Vol.2009-DBS-148 No.24
Vol.2009-FI-95 No.24
2009/7/28

N-gram と離散型共起表現を用いたワードサラダ型スパム検出手法の提案

森本浩介† 片瀬弘晶† 山名早人††・†††

インターネット上にウェブページが爆発的に増加し，インターネットから得られる情報が重要になっている．しかし，ウェブページの爆発的な増加につれてスパム行為を行うページも同様に増加し，インターネットから得られる情報の価値を下げている．スパム行為には様々な手法があるが，本論文では自動的に文章を生成するワードサラダに着目し，ワードサラダ型のスパムを効率的に検出する手法を提案する．ワードサラダ型スパムを検出するため，n-gram と離散型共起表現を用いてカルバック・ライブラー情報量に基づく文章のスコアを計算し，計算したスコアに基づき判定を行う．提案手法の評価実験を行った結果，既存手法と比較して F 値で 0.18 の性能の向上を確認できた．

Proposal of Word Salad Spam Detection Method using N-gram and Interrupted Collocations

Kosuke Morimoto†
Hiroaki Katase† and Hayato Yamana††・†††

Information on the Internet becomes important because of exploding Web page. However, Spam pages also have exploded and information from the Internet have become lower reliability. Though there are many Spamming methods, in this article we focus on “word salad” that creates text automatically, and we propose the effective method of word salad detection. We detect word salad by the score based on Kullback-Leibler divergence calculated with n-gram and interrupted collocation. As a result of experiment, our method improves 0.18 points in F-value from the existing method.

1. はじめに

ワードサラダとは機械が自動的に生成した文章のことである．インターネット上には人間が作ったものでない，自動的に生成された文章で構成されたウェブページが存在する．特に，最近はブログやソーシャルネットワーキングサイトなどの，誰でも簡単にウェブページを作成できるサービスを利用する形で，ワードサラダを作成する人が増えたため，自動的に生成された文章が増加している．自動的に文章を生成してウェブページを作成する目的は主に検索エンジンにおけるランキングを不正に上昇させようとする目的で行われる．その結果，ワードサラダでないページのランキングが下がることになり,インターネット上で得られる情報の信頼性などが低下する.従って，このような文章のページを検出することはインターネット上のウェブページの信頼性を向上させることにつながる．

自動的に文章を生成する手法は様々なものがある．[1]によると，よく検索される単語を並べる手法や，有名なページの文章をコピーするだけの手法など単純な手法のほかに，N 階マルコフ連鎖モデルを用いて文章を生成する手法がある．

N 階マルコフ連鎖モデルとは N 個の連続した系列から N+1 番目に出現する事象が確率的に決定するモデルであり，特に N＝１の時を単純マルコフ連鎖モデルという．

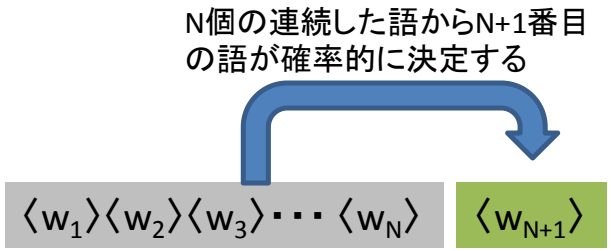

図 1 N 階マルコフ連鎖モデル

N 階マルコフ連鎖モデルに基づいて自動的に生成された文章は，構造上が正しい傾向にあり既存の手法では検出が難しい．また，文章をコピーする手法と異なり，生成された文章がオリジナルであることが検出をより困難にしている．図 2 は実際に 3 階

† 早稲田大学大学院 基幹理工学研究科
Graduate School of Fundamental Science and Engineering, Waseda University
†† 早稲田大学理工学術院
Science and Engineering, Waseda University
††† 国立情報学研究所
National Institute of Informatics

Vol.2009-DBS-148 No.24
Vol.2009-FI-95 No.24
2009/7/28

マルコフ連鎖モデルに基づき生成したワードサラダである．

東証第 1 部の音ですので, 今日もパパっとさばける料理上手でおしゃれな空間へ BARKS は、これは「ポニョ」でした朝から晩までずっと一緒にクエに行けます。2 階のトイレで泣きましたぁーてーウインドミル投球練習をして下さい私は頭でっかち。留学生らのときわ。2009 円＆送料無料スペシャル！イワです！

図 2　3 階マルコフ連鎖モデルによるワードサラダ

このように隣接した部分では意味が繋がっているが，少し離れた部分ではまったく意味が繋がらなくなっている．しかし，文法的には「てにをは」などのつながりはあまりおかしくない．このことが既存手法では検出を難しくしている．

