれるラベルを教えてくれる仮識別器に問い合わせ，得たラベルを教師データに追加する．

Active Learning の流れは以下の通りである．

1. ラベルなしデータをプールとして大量に貯蓄する
2. 現在のモデルにおいて，学習にもっとも有用だ考えられるデータをプールの中から選択する (query)
3. 選択したデータにラベル付けを行い，教師データ群に追加する
4. 2〜3 を任意の回数繰り返す

Query を選ぶ戦略には Expected Error Reducation をベースにした MM-MS と呼ばれる手法を採用している．Expected Error Reducation は訓練データ L で学習した学習器において，点 x におけるラベル y の確率を $P(y|x;L)$ とすると，訓練データに $(x_i, y_i)$ を追加したときの log-less は以下の式で得られる．

$$-\sum_{x\in U}\sum_{y} P(y|x, L\cup(x_i,y_i))\log P(y|x, L\cup(x_i,y_i))$$

これを $R(x_i, y_i; L)$ と置く．$y_i$ が未知であるため，現在の学習器での期待値を評価値とする．

$$E_{yi}[R(x_i, y_i; L)] = \sum_{y_i} P(y_i|x_i; L)R(x_i, y_i; L)$$

これの式を最小にするような $x_i$ を queryt とし，その期待値を正解ラベルとする．

しかし，Expected Error Reducation はクラス数×プールサイズ回学習器を更新しなくてはならないため，非常に低速である．これをベースにした MM-MS では，margin が小さい順に T 個抽出し，MCMI[min] が最小のものを query とすることで，高速で Active Learning を実行することができる．

以上のことから，大量の教師データを用意する必要がなくなり，少ない労力で高い分類精度を発揮することが期待できる．

# 3 実験

## 3.1 実験内容

実験の内容は個人情報流出事件を題材とし，それが内部の犯行か外部の犯行であるかを分類する．内部犯行である場合は，それを重要文書とする．

実験で使う文書群であるが，実際の裁判で提出された証拠文書を入手することは困難であるため，代替としてセキュリティ関連のニュースを取りまとめている Security Next の記事を利用する．

## 3.2 実験環境

学習データの件数を表 2 に，評価データの件数を表 3 に示す．

| | |
|---|---|
| 内部犯行のデータ | 200 件 |
| 外部犯行のデータ | 200 件 |
| データプールのサイズ | 3000 件 |

表 2: 学習データの件数

| | |
|---|---|
| 内部犯行のデータ | 400 件 |
| 外部犯行のデータ | 400 件 |

表 3: 評価データの件数

### 3.3 分類実験

実験では提案手法と従来手法の分類精度 (Accuracy) を比較した実験結果を示す.

なお，Predictive Cording のツールも入手できなかったため，文書分類の定番手法である Bag of Words + Naive Bayse と比較した．また，Active Learning の有効性を調べるため，提案手法での Active Learning があるかないかでの精度差も比較して実験を行った．実験結果を表 4 に示す.

| | |
|---|---|
| BoW + NB | 89.4% |
| 提案手法 (AL なし) | 91.7% |
| 提案手法 (AL あり) | 100% |

表 4: 実験結果

## 4 考察

提案手法は従来手法に比べて高い分類精度を出すことができた．また，Active Learning は非常に効果があることがわかり，Active Learning を利用することで，証拠文書を完璧に分類することができた．このことから，Active Leaning は得られるラベル付き文書が少ない場合，有効な手法であることが示された.

## 5 おわりに

本研究で E ディスカバリを効率化するための機械学習アルゴリズムを提案した．今後はあらゆるデータセットに対して安定した分類精度を出せるようにしていきたい.

## 参考文献


[1] eDiscovery Blog：「Predictive Cording とは？」, http://ediscoveryblog.ji2.co.jp/?p=1012

[2] Quoc Le, Tomas Mikolov, “Distributed Representations of Sentences and Documents”, http://cs.stanford.edu/ quocle/paragraph_vector.pdf,Google Inc, 1600 Amphitheatre Parkway, Mountain View, CA 94043.

[3] Yanhong YUAN, Liming HE, Li PENG, Zhixing HUANG,“A New Study Based on Word2vec and Cluster for Document Categorization”, http://www.jofcis.com/publishedpapers/ 2014_10_21_9301_9308.pdf, School of Computer and Information Science, Southwest University, Chongqing 400715, China.

[4] Bill MacCartney, “NLP Lunch Tutorial: Smoothing”, http://nlp.stanford.edu/ wcmac/papers/ 20050421-smoothing-tutorial.pdf, Stanford University, 21 April 2005.

[5] Mark Dredze, Koby Crammer, “Confidence-Weighted Linear Classification”, http://www.cs.jhu.edu/mdredze/ publications/icml_variance.pdf, ICML2008, p.3 pp.

[6] Koby Crammer, Alex Kulesza, Mark Dredze, “Adaptive Regularization of Weight Vectors”, http://www.cs.jhu.edu/%7Emdredze/ publications/nips09_arow.pdf, Crammer et al., 2009.

[7] Jason D. M. Rennie, Lawrence Shih, “Tackling the Poor Assumptions of Naive Bayes Text Classifiers”, http://people.csail.mit.edu/jrennie/papers/icml03-nb.pdf, Artificial Intelligence Laboratory; Massachusetts Institute of Technology; Cambridge, MA 02139.

[8] Yi Zhang, “Active Learning”, https://www.cs.cmu.edu/ tom/10701_sp11/recitations/ Recitation_13.pdf