3C-2

機械学習を用いたWeb上の産学連携関連文書の抽出

蔵川　圭†　　孫　媛†　　馬場　康維‡
国立情報学研究所†　　統計数理研究所‡

1.はじめに

科学技術政策立案のための情報源として産官学連携の実態を把握することの重要性が指摘されている[1]．Web 上の情報は実態把握のための有用な情報源の一つである．本研究は，そのようなWeb上の文書を収集・整理することを前提として，Web 上の大学や企業のプレスリリースに着目し，産学連携関連情報の収集および分類する手法の構築を目的とする．

本報では，Web から収集した文書を２つの観点から分類する実験を行った．一つは文書の産学連携関連かどうかの判別であり，もう一つは産学連携関連文書のトピックによる分類である．以降，これらの実験について述べる．

2.Web 上の産学連携関連文書の判別と分類

Web から文書を産学連携かどうか判別し，トピックに応じて分類する手順を示す(図 1)．

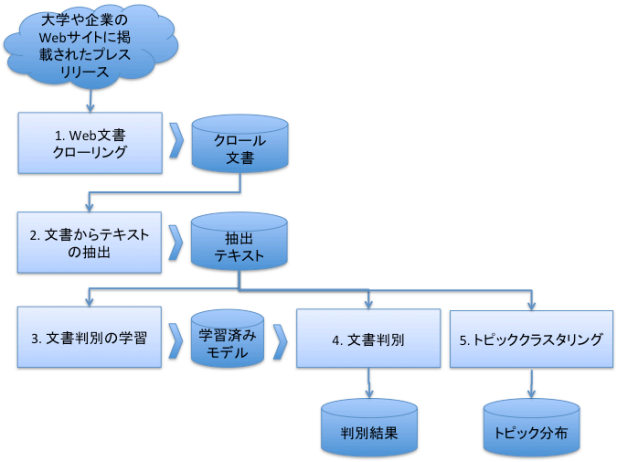

図 1 産学連携関連文書の判別とトピック分類

2.1. クローリングおよび文書判別

Web 上の大学や企業のプレスリリースは，定型的でフォーマルな文章によって，産学連携の実態を説明することが多い．産学連携の根拠は，多くは１～２文に現れ，「○○大学と株式会社△△が，××に関する研究成果を発表した．」などのように表現される．文書全体からすると産学連携の実態はごく一部の文に現れ，その他大多数の文は研究成果についての記述である．

文書の判別には二値分類器である SVM(Support Vector Machine)[2]を用いる．SVM の入力となる特徴ベクトルは，図 2に示すように上述の産学連携関連文書の特徴を考慮して定義する．Web から HTML 文書をクロールし，HTML タグ除去し，メニュー項目などのノイズ除去のために句点をもつ一連の文のみを抽出し，抽出したテキストを元に特徴ベクトルを構成する．

	Features	Description
(1)	BoW	Bag of Words. 形態素解析器Mecabを用いて，文を形態素に分割.
(2)	BoW(N)	Bag of Words. 名詞のみ抽出.
(3)	BoW(N-3)	Bag of Words. 名詞-一般，名詞-固有名詞，名詞-サ変接続に限定して名詞を抽出.
(4)	K(14)	14個の関連キーワード. 研究，開発，実験，成功，発見，開始，受賞，表彰，共同，協同，協力，産学，産官学，連携
(5)	K(18)	18個の関連キーワード. K(14)に加えて，受託，委託，締結，研究員
(6)	K(18)+NM	K(18)のキーワードと文中後接形態素の品詞の組．ただし，研究，開発，実験，成功，発見，開始，受賞，表彰，受託，委託，締結の11キーワードに対して，後接形態素の品詞を動詞-*，助動詞-*，名詞-サ変接続に限定．共同，協同，協力，産学，産官学，連携，研究員の7キーワードに対し，後接形態素の品詞は任意.
(7)	ORG	係り受け解析器Cabochaの固有表現抽出機能で形態素に，組織タグが付与されているかどうか.
(8)	Corp.	会社を表すキーワード．株式会社，(株)，(株
(9)	Univ.	大学を表すキーワード．大学，大
(10)	C.+U.	一文に，Corp.およびUniv.のキーワードが同時に存在するかどうか.