本論文では，従来手法である n-gram を単独で用いたワードサラダの検出手法[1]の弱点を解消するための新しい手法を提案する．提案手法では，N 階マルコフ連鎖モデルに基づいて生成されたワードサラダの「N 番目に生起する語は直前の N-1 個の語によってのみ決定する」という特徴に着目し，N+1 個以上離れた語の関連の強さを検出することでワードサラダ型スパムを検出する．

具体的には大きな文書集合が与えられた時, 解析により形態素単位の n-gram と離散型共起表現が得られる．ここで，文書集合中の出現頻度が高い n-gram と離散型共起表現は文書的に自然な表現であると考えられ，逆に出現頻度が低い表現は不自然なものであると考えられる．N 階マルコフ連鎖モデルによって生成されたスパムでは，直前の N-1 個の語の列の情報に基づいてのみ次に生起する語が決定する．従って，ある語 W と W からの位置が N 番目より前の語との関係性については考慮されていないと考えられる．従って，ワードサラダが N 階マルコフ連鎖モデルに基づいて生成されていた場合，$N+1 \leq n$である n-gram と，離散型共起表現を用いればマルコフ連鎖モデルに基づいて生成されるスパムを検出することができる．そこで，大きな文書集合を解析して得られた n-gram と離散型共起表現の出現頻度に基づいて文章を形態素解析して得られる n-gram と離散型共起表現についてスコアを計算し，その計算した n-gram と離散共起表現のスコアに基づいて文章全体のスコアを計算することで，入力された文章がワードサラダ型スパムかそうでないかの判断を行う．本手法を用いてワードサラダ型スパムの検出について実験した結果，従来手法に比べて F 値において最大 0.18 の性能の向上を確認できた．

2. 関連研究

2.1 カルバック・ライブラー情報量によるワードサラダ検出手法[1]

2008 年に Orange Labs / ENST Paris の Lavergne らは n-gram とカルバック・ライブラー情報量を用いた自動生成文の検出手法を提案している．[1]は n-gram とカルバック・ライブラー情報量をもとにした文章のスコアを計算し，ワードサラダの判定を行う．

2.1.1 n-gram と n-gram 言語モデル[2]

n-gram とは文字列 S 中の長さ n の連続した部分文字列のことである．例えば「abracadabra」という文字列を文字単位の 3-gram で切り出すと図 3 のようになる．

n-gram は何を単位とするかで，切り出し方が変わってくる．一般的に文字を単位とするか，単語を単位とするかのいずれかで n-gram を切り出すことが多い．単語を単位とする場合，英語の場合は基本的に文章を構成する単語と単語の間がスペースで区切られて記述されるので，スペースを基準として単語を切り出せばよい．日本語のような単語と単語の間にスペースを置かない言語で書かれた文章を単語単位で n-gram を切り出す場合は，事前に形態素解析を行う必要がある．

abr
bra
rac
aca
cad
ada
dab
abr
bra

図 3　3-gram の例

また，n-gram 言語モデルとは n-1 個の文字列から n 番目の語を推定するモデルである．このモデルは機械翻訳や音声認識で広く用いられ，情報検索の分野でも用いられる．例えば，n=3 のとき，長さが 3 以上の文字列 $w_1 \cdots w_k$ は

$$p(w_1 \cdots w_k) = p(w_1)p(w_2 \mid w_1) \cdots p(w_k \mid w_{k-2} w_{k-1}) \quad (1)$$

の確率で生起する．p(w|h)は h という文字列（単語列）から文字（単語）w が生起する条件付き確率である．

2.1.2 カルバック・ライブラー情報量

カルバック・ライブラー情報量(Kullback-Leibler divergence)は相対エントロピーとも

Vol.2009-DBS-148 No.24
Vol.2009-FI-95 No.24
2009/7/28

呼ばれ，2 つの確率分布 P と Q の差の尺度であり，

- P と Q が離散確率分布のとき

$$D_{KL}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \quad (2)$$

- P と Q が連続確率分布のとき

$$D_{KL}(P \| Q) = \int_x P(x) \log \frac{P(x)}{Q(x)} \quad (3)$$

という式で定義される．相対エントロピーは常に非負であり，$D_{KL}(P\|Q)=0$ となるのは P=Q のときのみである．n-gram は先頭要素と末尾要素に密接な関係がある．例えば，「犬猿の仲」は形態素解析を行うと「犬猿」「の」「仲」と分解される．「の」は多数の語句と接続可能だが，「犬猿」という言葉の後に「の」が来た場合は「仲」という語が来る場合が非常に多いと考えられる．しかし，自動生成された文章ではこのような表現が崩れることがある．この特徴に基づき，カルバック・ライブラー情報量を計算する．

