n-TF*IDFによる情報検索

樋山友理香 三浦 孝夫

† 法政大学 理工学部創生科学科 〒 184-8584 東京都小金井市梶野町 3-7-2 E-mail: †yurika.hiyama.5p@stu.hosei.ac.jp, ††miurat@hosei.ac.jp

あらまし 情報検索では 1-Gram の TF*IDF を用いることが多い. 特に IDF(逆文書頻度) の算出に, 語の生起確率の情報量から推定する方法が知られている。本研究では n-Gram を基本とする TF*IDF による情報検索の提案とその評価を行う。初めに, 生起確率から n-IDF の推定が可能であり, 効率よい情報検索が可能であることを示す。また, 本研究では語間のマルコフ性を用いた情報検索手法や,2-TF*IDF の算出が大幅に簡素化されることを示す。

キーワード 自然言語処理, 情報検索, 情報量, マルコフ遷移, n-TF*IDF

1. 前書き

近年、インターネットの普及に伴い日々膨大な文書データが生成されており、ユーザーが意図している文書を抽出することがますます困難になっている. 現在、情報検索では TF*IDF が主要な働きをしており、その算出は主に 1-Gram が用いられている.

本稿では,n-GramTF*IDF (以下 n-TF*IDF) に着目し, より精度の高い情報検索が可能とする環境を構築する.

この基礎には、本研究で提案する、n-TF*IDF の計算簡略化がある。ここで、n-TF*IDF の計算の簡略化を提案する。 2 章でn-Gram の基礎概念について述べ、3 章で情報検索に用いる従来の 1-GramTF*IDF,IDF の推定とマルコフ遷移について述べる。 4 章で具体的な提案手法を述べ、5 章で実験結果を示し、6 章で結論を述べる。

2. *n*-Gram

n-Gram とは n 個の連続した語がなす意味表現のことであり,N 個で単一のオブジェクトを表現する。例えば (ドナルド・ダッグ) (ドナルド・トランプ) (ドナルド・キーン) は共通語 "ドナルド"を構成しているが,その語だけで意味を生成することができない。n-Gram は意味が構成する語の意味とは全く異なるため,1-Gram を用いた情報検索でオブジェクトの区別がつかない。このことから n-Gram によって 1-Gram より複雑な情報を与えて情報検索可能となる。

2-Gram(n-Gram) の抽出には 2 つの問題点がある。第 1 に,抽出の困難さである。2-Gram は「単なる 2 語の並び」ではなく,2 語で独立した意味を成すことの区別が難しい。例えば「東京」の「大学」と「東京大学」では意味が異なっている。第 2 に,2-Gram として複数意味を持つ場合(語義曖昧性)がある。例えば「茶飲み友達」とは,お茶を飲みながら会話を楽しむ真柄の友達という意味であるが,必ずしも毎回お茶を飲んでいるわけではない。しかし文脈でお茶を飲むかどうかの判別は容易ではない。第 2 の問題は n-Gram に限るものではないので,ここでは論じない。一般に n-Gram は熟語,連語,慣用句などのような語彙として辞書に含まれることが多く,また特殊な出現特

性を有することが多い. 従って n-Gram の検出は, 辞書操作を伴うか, または (n-Gram 抽出のための) 統計的フィルタを用いてなされることが多い. とくに n-Gram の検出には自己相互情報量 (Pairwise Mutual Information, PMI) が効果的であることが知られる [3] [4] [5] [6]. σ をしきい値, 連続出現する語列をa,bとするとき、PMI を以下の式で定義する.

$$-log\frac{P(ab)}{P(a)P(b)} + \sigma > 0$$

これは 語 a,b のつながりの強さを示す指標になり, 値が大きいほど n-Gram として機能している可能性が高くなる.

3. 情報検索

TF*IDFとその計算について言及する. TF(Term Frequency, 語頻度) は語 w の出現数 f(w) を表し, 重要語や慣用的な語では頻出することが多い. 語の生起確率 P(w) の算出は次のように定義する.

$$P(w) = \frac{f(w)}{\sum_{i} f(w_i)}$$

IDF(逆文書頻度) は、語wが一度でも出現する文書数 N_i の逆数を対数化したものであり、次のように定義される。ここでNは総文書数を表す。

$$IDF(w) = log \frac{N}{N}$$

 ${
m IDF}(w)$ の値が大きい語 w は特定の文書に生じやすいことを示す。 両者の積 ${
m TF}^*{
m IDF}$ は 文書中で語 w が果たす重要度を示す値である。例えば「思う」「私」など、どの文書にも出現する語は ${
m TF}$ 値が大きいが ${
m IDF}$ 値が小さくなるため ${
m TF}^*{
m IDF}$ 値が小さくなる。 この値が高いほど語 w_i はその文書検索には重要であるといえる。 効果的な情報検索に ${
m TF}, {
m IDF}$ の特性を利用しやすい。 語 w_i の ${
m TF}^*{
m IDF}$ は次のように定義される。

$$f(w_i) * log \frac{N}{N_i}$$

IDF を算出するためには、語w毎に それが出現する文書を数える必要がある。しかし (SNS など) 短い間に大量に投稿がされる状況データでは、文書数の変動が大きく、計算が困難である。そ