図 2 産学連携関連文書を判別する特徴ベクトル要素

2.2. トピックによる分類

トピックによる分類は，同一の抽出したテキストに対して LDA (Latent Dirichlet Allocation)[3]を用いる．

3.実験

大学または企業のプレスリリースサイトをクロールし，クロールした文書に対して SVM による判別および LDA によるトピック分類をそれぞれ行った．

3.1. データセット

組織	クロールしたHTML		実験に利用するHTML	
	正例	負例	正例	負例
東北大学	44	499	44	44
東京大学	106	848	106	106
京都大学	40	329	40	40
東京工業大学	37	343	37	37
日立製作所	103	450	103	103
合計	**330**	**2469**	**330**	**330**

図 3 実験に用いた文書

Machine learning based extraction for university-industry relation documents on the Web
†National Institute of Informatics
‡The Institute of Statistical Mathematics

Test ID	TF-IDF Feature Element										Kernel
	BoW	BoW(N)	BoW(N-3)	K(14)	K(18)	K(18)+NM	ORG	Corp.	Univ.	C.+U.	
1-1	✔										Linear
1-2		✔									Linear
1-3			✔								Linear
2-1				✔							Linear
2-2				✔							Polynomial
2-3				✔							RBF
3-1					✔						Linear
3-2					✔						Polynomial
3-3					✔						RBF
4-1						✔					Linear
4-2						✔					Polynomial
4-3						✔					RBF
5-1						✔	✔				Linear
5-2						✔	✔				Polynomial
5-3						✔	✔				RBF
6-1						✔	✔	✔	✔		Linear
6-2						✔	✔	✔	✔		Polynomial
6-3						✔	✔	✔	✔		RBF
7-1						✔	✔	✔	✔	✔	Linear
7-2						✔	✔	✔	✔	✔	Polynomial
7-3						✔	✔	✔	✔	✔	RBF
7-4						✔	✔	✔	✔	✔	RBF(γ tuned)
8-1						✔		✔	✔	✔	Linear
8-2						✔		✔	✔	✔	Polynomial
8-3						✔		✔	✔	✔	RBF
8-4						✔		✔	✔	✔	RBF(γ tuned)

図 5 SVM による判別を行うためのテストケース

Test ID	Accuracy	Precision	Recall	F-measure
1-1	61.21	64.04	42.12	47.28
1-2	60.61	63.75	40.00	45.54
1-3	61.52	67.44	40.00	46.72
2-1	67.58	72.02	61.52	63.70
2-2	58.03	69.76	23.33	34.45
2-3	66.51	62.53	86.37	71.89
3-1	68.18	72.02	63.33	64.78
3-2	57.88	69.00	23.03	34.08
3-3	66.67	62.22	88.18	72.43
4-1	70.61	74.66	63.64	67.40
4-2	-	-	-	-
4-3	70.76	65.49	90.30	75.66
5-1	70.61	74.61	63.64	67.31
5-2	-	-	-	-
5-3	70.76	65.49	90.30	75.66
6-1	-	-	-	-
6-2	-	-	-	-
6-3	70.15	64.64	93.64	76.09
7-1	78.94	85.03	71.82	77.16
7-2	-	-	-	-
7-3	71.82	65.73	94.85	77.35
7-4	79.85	78.51	83.94	80.86
8-1	78.79	85.01	71.52	76.99
8-2	-	-	-	-
8-3	72.27	66.07	**94.85**	77.61
8-4	**80.15**	78.81	83.94	**81.05**

図 5 テスト結果

大学または企業のプレスリリースサイトを wget によりクロールし，人手による判定を行った上で，図 3に示すように，実験のために正例と負例が同数になるようにデータセットを準備した．

3.2. SVM による文書判別

定義した特徴ベクトル要素がどの程度有効であるか判断するため，図 5に示すようにテストケースを用意した．ここでは，特徴ベクトルにおける要素の組み合わせと同時に，カーネル関数の選択に関しても考慮する．