{p(· |h)}を n-gram 言語モデルの集合とする．h′をhから先頭を取り除いた文字列とする．このモデルについてカルバック・ライブラー情報量を次のように定義する．

$$KL(p(\cdot \mid h) \| p(\cdot \mid h')) = \sum_w p(w \mid h) \log \frac{p(w \mid h)}{p(w \mid h')} \quad (4)$$

式(4)は先頭が取り除かれたときにどの程度情報量が減ったかを示す尺度である．つまり，先頭の単語と末尾の単語の相関性を示す尺度となり，この値が大きければ相関性が強いということになる．また，p(w|h) = p(w|h′)のときに 0 となりこのときは情報量が減らなかったことになる．また，p(x|y)は y という文字列から x の生起する条件付き確率である．

2.1.3 スコアの計算方法

前節で述べたような先頭と末尾の文字の間の関係のスコアを計算するために，点カルバック・ライブラー情報量(pointwised Kullback-Leibler divergence)を定義する．これはhから生起する可能性のある全てのwについて個々に定義する．

$$PKL(h,w) = p(w \mid h) \log \frac{p(w \mid h)}{p(w \mid h')} \quad (5)$$

全てのwについて点カルバック・ライブラー情報量を足すとカルバック・ライブラー情報量となる．PKL(h, w)はhの先頭文字とwに強い依存関係がある場合に値が大きくなる．つまり，PKL(h, w)が小さくなるということはhの先頭文字とwにあまり関係がないということになる．このような小さな値に対してペナルティを与える．そのスコアは以下のように定義する．

$$S(h,w) = \max_v PKL(h,v) - PKL(h,w) \quad (6)$$

式(6)はhの先頭文字とwの弱い関係が多い場合に大きくなる．そして，最終的な文書Dのスコアは D中の全ての n-gram についてS(h, w)を計算し，その平均とする．

2.1.4 問題点

この手法は Google 1T 5-gram3)を用いて英語版 Wikipedia3)のスナップショット・English EU parliament proceedings を元に作ったワードサラダと通常の文書についてスコアを計算し，フィルタリングを行った結果を F 値で評価している．しかし，F 値は大きなばらつきがあり，最大で 0.98 と高い値を示している一方で 0.32 と，実験ごとに差がでている．これは，例えば 3-gram のモデルを用いてワードサラダを生成していた場合，3-gram についてのカルバック・ライブラー情報量はワードサラダでない文章とワードサラダの文章との間で差が表れないことを意味する．

2.2 ウェブページコンテンツ解析による検出手法 3)

2006 年にカルフォルニア州立大学の Alexandros らはウェブページのコンテンツを解析することでスパムを検出する手法を提案している．[5]では，ウェブページの内容から特徴を抽出し，機械学習を用いてスパムページを検出している．

2.2.1 特徴量

[5]では，ウェブページからページ中の単語数・ページタイトル中の単語数・単語の平均長・アンカーテキスト数・見えているコンテンツの割合・圧縮率・ページ中に一般的によく使われる単語が含まれる割合・一般的によく使われる単語の割合・独立な n-gram の割合・条件付き n-gram の割合の 9 個の特徴量を抽出している．

2.2.2 検出手法

前節で説明した 9 個の特徴量をもとに，機械学習アルゴリズムを用いて，ページがスパムかスパムでないかの分類を行っている．実験に用いられた機械学習アルゴリズムは決定木・ルールベース法・ニューラルネットワーク・サポートベクターマシンの 5 個である．これらのうち，最も分類精度がよかったアルゴリズムは決定木の C4.5 であった．C4.5 とは最も分類精度が高い特徴量から順番に木を構築するアルゴリズムである．また，さらに分類精度を上げるため，Bagging と Boosting を用いた．Bagging を用いることで精度は上昇し，Boosting を用いることで，Bagging よりもさらに精度が向上した．

2.2.3 問題点

[5]の手法は機械学習アルゴリズムを用いてスパムかスパムでないかの分類を行っているため，スパムページに使われる単語やスパムページの構成が変化すると再学習を行わなければならない．[6]によると，スパムはよく検索される語句を用いるため，