のため、語の生確起率 p(w) の情報量 $-\log p(w)$ で推定できることが知られている [1]. この IDF の推定は TF の情報のみで算出可能であり、しかも推定精度がよい [7]. この結果、TF*IDFは TF 値のみで計算でき、計算が大幅に簡略化できる。 n-IDFの算出も、n-Gram を含む文書を数えるため、計算時間及び記憶域量が 1-Gram 以上に大きくなることから、その値の推定が可能ならば大幅な処理効率向上になる.

語列 $w,w_1,w_2,...,w_n$ を考える. w がマルコフ性を有するとは, w が直前の語 w_1 だけに依存して確率的に生起するときをいう.

$$P(w \mid w_1 w_2 ... w_n) = P(w \mid w_1)$$

本研究では、すべての語がマルコフ性を有すると仮定する.文書中に語 a,b が出現しているとき、この後続頻度をすべて計算しておくとする. すなわち $f(b \mid a)$ は a の直後に b が生じる頻度を表すとする. 定義より $TF(b) \ge f(b \mid a) \ge 0$ である. $P(b \mid a)$ によって a の直後に b が生起する確率を表す. このとき、マルコフ性の仮定から a がどのような状況で生起していても b の生起確率は変わらない. 2-TF 値 を推定することができる. 例えば、2-Gram w_1w_2 で w_1 = ドナルド であり、文書中に 10 回生じるならば、 w_2 = トランプ が生じる確率 $P(w_2|w_1)$ が 0.6 ならば、この 2-Gram の TF 値は TF(ドナルド、トランプ) = 10×0.6 = 6 であろう. 同様に、語 w_1 = ドナルド の生起確率 $P(w_1)$ = 0.2 ならば $P(w_1w_2)$ = $P(w_2|w_1)P(w_1)$ = 0.6×0.2 = 0.12 である.

4. 提案手法

本稿では 2-IDF の推定手法, およびマルコフ性を用いた 2-TF*IDF の推定手法を提案する.

本研究では、2-Gram 候補を抽出するため、[3] に従って、PMI フィルタを利用する.共起頻度を調べるため、1-Gram リストを利用するが、例えば形態素解析を用いて語の構成や語幹生成などの事前処理を行っておく.1-Gram 同士の共起頻度に閾値を設け、一定以上の共起頻度を有するものとする.PMI 統計フィルタ処理によって、語のつながりの強い共起語を抽出する.

1-IDF 推定のために情報量を利用した手法 [1] と同様, 本研究では、これを 2-Gram w_1w_2 に適用して、確率的に 2-IDF を推定することを提案する。 すなわち 2-Gram 生起確率 $P(w_1w_2)$ に対して $-logP(w_1w_2)$ を 2-IDF (w_1w_2) と推定してよい、本研究では、この値が(真の)2-IDF と同等の効果を有することを検証する。

本研究では、連続する 2 語のマルコフ性を仮定して、2-TF*IDF 計算の簡略を提案し、その有効性を検証する。 2-Gram を ab とするとき、マルコフ性の下で条件確率 $P(b\mid a)$ は 出現頻度 $f(b\mid a)$ を用いて次のように算出できる:

$$P(b \mid a) = \frac{f(b \mid a)}{\sum_{b} f(b \mid a)}$$

このとき,2-TF および 2-IDF を以下のように推定する.

$$2 - TF : f(w_1 w_2) = f(w_1) * P(w_2 \mid w_1)$$

$$2 - IDF : IDF(w_1w_2) = -logf(w_1f_2)$$
$$= -log(P(w_1) * P(w_2 \mid w_1))$$

これら式では、2-TF、2-IDF 共に 1-Gram 頻度とマルコフ性の情報のみで推定しており、この結果 2-TF*IDF も推定できることに注意する.