形態素解析器の実装として MeCab[4]，SVM の実装として SVM^{light}[5]を使用し，データセットに対し 10 分割交差検定を行った．結果を図 5に示す．

3.3. LDA によるトピック分類

LDA の実装は，MALLET[6]を用いた．ここではトピック数を 60，Gibbs Sampling の回数を 2000 とした．また，ストップワードとして，キーワードとして不適切なトピック中の重みの高い語を指定した．

topic# [K=60]	Word (weight) [Rank>21]
37	研究 (326) 連携 (232) 的 (229) 社会 (199) 東京大学 (186) 大学 (147) 共同 (134) 学 (122) 企業 (105) 者 (100) 分野 (99) 産学 (89) 推進 (87) 東京 (72) 等 (70) 教育 (69) 化 (69) 活動 (68) テーマ (67) 大学院 (65)
43	温度 (130) プラズマ (73) 超電導 (69) 表面 (56) 化 (54) 粒子 (54) 装置 (46) 線 (40) 低温 (36) 分布 (33) 材料 (33) ガス (31) 材 (28) 反応 (28) 分光 (27) ホウ素 (25) 電流 (25) 処理 (25) エッチング (24) 高温 (23)
49	データ (85) 解析 (81) データベース (77) シミュレーション (75) ストレージ (72) 検索 (71) 計算 (66) システム (62) 高速 (54) 規模 (51) 文書 (48) 処理 (45) メッシュ (34) エンジン (30) 性能 (30) スパコン (29) 的 (29) 形状 (28) HDD (28) ソフトウェア (27)
55	研究 (1313) 技術 (875) 開発 (709) 科学 (334) 年 (270) 世界 (240) 共同 (214) 法人 (209) 大学 (205) プロジェクト (202) センター (183) 機構 (177) 産業 (174) 株式会社 (173) 成果 (157) 的 (157) 事業 (155) 利用 (140) 行政 (132) 研究所 (118)
59	東京大学 (315) 発表 (243) 月 (227) 必要 (202) 日 (197) ページ (196) 設定 (190) オン (186) リロード (181) JavaScript (176) ブラウザ (174) 表示 (174) 研究 (144) 開催 (118) 年 (109) 者 (103) 下記 (90) 研究所 (89) 内容 (79) 記者 (75)

図 6 LDA によるトピック例

図 6に，出力されたトピックの例を示す．トピック 43 や 49 のように重みの高い語として研究内容を示す場合がほとんどであるが，37 や 55 のように産学連携関連のキーワードが並ぶものもあった．また，59 のようにノイズもあった．

産学連携を示すトピック 55 に着目し，図 7に示すように正例文書と負例文書ごとにトピック確率をランク順で並べたところ，グラフ上正例の確率が全体的に高いことが見て取れた．

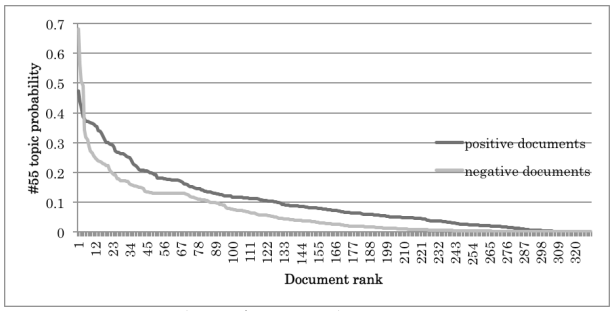

図 7 正例文書と負例文書のトピック確率

4.まとめ

SVM による判別は特徴ベクトルの定義どおりに精度を高めることができた．トピック分類は通常，内容に関する分類に用いられるが，本実験により産学連携研究開発文書の判別に作用できる可能性が示唆された．今後の展開として，LDA の判別性能を検討する．

参考文献

[1] Leydesdorff, L., Meyer, M.: The triple helix of university-industry-government relations. Scientometrics 58(2):191-203 (2003)
[2] Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, New York, NY, USA (1995)
[3] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993-1022 (2003)
[4] http://mecab.sourceforge.net
[5] http://svmlight.joachims.org
[6] http://mallet.cs.umass.edu