Vol.2009-DBS-148 No.24
Vol.2009-FI-95 No.24
2009/7/28

再学習を頻繁に行う必要があると考えられる．

2.3 Suffix Array によるコピーコンテンツ検出手法[5]

2008 年に総合研究大学院大学の竹田らは Suffix Array を用いたコピーコンテンツの検出手法を提案している．

2.3.1 Suffix Array

Suffix とは文字列 S のある位置から始まる，文字列 S の末尾までの部分文字列のことである．例えば「abracadabra」という文字列の Suffix は図 4 の左側ようになる．なお，文字列の左側の数字は各 Suffix が元の文字列 S の何文字目が開始位置であるかを示している．次に，Suffix Array とは文字列 S の Suffix を辞書順でソートした際のインデックスの配列である．従って，「abracadabra」の Suffix をソートすると図 4 の右側のようになる．図 4 からもわかるように「abracadabra」の Suffix Array は「11, 8, 1, 4, 6, 9, 2, 5, 7, 10, 3」となる．

1 : abracadabra	11 : a
2 : bracadabra	8 : abra
3 : racadabra	1 : abracadabra
4 : acadabra	4 : acadabra
5 : cadabra	6 : adabra
6 : adabra	9 : bra
7 : dabra	2 : bracadabra
8 : abra	5 : cadabra
9 : bra	7 : dabra
10 : ra	10 : ra
11 : a	3 : racadabra

図 4　Suffix Array の構築例

2.3.2 検出手法

まず，複数の文書から前節で述べた Suffix Array を構築する．Suffix Array の中に閾値以上の長さの一致領域(以下，コピー領域)がある場合，時系列順で見たときに後に作成された領域をコピー領域とみなす．また，文書中のコピー領域の割合(以下，コピー率)が閾値以上ならばその文書をコピー文書とみなす．

2.3.3 問題点

7)によると，ワードサラダ型スパムブログなどのコピーコンテンツが占める割合が低いスパムブログが検出できなかった．これはワードサラダ型スパムは短いフレーズを切り出して文章を構成しているために検出できなかったものと考えられる．

3. 提案手法

本節では，n-gram と離散型共起表現を用いたワードサラダ型スパムの検出手法について説明する．

3.1 提案手法の概要

1)では n-gram のみを用いて，カルバック・ライブラー情報量に基づくスコアの計算を行っている．n-gram はワードサラダスパムと人間が記述した文章の差異を，小さな粒度で比較する指標として用いることができるが，連続した領域の特徴しか評価することができない．つまり，<A><B><C><D>いう連続した語の列を 3-gram を用いて評価する場合<A><B><C>と<B><C><D>という順の共起が適切かどうかを評価することができても，<A>と<D>という離れて共起する表現が適切であるかどうかを評価できない．そこで，本手法では，連続した領域を評価できる n-gram に加え，離散した領域を評価できる離散型共起表現を用いて，文章がワードサラダか否かの判定を行う．

3.2 離散型共起表現

離散型共起表現とは，1 文中で離れた位置で共起する 2 つ以上の語句から成り立つ表現である．例えば「～いかにも～らしい」や「つまり～である」や「もし～ならば～である」というような表現である．このような表現は文中でどれだけ離れていても関係がある．したがって，n-gram では検出できない文の妥当性を評価できると考えられる．

なお, 本稿では, 離散型共起表現は 1 文中で完結しているものと仮定する. つまり, 複数の文章にまたがって共起する表現は考えないものとする．

3.3 スコアの計算方法

3.3.1 n-gram に基づくスコア

1)の手法ではスコアを算出するためのコストが大きく，規模の大きなデータを扱うことに不向きである．これは，S(h,w)を求めるときに max PKL(h,v)を求める必要があるため，PKL(h,v)をすべての v について求めなければならないからである．そこで，

$$ngramscore(h,w) = PKL(h,w) \tag{7}$$

を与えられた文書の各 n-gram について計算する．この手法ならば，1 つの w についてのみ求めればよいので計算コストが平均で 1/|v|になる．ここで，h とは長さが n の n-gram における長さが n-1 の接頭辞であり，h′ は長さが n-2 の h の接尾辞である．また，w は n-gram の末尾の語句である(図 5).

Vol.2009-DBS-148 No.24
Vol.2009-FI-95 No.24
2009/7/28

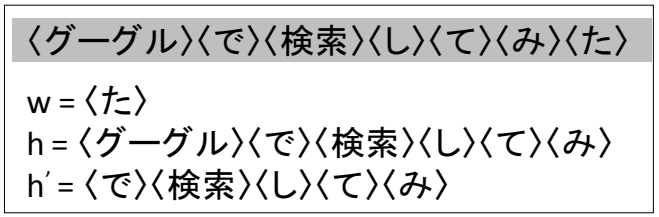

図 5　文の分割例

ngramscore は個々の n-gram についてのスコアであり，これを文章のスコアにするために全ての n-gram のスコアの平均を計算する．文章中の n-gram の集合を N とし，

