6P-03

確率的TF-IDFを用いた特徴語抽出と文書検索

三浦 大輝　　三浦 孝夫
法政大学理工学部創生科学科
東京都小金井市梶野町 3-7-2

1 前書き

本研究ではツイート文を利用し, 特徴語抽出および検索を試みる。Twitter サービスは SNS のひとつであり, 多くの利用者と, ツイート文と呼ばれる (140 文字以内の) 短い文章を共有する。利用者は, ”今どこどこで何食べてる！” といった単純文や, ”iPhoneX 買った”等の忘備録などを, Twitter システムに投稿する。自分の投稿とそれを ”フォロー” した利用者の投稿が時系列順 (タイムライン) に表示され, 投稿およびそのコメントを通じてコミュニティが生成される。ツイート文に対するキーワード検索が可能であるが限定的であり, 例えば情報検索に欠かせない特徴語抽出や有効な検索手法は有効でない。情報検索の観点から見れば,SNS の文書には (1) サイズが小さい (2) 文書数の更新が多い (3) 正確でない語彙や構文が多い, という特徴がある。このような SNS 文書に従来の情報検索の手法の一つである TF-IDF を用いる場合, IDF の計算が困難という問題がある。

本研究ではツイート文を単位とする情報検索を目的として, 単語 w の出現確率 $P(w)$ により IDF を推定し [1], 確率的 TF-IDF による文書処理の手法を提案する。リアルタイムに更新される Twitter などの文書は IDF の計算に必要な文書頻度を数えることが非常に困難である。本研究では特徴語抽出と文書検索を通して確率的 TF-IDF による文書処理の手法を提案する。

2 TF-IDF モデルと提案手法

TF-IDF は, 文書中に含まれる単語の重要度を評価する手法の 1 つであり, 主に情報検索やトピック分析などの分野で用いられている [2]。語 w の文書集合 D における語出現頻度 (Term Frequency, TF) は D 全体での w の相対出現頻度であり, 単語の頻度の全文書の総単語頻度での比である。一般的な語ほど出現数は大きい。一方, w の文書頻度 (Document Frequency, DF) とは w を含む文書数 N_w であり, 大きいほど特徴を失う。逆文書頻度 (Inverse Document Frequency, IDF) は総文書数 N に対する DF の割合の逆数対数 $\log N/N_w$ であり, 特徴的であるほど大きな値を示す。IDF は文書の重みを考慮した一種のフィルタとして働き, 多くの文書に出現する語（一般的な語）は特徴が下がり, 特定の文書にしか出現しない単語の特徴を上げる指標となる。TF×IDF は両者の特徴を有するため, 情報検索では重み付き語として利用される。

Aizawa らは TF-IDF 概念をモデル化するため, エントロピの概念を利用した定式化を提案している [1]。文書集合 $D = \{d_1, .., d_N\}$, 語 w_i の d_j での頻度 TF を f_{ij}, D での出現頻度を f_{w_i}, D での総語出現頻度を F, w_i を含む D の文書集合を D_{w_i}, その文書数を N_i とする。このとき D のエントロピ $H(D) = \Sigma_{j=1,..,N} P(d_j)\log(1/P(d_j))$, 語 w_i を含む文書集合のエントロピ $H(D|w_i) = \Sigma_{j=1,..,N} P(d_j)\log(1/P(d_j|w_i))$ と定義すると, その相互情報量 $I(D, D_w) = H(D) - H(D_w)$ は $\Sigma_i(\Sigma_d(f_{ij}/F))\log(N/N_i)$ となる。$\Sigma_d(f_{ij}/F)$ は 語頻度 TF であり, これに対する重み $\log(N/N_i)$ は逆文書頻度 IDF に対応する。

語の重要度を計算するため, 分野を特定した重要語の抽出あるいは重み付けの観点から, 滝川は TF-IDF の拡張および相互情報量の利用を提案した [3]。

本研究では, TF はそのまま計算し, IDF を確率推定する。このため, 全ての単語出現が多項分布に従うと仮定する。この生成確率を用いて確率的 TF-IDF を計算し, 閾値以上の単語を特徴語として抽出する。文書検索では, まず各文書ごとに TF-IDF 及び確率的 TF-IDF を算出して文書ベクトルを作成し, さらに質問文書に対しても同様に文書ベクトルを作成して, 余弦類似度を用いた検索を行う。

語 w の生成確率分布 $P(w)$ を推定するため, 本稿では (検索文書集合とは独立の) 学習文書内の単語集合が多項分布に従うと仮定する。ゼロ頻度問題を防ぐために最大事後確率 (MAP) を推定する。