5. 実 験

5.1 実験準備

本稿では CD-毎日新聞 2017 年版に採録されている 1 月 1 日から 1 月 14 日の 2 週間分のデータを抽出する. このうち,1 文書中総語数 200 語以上のデータサイズが大きい文書を使用する. 1 つの記事を 1 文書としたとき,文書数は 482 となる. 1-Gram 語を抽出するため,形態素解析ソフト Mecab により形態素解析を行う. Mecab が扱う 69 の品詞体系のうち 2-Gram になる形態素を 17977 語抽出する.

次に文書内で語が (w_1,w_2) と連続して出現している語を抽出し、共起頻度 $(2\text{-}\mathrm{TF})$ を求める。共起語の頻度に対して閾値を設定し、閾値以上の語を抽出する。 PMI フィルタを使い、閾値以上のものを抽出する [3]. 抽出した (w_1w_2) の生起確率を求め、IDFを推定する。 1-Gram 頻度とマルコフ性で $2\text{-}\mathrm{TF}^*\mathrm{IDF}$ を推定する。 予め連続する 2 語 (w_1w_2) の遷移表を作成する。次に w_1 の頻度を求め、 $2\text{-}\mathrm{TF}^*\mathrm{IDF}$ を推定する。

実験結果を評価するため質問ベクトルをランダムに生成し、 余弦類似度を用いて情報検索をする。そして検索結果の上位 10 位、上位 20 位を抽出する。2-IDF と推定 2-IDF の分布を比較し、 適合度検定を用いて結果分布が類似しているか評価する。また、 2-TF*IDF の情報検索結果を基準として、2-TF と推定 2-IDF を 用いた推定 2-TF*IDF の情報検索結果を適合率、ケンドール順 位相関係数を用いて評価する。2-TF、2-IDF と推定 2-TF、推定 2-IDF を比較し分布の類似性を検証する。また、2-TF*IDF の 情報検索結果を基準として、推定 2-TF*IDF の情報検索結果を 適合率、ケンドール順位相関係数を用いて評価する。

5.2 結 果

共起語として延べ語数 130796 語が生起しており、はじめに 頻度 3 以上の共起語を抽出し、次に $\sigma=11$ として PMI フィルタより抽出する. この結果見出し語数 3528 語となり、延べ語数はフィルタをする前の 15 %になる. 抽出した語の総頻度は平均 5 回である. 抽出した語で共起頻度が多い上位 10 個を図 1 に示す. 園田 [3] によると、推定精度は 68 %である.

Ì	w1	w2	共起頻度	w1	w2	共起頻度
	委員	共	129	東京	五輪	80
	党	県	108	次期	大統領	71
	党	地区	95	東海	大仰	70
	東	福岡	94	日	*	67
	トランプ	次期	82	記者	会見	62

図 1 抽出した 2-Gram

2-IDF を生起確率の情報量から推定する. 2-IDF と推定 2-IDF の分布の結果を表 1 に示す.

表 1 2-IDF と推定 2-IDF の誤差

衣 I Z-IDF C推足 Z-IDF の										
最小	0.0059	平均	0.67							
最大	2.36	分散	0.22							

1 	表	2	推定	2-IDF	情報検索結果
--	---	---	----	-------	--------

1		適合度	ケンドール
	上位 10 位	1	0.822
	上位 20 位	0.95	0.884

誤差の平均と分散が小さく, 推定 2-IDF が 2-IDF に近似している. 有意水準 5%で適合度検定し, カイ2乗値が520.78, 棄却域が3667.30で棄却されないことから類似した分布である.

ランダムに生成した質問文書の一部を図 2 に示す. 2-TF*IDF,2-TF と推定 2-IDF を用いた推定 2-TF*IDF の情報 検索の結果を図 3 に示す. 表 2 より上位 10 位, 上位 20 位とも

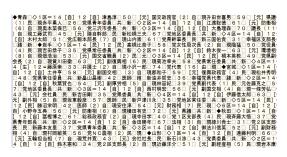


図2 質問文書

TOP	IDF(文書会号)	simcos	推定IDF(文書會号)	simcos	TOP	IDF(文書会号)	simcos	推定IDF(文書音号)	simcos
1	56	1	56	1	11	55	0.501485	55	0.572585
2	64	0.8245	61	0.811441	12	62	0.02065	62	0.028068
3	59	0.823671	64	0.809351	13	247	0.013584	421	0.019816
4	66	0.818629	59	0.809304	14	67	0.012127	247	0.019365
5	61	0.815757	66	0.799066	15	8	0.011593	8	0.016238
6	65	0.806143	65	0.792504	16	421	0.011485	212	0.014041
7	57	0.715446	57	0.716544	17	212	0.011057	216	0.01309
8	58	0.634607	60	0.667331	18	216	0.010303	67	0.012905
9	60	0.620947	58	0.664298	19	171	0.008961	425	0.012123
10	63	0.597647	63	0.640309	20	355	0.007752	355	0.011805