$$avgngramscore = \frac{1}{|N|}\sum_{(h,w)\in N} ngramscore(h,w) \quad (8)$$

で表わされる avgngramscore を文章のスコアとして定義する．

3.3.2 離散型共起表現に基づくスコア

共起する表現は語句の出現順に意味があると考えられる．例えば，図 6 で言うと「もし」と「なら」の出現順が逆では意味がおかしくなってしまう．従って，離散型共起表現は生起する順番を考慮する．離散型共起表現も n-gram と同様に，初めに出現した語句と対応した語が後方で生起する．従って，離散型共起表現に基づくスコアもカルバック・ライブラー情報量を用いて計算する．前方で生起した語を w_a，後方で生起した語を w_b とすると，

$$collocationscore(w_a, w_b) = p(w_b \mid w_a)\log\frac{p(w_b \mid w_a)}{p(w_b)} \quad (9)$$

として離散型共起表現のスコア collocationscore を定義する．この値は大きいほど w_a と w_b の相関性が強く，w_a が生起した場合，その後方で w_b が生起する確率が高いということを示している．図 6 を用いて説明すると，〈もし〉という語が前方で生起しているならば，〈なら〉という語が後方で生起している確率が高いと考えられるが，〈ので〉という語は〈もし〉という語の後方で生起している確率は低いと考えられる．従って，この場合〈もし〉と〈なら〉の離散型共起表現のスコア collocationscore(“もし”, “なら”)が〈もし〉と〈ので〉の離散型共起表現のスコア collocationscore(“もし”, “ので”)よりもスコアが高くなる．

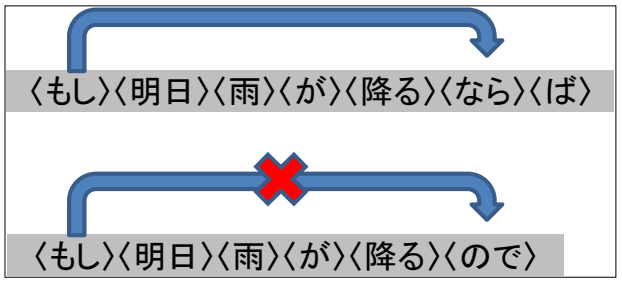

図 6　離散型共起表現の例

文章の離散型共起表現のスコアも n-gram のスコアと同様に文章中の全ての collocationscore の平均とする．文章中の離散型共起表現の集合を C とし，

$$avgcollocationscore = \frac{1}{|C|}\sum_{(w_b|w_a)\in C} collocationscore(w_b \mid w_a) \quad (10)$$

で表わされる avgcollocationscore を文章の離散型共起表現のスコアと定義する．

3.4 スコア計算の概要

文章のスコア計算の手順を説明する．

1. 文章を形態素解析し，分かち書きした形式にする．
2. 分かち書きした本文から n-gram と離散型共起表現を生成する．
3. 抽出した n-gram と離散型共起表現についてそれぞれのスコアを計算する．

4. 実験と評価

4.1 データセット

4.1.1 n-gram

本手法では n-gram データとして日本語版 Google n-gram3)を用いた．[8]は Web から抽出した日本語データから n-gram(1～7-gram)とその出現頻度をカウントしたデータである（表 1）．抽出対象は一般に公開されているウェブページで，Google がクロールしたものである．また，抽出は文ごとに単語を単位として行われている．日本語の文は単語間に空白が入らず語と語の境界が自明でないので，前処理として文を単語単位に分割する必要がある．文を分割するために形態素解析を行った．形態素解析には mecab-0.96[9]と mecab-ipadic-2.7.0-2007-0801[9]を用いた．

Vol.2009-DBS-148 No.24
Vol.2009-FI-95 No.24
2009/7/28

表 1　日本語版 Google n-gram

総単語数	255,198,240,937
総文数	20,036,793,117
1-gram	2,565,424
2-gram	80,513,289
3-gram	394,482,216
4-gram	707,787,333
5-gram	776,378,943
6-gram	688,782,933
7-gram	570,204,252

4.1.2 離散型共起表現

日本語版 Wikipedia[10]の 2008 年 11 月 27 日分のスナップショット(純記事数:545055)から以下の手順で離散型共起表現を抽出した.

1. [10]から本文データを抽出
2. 抽出した本文を形態素単位で分かち書き
3. 分かち書きした文から，1 文中で形態素間の距離が 2 以上である形態素の組み合わせを離散型共起表現として抽出する.
4. 3 で抽出した離散型共起表現の出現回数用いてカウントする.

この手法で抽出した離散型共起表現は 265,995,767 個であった.この中から出現回数が 20 回以上のもの抽出し，最終的に 10,904,371 個現をスコアの計算に用いた.