3 実験

3.1 準備

本稿で提案する手法の優位性を示すため, 特徴語抽出および情報検索結果を検証する。本実験では,20,000 件のツイート文を対象としてコーパス D を構成する。このため 統計解析ツール R のパッケージ (twitterR) を用い, ツイート文 (索引語数 11,117 語, 2017 年 11 月 20 日) 20,000 件を収集する。このコーパスには「人工知能」,「仮想通貨」,「ガンダム」,「モンスト」に関する文書が 5,000 件ずつ含まれている。収集した文書を MeCab により形態素解析して自立語*を抽出し, 一般的な不要語や機能語 (”「”, ”(”などの記号) は除去しない。特徴語抽出および文書検索の実験では, コーパスとして 5,000 件, 10,000 件, 15,000 件, 20,000 件を対象として実験する。

文書集合 D に生じる語 $w_1, .., w_N$ について, 各単語 w_i の出現頻度を f_{w_i}, 語総出現頻度を $F = \Sigma = i = 1, .., N f_{w_i}$ として, $P(w_i) = (f_{w_i} + 1)/(F + 1)$ により MAP 推定値を算出する。特徴語抽出ステップでは, 推定した $P(x)$ を用いて $\log(1/P(x))$ を IDF として推定し, 確率的 TF-IDF $f_{w_i} \times \log(1/P(w_i))$ を定める。同時に, 評価のために従来の TF-IDF も計算する。語 w_i の文書数 N_i を求め上位 50 番目までの語を特徴語とする。従来の TF-IDF の抽出結果を正解として確率的 TF-IDF の抽出結果の適合率を評価する。確率的 TF-IDF を用いた文書検索を行う。このため, 推定した IDF による確率的 TF-IDF を用い, 文書 d を 50 次元ベクトル化する。従来の TF-IDF でも同様に文書ベクトルを作成する。

質問ベクトルも同様に従来の TF-IDF と確率的 TF-IDF それぞれで文書ベクトル化する。

作成した文書ベクトルを用いて質問ベクトルと検索対象の文書で余弦類似度を計算し, 検索する。評価は従来の TF-IDF と確率的 TF-IDF のそれぞれの検索上位を, 共有率と順位相関によって行う。

Feature Word Extraction and Document Retrieval Using Probabilistic TF-IDF
Masaki MIURA　Takao MIURA
Dept.of Advanced Science, HOSEI University

*名詞, 動詞, 形容詞, 形容動詞, 副詞

3.2 実験結果

特徴語抽出の結果を, 各文書数ごとの適合率と特徴語の数, および特徴語として抽出された単語の TF-IDF 値と確率的 TF-IDF 値の誤差平均を表 1 に示す。特徴語抽出では, すべてのコーパスにおいて適合率が 90%を超える。また, 推定した TF-IDF 値の誤差は実際のそれと最大 3560.641 である。

コーパス	正解数	適合率 (%)	TF-IDF 平均誤差
5,000	47	94	890.283
10,000	48	96	3452.163
15,000	48	96	3230.605
20,000	49	98	3560.641

表 1: 特徴語数と TF-IDF 誤差

確率的 TF-IDF を用いた文書検索の結果として, 検索上位 50 位の共有率, および文書数ごとの共有した文書における順位相関係数を表 2 に示す。

コーパス	共有数	共有率 (%)	順方向数	逆方向数	組合せ数	順位相関係数 (%)
5,000	30	60	405	30	435	0.86
10,000	39	78	725	16	741	0.95
15,000	46	92	1026	9	1035	0.98
20,000	49	98	1171	6	1176	0.99

表 2: 検索上位 50 位の共有率と順位相関

表 2 より, 共有率および順位相関係数は文書数が増えるごとに上昇していて, 共有率は 60%最大 98%まで上昇し, 順位相関係数は 0.86 から最大 0.99 まで上昇している。

3.3 考察

実験結果に対し, 特徴語抽出について考察する。表 3 に, 文書数 20000 件における抽出された特徴語を確率的 TF-IDF 順に上位 50 個を示す。上位 20 件で ”交換”(11)/”注文”(13) のみ順序が異なっている。

表 3 より, 抽出された特徴語の, TF-IDF と確率的 TF-IDF の最大誤差は 9766 である。この値は表 1 で示した 20,000 件における推定値の誤差の約 3 倍であり, これは単語の出現頻度が増すにつれて各単語の TF-IDF の推定値の誤差が大きくなっていること, 実測値よりも推定値の方が大きい値として計算されていることが影響していると考えられる。特徴語を抽出する場合, 大きな確率的 TF-IDF 値に関しては不要語として考慮する必要があることを示している。

確率的 TF-IDF を用いた文書検索について, 文書数 20000 件における文書検索結果では, 上位 20 位まではほぼ同一であり, 順序が異なる文書 (16 位/17 位) があるのみである。