図3 2-IDFと推定2-IDFを用いた情報検索結果

適合率とケンドール順位相関係数が高い。特に図 3 で上位 10 位は類似度 0.6 以上の類似文書からなり、推定 2-IDF ですべての文書を検出している。

遷移表は 2126 個からなる. ランダムに抽出した文書 9 の一部 (図 4) と, その 2-TF と推定 2-TF の結果を表 3 に示す. 2-IDF と推定 2-IDF の分布の結果を表 4 に示す. 推定 2-TF と推定

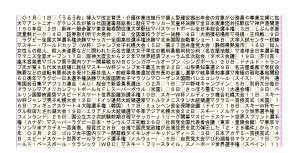


図4 文 書 9

2-IDF は誤差の平均と分散が小さく,2-TF と 2-IDF に値が近似している. 推定 2-TF を有意水準 5 %で適合度検定し,カイ2 乗値が 153.26, 乗却域が 170.80 で棄却されないことから類似

表 3 2-TF と推定 2-TF の誤差 表 4 2-IDF と推定 2-IDF の誤差

最小	0	平均	0.82	最小		0.00023	平均	1.1
最大	6	分散	1.19	最大		3.24	分散	0.3

表 5 推定 2-TF*IDF 情報検索結果

	適合度	ケンドール
上位 10 位	1	0.56
上位 20 位	0.85	0.74

表 6 推定 2-IDF の成功例と失敗例

成功例:(安	倍, 内閣)	失敗例:(健康,診断)			
推定値誤差	-0.12	推定值誤差	-1.54		
共起頻度	23	共起頻度	3		
文書数	10	文書数	1		

分布である. 推定 2-IDF を有意水準 5 %で適合度検定し,カイ 2 乗値が 1027.87, 棄却域が 3667.30 で棄却されないことから類 似分布である.

次に,2-TF*IDF と, 推定 2-TF*IDF を情報検索する. 得られた結果の上位 20 位を図 5 に示す. 図 5 より, 適合度とケンドー

TOP	TFIDF	simcos	マルコフ	simcos	TOP	TFIDF	simcos	マルコフ	simcos
1	56	1	56	1	11	55	0.501485	55	0.653212
2	64	0.8245	65	0.831931	12	62	0.02065	247	0.028326
3	59	0.823671	64	0.830028	13	247	0.013584	421	0.016322
4	66	0.818629	59	0.829795	14	67	0.012127	355	0.015808
5	61	0.815757	60	0.827469	15	8	0.011593	425	0.014356
6	65	0.806143	61	0.802033	16	421	0.011485	8	0.012105
7	57	0.715446	57	0.796488	17	212	0.011057	62	0.011238
8	58	0.634607	66	0.795104	18	216	0.010303	333	0.008028
9	60	0.620947	58	0.707949	19	171	0.008961	433	0.007784
10	63	0.597647	63	0.693409	20	355	0.007752	67	0.007323

図 5 2-TF*IDF と推定 2-TF*IDF の情報検索結果

ル順位相関係数を表 5 に示す. 表 5 より上位 10 位,上位 20 位 では情報検索の順位が前後しているが,適合率が高い. 特に図 5 の上位 10 位では類似度が約 0.6 以上の類似文書をすべて検出している.

5.3 考 察

2-IDF と推定 2-IDF がうまく近似できる語とできない語の 例を表 6, その語を含む文書例の一部を図 6, 図 7 に示す.

安倍晋三首相は年頭の配者会見を行い、米トランブ政権誕生などの国際情勢を踏まえ「変化の一年が予想される」と強調した。 国際情勢が転機を迎える年だけに、安倍政権にも従来にない発想が求められる。 経済政策などの実績を十分検証し、外交、内政とも変化に対応できる柔軟さを求めたい。□首相は2005年の「劉政解教」や1993年の自民党野党施落を引き合いに「酉(とり)年は波治の転換点となってきた」と語った。政州の選挙イヤーや英国の欧州連合(GU)離賊交渉なども念頭に、変化への備えを説いたということだろう。□外部環境の急変への対応はもちろん大切だ。た、放・販権に選聘してから5年目を迎え、安<mark>倍内閣</mark>も政策を再点検する事務を迎えている。□経政策に関し首相は「「先が見えない時こぞ」大切なことはぶれないことだ」と述べ、経済最低光を継続して金融緩和、積極財政など従来路線を維持すると強調した。だが、デフレ税却に向けて掲げた2%の物価上昇率などの目標は、連成には遠い状況に