4.2 実験方法

収集したブログから 3 階・4 階マルコフ連鎖に基づくワードサラダをそれぞれ 1,000 個生成した．また，非ワードサラダとして収集したブログと[10]から，無作為に 1,000 個の文章を抽出した．ワードサラダと非ワードサラダをそれぞれ 1,000 個ずつ，合計 2,000 個について，以下の 5 種類の手法で検出実験を行う.

- 従来手法 1. 3gram
- 従来手法 2. 4gram
- 提案手法 1. 離散型共起表現
- 提案手法 2. 3gram と離散型共起表現
- 提案手法 3. 4gram と離散型共起表現

また，実際にスパムスコアの計算を行ったところ，n-gram のスパムスコアの分布と離散型共起表現のスコアの分布のスケールに差が出たので，2 つのスコアを足し合わせるときにスケールが小さい離散型共起表現の方に重みづけを行った．重みは n-gram に基づくスコアの平均と離散型共起表現に基づくスコアの平均の比である.

4.3 結果

既存手法と提案手法の再現率・適合率の関係，F 値，ROC 曲線を掲載する．なお，図 8 と図 10 において，再現率が下がったところで適合率も下がっているのは，閾値を大きくすると，閾値を超えなくなるワードサラダが増えていくが，さらに閾値を大きくすると，ワードサラダが閾値を超えなくなる割合より，ワードサラダでない文書が閾値を超えなくなる割合が大きくなるためである.

4.3.1 3 階マルコフ連鎖モデルに基づくワードサラダの検出実験

本節では 3 階マルコフ連鎖モデルに基づいて生成したワードサラダと非ワードサラダで検出実験を行った結果を示す．図 7 は各手法の再現率と適合率を表わしている．図 8 は各手法の ROC 曲線である．表 2 は各手法の F 値の最大値を表している.