実際, 文書検索では, TF-IDF と確率 TF-IDF での類似度の誤差は 0.02 でありほぼ一致する。これは実験結果 (表 2) より共有率と順位相関が共に高いことから, TF-IDF 値と確率的 TF-IDF 値の差が大きくても検索結果への影響が小さいためであると考えられる。

4 結論

従来の単純な文書頻度の計算と異なって, 逆文書頻度 (IDF) を語 x の出現確率 $P(x)$ から推定可能である。本研究では, 文書集合の安定的な管理が困難な SNS などに, これまでの TF-IDF 技法による情報検索が可能であることを示した。実際, 特徴語抽出では適合率 90%以上, ツイート文検索では検索上位で共有率 90%以上, 順位相関係数 0.99 を達成できる。このことにより, 確率的 TF-IDF に基づく文書検索処理が十分実用性に耐えられることを実験で確認した。

確率的 TF-IDF の今後の課題として, ツイート文やリツイート文など, 内部構造を含む文書関連に対して 確率的 TF-IDF が効果的となモデルの拡張, また検索手法と更新を伴う状況下での高速なモデル計算を構築する必要がある。

順位	特徴語	TF-IDF	特徴語	確率的 TF-IDF
	TF-IDF			確率的 TF-IDF
順位	語	tfidf	順位	語	確率 tfidf
1	機動	3153.395	1	機動	12919.4795
2	戦士	3124.4918	2	戦士	12801.0608
3	開発	2343.6378	3	開発	8241.9411
4	ルンバ	2681.7956	4	ルンバ	7756.4244
5	使う	2034.71	5	使う	6832.7586
6	買取	2012.327	6	買取	6702.4981
7	搭載	1980.3288	7	搭載	6595.9212
8	スピーカー	1793.6553	8	スピーカー	5719.6229
9	機能	1789.9417	9	機能	5707.781
10	商品	1771.3738	10	商品	5648.5717
11	交換	1767.2177	11	認識	5553.8367
12	認識	1752.967	12	音声	5530.153
13	インターネット	1748.958	13	注文	5518.3111
13	起こる	1748.958	14	起こる	5506.4692
15	音声	1745.4917	14	インターネット	5506.4692
16	注文	1741.754	16	ペット	5459.1018
17	ペット	1733.9132	17	交換	5435.418
18	飼い主	1722.6296	18	飼い主	5423.5762
18	まねる	1722.6296	18	まねる	5423.5762
20	ヨウム	1718.8684	20	ヨウム	5411.7343
20	ロンドン	1718.8684	20	ロンドン	5411.7343
20	珍事	1718.8684	20	珍事	5411.7343
23	販売	1689.05	23	販売	5139.3713
24	アカウント	1653.973	24	アカウント	5115.6875
25	好き	1643.0496	25	小隊	4902.5339
26	小隊	1602.6772	26	知る	4843.3245
27	知る	1591.7545	27	オーブ	4618.329
28	オーブ	1540.985	28	すごい	4263.0729
29	すごい	1442.6399	29	くださる	4203.8636
30	くださる	1428.9998	30	意味	4192.0217
31	意味	1424.9744	30	お願い	4192.0217
31	お願い	1424.9744	32	フォロー	4085.4449
33	フォロー	1423.8613	33	もん	4061.7612
34	もん	1392.7319	34	戦場	4002.5518
35	戦場	1378.2076	35	量子	3990.7099
36	量子	1374.1301	36	具合	3967.0262
37	具合	1365.975	37	平和	3943.3425
38	平和	1363.4167	38	想像	3919.6587
39	想像	1355.228	38	若い	3919.6587
39	若い	1355.228	40	部屋	3884.1331
41	できる	1349.5729	40	きれる	3884.1331
42	部屋	1348.3666	40	おっさん	3884.1331
43	地雷	1344.2557	43	地雷	3872.2913
43	汚い	1344.2557	43	汚い	3872.2913
45	きれる	1342.945	43	できる	3872.2913
45	おっさん	1342.945	46	無茶	3860.4494
47	見る	1341.8034	47	除去	3848.6
48	無茶	1340.1448	48	人命	3848.6075
49	除去	1336.034	49	見る	3765.7144
50	人命	1336.034	50	公開	3742.0307

表 3: 抽出された特徴語上位 50

参考文献

[1] Aizawa,A.:An Information-Theoretic Perspective of Tf-idf Measures, Information Processing and Management, Vol.39, No.1, pp.45-65 (2003)

[2] Salton, G., Fox, E.A. and Wu, H.: Extended Boolean Information Retrieval, CACM, Vol.26, No.11, pp.1022–1036 (1983)

[3] Takigawa, M. and Yamana, H.: 特定分野を対象とした単語重要度計算手法の提案と Twitter における専門性推定への適応, 第 15 回情報科学技術フォーラム (FIT), RD-001 (2016)