図 6 成功例の語を含む文書



図 7 失敗例の語を含む文書

表 7 遷移表を用いた推定 2-TF の成功例と失敗例

成功例:(元	,財務)	失敗例:(民,公明党)		
推定值誤差	+0.41	推定值誤差	+6.33	
共起頻度	2	共起頻度	1	
w ₁ 頻度	28	w_1 頻度	29	
遷移確率	0.08	遷移確率	0.24	

表 6 より (安倍, 内閣) は出現する文書数が 10 件と多い.これは文書データ 482 件の中で政治に関する記事が多いため他の文書でも出現しやすかったといえる. 一方、(健康、診断) は文書数が 1 件となり、他に似ている記事がない. 語が複数文書に出現しているため、2-IDF が情報量で近似できると考える.

推定 2-TF では $\mathrm{TF}(w_1)$ の値と遷移確率が大きいとき誤差 +4 以上が生じている.ここでは特に $\mathrm{TF}(w_1) > 20$, $P(w_2 \mid w_1) > 0.2$ で推定 2-TF は不正確になる.しかし, $\mathrm{TF}(w_1) > 20$ を含む 文書は 482 件中 44 件 (9 %) である.図 8 で $w_1 > 20$ である語を含む文書の例を示す. $\mathrm{TF}(w_1 > 20)$ の語はほとんどが遷



図 8 $w_1 > 20$ の単語を含む文書

移確率 0.2 以下である.これは語のほとんどが 10 種ほどの語に続くためである.図 9 に遷移表の例を示す. $\mathrm{TF}(w_1)$ が大き

w_1	w_2	$P(w_2 w_1)$	w_1	w_2	$P(w_2 w_1)$
	経	0.098901		成長	0.3
	文科	0.082418		産業	0.1625
	総務	0.082418		先行き	0.0875
	厚	0.082418		制裁	0.075
	国交	0.082418		学者	0.0625
	衆	0.054945		動向	0.05
	財務	0.054945	経済	減速	0.0375
	外相	0.049451		恩恵	0.0375
元	農相	0.043956		活性	0.0375
	法相	0.043956		為答	0.0375
	参院	0.038462		再生	0.0375
	銀行	0.038462		けん引	0.0375
	農水	0.027473		混乱	0.0375
	参	0.021978			
	府議	0.021978			
	外務	0.021978			
	財務省	0.021978			

図 9 遷移先が多い語

くても遷移確率が小さいため、推定頻度が大きくなりすぎない、よって誤差が少なく 2-TF を推定できる。推定 2-TF が 2-TF にうまく近似できる語とできない語の例を表 7 に示す。表 7 より (元, 財務) では (元) の頻度が大きいが遷移確率が小さいため推定値の誤差が小さくなる。一方 (健康, 診断) では w_1 の頻度が同じくらいであるが遷移確率が大きくなっている。そのため推定値誤差が大きくなると考えられる。

6. 結 論

本研究では情報検索に用いられる TF*IDF を 2-Gram まで拡張し、計算の簡略化を提案した。 2-Gram の抽出では、共起頻度と PMI の閾値を設定し推定精度 68 %で抽出できた。 また計算が困難な 2-IDF は、語生起確率の情報量から近似できることを示した。 2-TF*IDF は 1-Gram 頻度と遷移表で値を推定し、類似文書は適合率 1 で情報検索することができた。

文 献

- Aizawa, A: "Information-Theoretic Perspective of Tf-idf Measure" Processing and Management, Vol.39, No.1, (2003)
- [2] Lukas,H, and Vladik,K.: "Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (tfidf) Heurustuc(and Variations Motivated by This Explanation)." International Journal of Computer and Systems (2014)
- [3] 園田匠,島田諭,三浦孝夫: "条件付きコロケーションの抽出",第 5回データ工学と情報マネジメントに関するフォーラム (DEIM), 福島/郡山, Mar., 2013
- [4] 國田匠, 三浦孝夫: Mining Japanese Collocation By Statistical Indicators, 15th International Conference on Enterprise Information Systems (ICEIS13), Angers, France
- [5] 園田匠, 三浦孝夫: Extracting Conditional Collocation, the International Workshop on Social Media Utilization Environment(SMUE), the 9th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 2013), Kyoto, Japan
- [6] 國田匠, 三浦孝夫: Conditional Collocation in Japanese, Australasian Document Computing Symposium (ADCS 2013), Brisbane, Australia
- [7] 三浦大輝, 三浦孝夫: "確率的 TF-IDF を用いた特徴語抽出と文 書検索." 2018 年情報処理学会全国大会, 2018, 東京, 6P-3