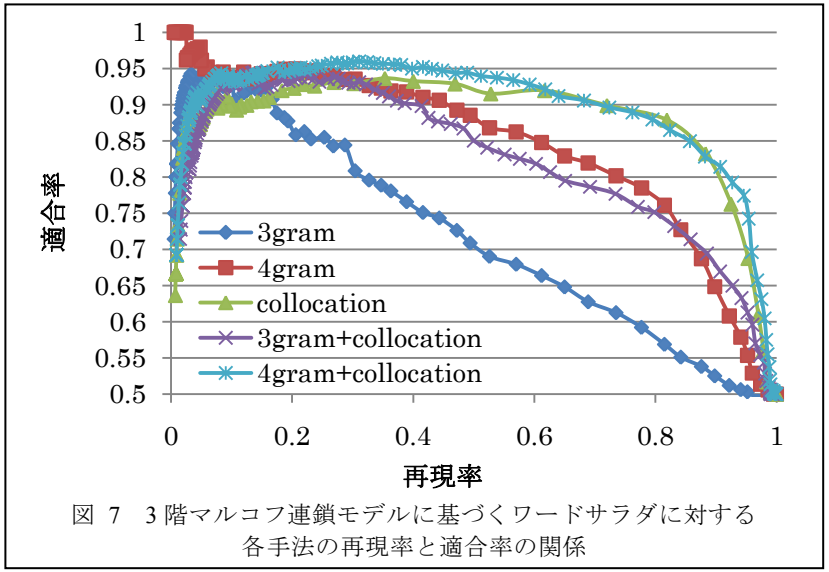

図 7　3 階マルコフ連鎖モデルに基づくワードサラダに対する各手法の再現率と適合率の関係

Vol.2009-DBS-148 No.24
Vol.2009-FI-95 No.24
2009/7/28

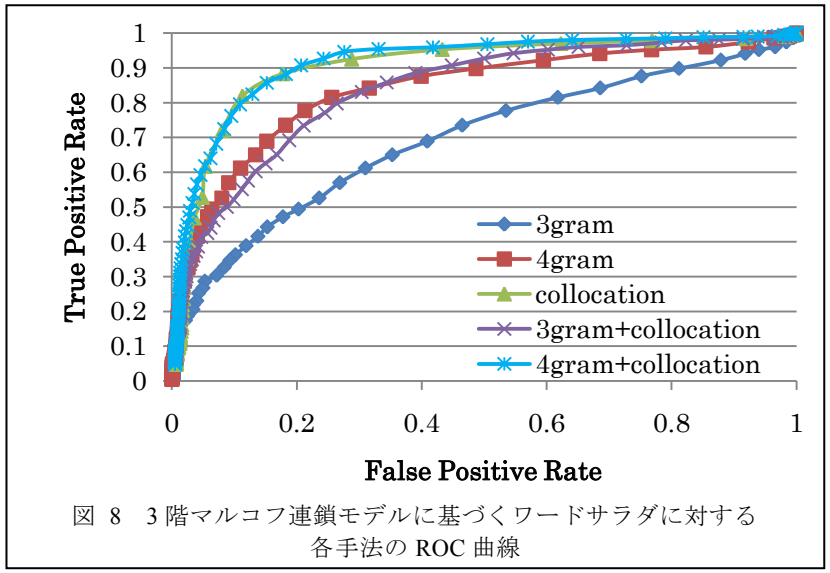

図 8　3 階マルコフ連鎖モデルに基づくワードサラダに対する
各手法の ROC 曲線

表 2　3 階マルコフ連鎖モデルに基づくワードサラダに対する
各手法の F 値の最大値

	F 値
3-gram	0.672145
4-gram	0.787059
collocation	**0.856866**
3-gram + collocation	0.779292
4-gram + collocation	**0.858089**

4.3.2 4 階マルコフ連鎖モデルに基づくワードサラダの検出実験

本節では 4 階マルコフ連鎖に基づいて生成したワードサラダと非ワードサラダで検出実験を行った結果を示す．図 9 は各手法の再現率と適合率を表わしている．は各手法の ROC 曲線である．表 3 は各手法の F 値の最大値を表している．

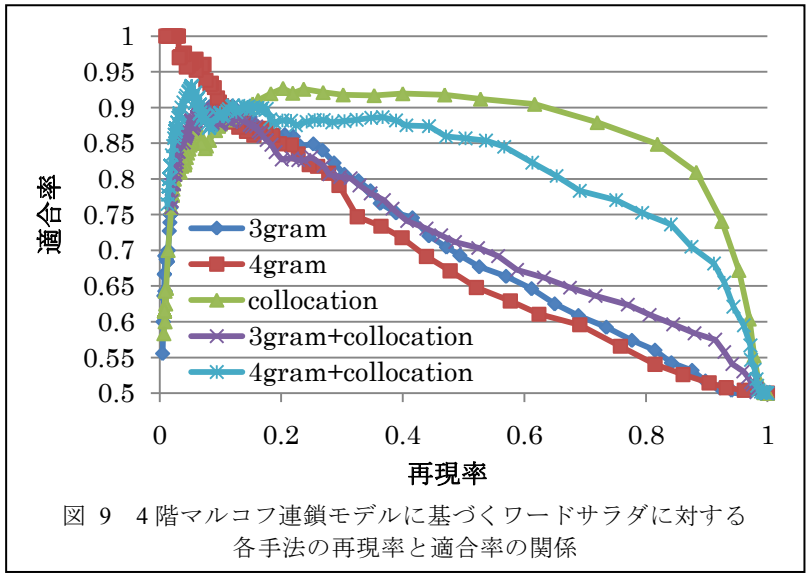

図 9　4 階マルコフ連鎖モデルに基づくワードサラダに対する
各手法の再現率と適合率の関係

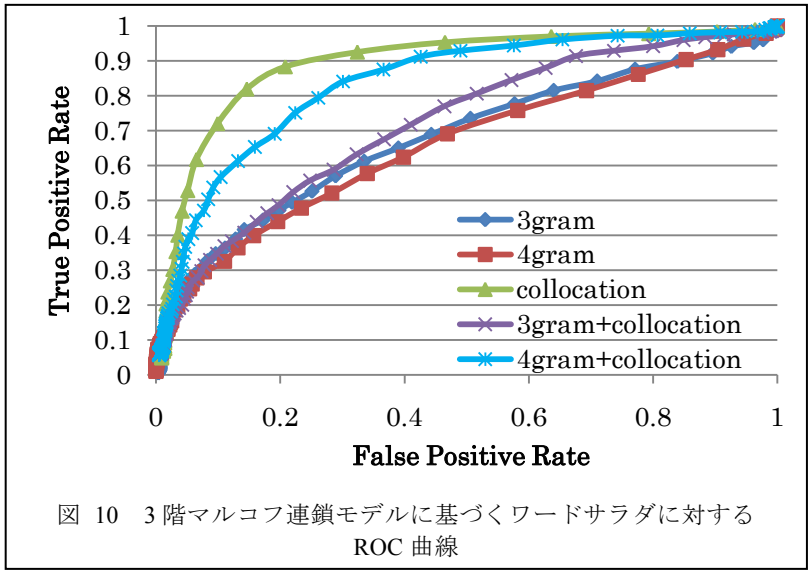

図 10　3 階マルコフ連鎖モデルに基づくワードサラダに対する
ROC 曲線

Vol.2009-DBS-148 No.24
Vol.2009-FI-95 No.24
2009/7/28

表 3　4 階マルコフ連鎖モデルに基づくワードサラダに対する
各手法ごとの F 値の最大値

	F 値
3-gram	0.666667
4-gram	0.666667
collocation	**0.844572**
3-gram + collocation	0.705792
4-gram + collocation	0.785247

4.4 評価

表 2・表 3 から，N 階マルコフ連鎖で生成したワードサラダを，n≤N である n-gram を用いて検出することは困難であるということが確認できた．

また，表 2・表 3 から N+1≤n である n-gram を単独で用いた場合(F 値：0.787059)，n≤N である n-gram を用いた場合(F 値：0.672145)よりも検出精度が良いことも確認できた．

図 8・図 10 から，離散型共起表現を用いた場合，単独でも精度よく検出することができるということが確認できた．また，表 2 と表 3 から，離散型共起表現は N に依存せずに検出することが可能であるということがわかった．

さらに，表 2 から N+1≤n である n-gram と離散型共起表現を組み合わせた場合，n-gram を単独で用いる場合よりも F 値でおよそ 0.18 の性能の向上が確認できる．また，図 8 から，ROC 曲線の下部の面積が大きいことからも，提案手法がワードサラダの検出に有効であることが確認できる．

5. おわりに

本稿では，n-gram と離散型共起表現に基づいたスコアを計算することによって，文章が人によって書かれたものと機械的に生成された文章の差異という点から，ワードサラダ型スパムブログを検出する手法を提案し，実験を行った．

N 階のマルコフ連鎖で生成されたワードサラダに対し，N+1≤n である n-gram と離散型共起表現を組み合わせることでワードサラダ型スパムブログを F 値で 0.858089 と精度よく検出することが可能である．また，離散型共起表現は N にあまり依存せず，ワードサラダ型スパムブログを検出することが可能であることが確認できた．従って，離散型共起表現はワードサラダ型スパムの検出に有効であるといえる．

謝辞　本論文を作成するに当たり山名早人研究室の諸先輩方のアドバイスに深く感謝いたします．本研究の一部は科学研究費補助金（基盤研究（B））「検索エンジンの信頼性解析」（課題番号 21300038）によるものである．

参考文献

1) Thomas Lavergne, Tanguy Urvoy and Francois Yvon : Detecting Fack Content with Relative Entropy Scoring, CEUR Workshop Proceedings, Vol.377 (ECAI'08 Workshop on Plagiarism Analysis, Authorship Identification and Near-Duplication Detection), pp. 27-31 November 2008
2) Christopher D. Manning and Schutze Hinrich, “Foundations of Statistical Natural Language Processing”, MIT Press, 1999
3) Thorsten Brants and Alex Franz: Web 1T 5-gram corpus version 1, Linguistic Data Consortium, Philadelphia, 2006
4) 英語版 Wikipedia: http://en.wikipedia.org/
5) Alexandros Ntoulas, Marc Najork, Mark Manasse. and Dennis Fetterly : Detecting Spam Web Pages through Content Analysis, WWW 2006, pp.83-93, Edinburgh, Scotland, May 23-26, 2006
6) 佐藤有記，宇津呂武仁，福原知宏，河田容英，村上嘉陽，中川裕志，神門典子 ： キーワードの時系列特性を利用したスパムブログの収集・類型化・データセット作成, Proceedings of Data Engineering Workshop, A10-2, pp.1-8, 2008
7) 竹田隆治，高須淳宏 ： 日本語 splog の現状と対策”，電子情報通信学会東京支部学生会研究発表会，p.241，2008
8) 工藤拓，賀沢秀人: Web 日本語 N グラム第 1 版，言語資源協会発行, 2008
9) MeCab: http://mecab.sourceforge.net/
10)　日本語版 Wikipedia: http://ja.wikipedia.org/

正誤表

ページ	行	正誤	
3	7	誤	英語版 Wikipedia3)
		正	英語版 Wikipedia[4]
3	13	誤	ウェブページコンテンツ解析による検出手法 3)
		正	ウェブページコンテンツ解析による検出手法[5]
3	14	誤	Alexandros
		正	Ntoulas
4	2	誤	Suffix Array によるコピーコンテンツ検出手法[5]
		正	Suffix Array によるコピーコンテンツ検出手法[7]
5	13	誤	日本語版 Google n-gram3)
		正	日本語版 Google n-gram[8]
7	3-4	誤	は各手法の ROC 曲線である.
		正	図 10 は各手法の ROC 曲線である.