語の出現間隔の分析に基づく単一文書からのキーワード抽出の試み

三浦 準也† 小原 佑斗† 吉田 光男† 梅村 恭司†

†豊橋技術科学大学 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: †miura.junya.pv@tut.jp, y173321@edu.tut.ac.jp, yoshida@cs.tut.ac.jp, umemura@tut.jp

あらまし Keyword extraction methods usually need document frequency. Therefore, these methods cannot extract keywords from a single document because a set of documents is necessary to calculate the document frequency. In this study, we propose a method to extract keywords from a single document not using document frequency but the recurrence interval of words in a document. Each word has a different distribution of counts of the recurrence interval in a document. Thus, we assumed that the quartile of the distribution can be used as a feature value. We define a score using the quartile as a keyword-likeness. We tried to extract keywords from some single documents such as a story, using our defined score. As a result, we found that the proposed method works as expected, and that recurrence intervals of words are useful for keyword extraction.

キーワード キーワード抽出、反復出現、語の出現間隔、語義分析

1. 序論

1.1. 研究の背景

キーワード抽出は，ドキュメントに含まれるキーワードを自動で抽出するタスクであり，情報検索において重要な技術の一つである。具体的にどの語がキーワードであるかを客観的に決めることは困難であるが，本研究においては，キーワードとはそのドキュメントを特徴づける語であると定義する。例えば登場人物名やタイトルに含まれる語などが考えられるが，少なくとも登場人物名はキーワードであるとして議論を進める。

文書には論文や新聞記事，小説など，様々な種類があるが，これらは 2 種類に分けることができる。一つは，新聞や辞典など，複数のドキュメントの集合として構成されるものである。もう一つは，小説や物語文など，全体で一つのドキュメントと言えるもので，これを単一文書と呼ぶ。

キーワード抽出の既存手法の多くは，ドキュメント集合からキーワードを抽出するものである。特に，ドキュメント頻度を用いる手法が主流である。例えば，逆ドキュメント頻度 (inverse document frequency) または IDF と呼ばれる値を用いる手法 [1] や，ドキュメント頻度および反復度を算出してキーワード抽出を行う手法 [2] がある。しかし，単一文書ではドキュメント頻度が計算できないため，このような手法を使うことができない。ところが，人間は一つのドキュメントだけを見てどの語がキーワードであるかを考えることが可能である。そのため，他のドキュメントと比較することなくキーワードらしさを定義することが可能なはずである。単一文書から語の特徴を分析する方法としては，言語情報を用いる方法や語の共起情報を用いる方法などが考えられるが，本研究では語の出現間隔に注目した。本論文では，語の出現間隔について分析し，単一のドキュメントからキーワードを抽出する手法を提案する。

1.2. 研究の目的

本研究の目的は，ドキュメント頻度を用いずに，ひとつのドキュメントからキーワードを抽出する手法を提案することである。本手法は，語の出現位置の偏りに注目し，出現間隔の分布の四分位数を用いることを特徴としている。

1.3. 関連研究

本論文で提案する手法は，ドキュメント内の語の出現位置の偏りに注目したものであるが，キーワードが出現するドキュメントの偏りに注目した研究は多く行われている。例えば，[3] では，あるドキュメントに単語が 1 回出現しているとき，その単語が 2 回以上出現している条件付き確率は，その単語がキーワードであるとき顕著に高いことを報告している。このことを利用したキーワード抽出の手法の一つに [2] がある。また，ドキュメント集合における大域的な語の繰り返しをモデル化することで，語の特徴を捉える試みも行われている [4] 。これらの研究は，主にドキュメント集合を対象とした分析である点が本手法と異なるが，語の繰り返しを分析することで語の特徴を捕えようとしている点が本手法と類似しており，本研究における主張の根拠となる。また，[5] はドキュメント内の語の出現間隔をモデル化しており，こちらはより直接的に本手法の理論的裏付けとなり得ると考えられる。

また，単一文書に対して適用できるキーワード抽出の手法として，文章中における語の前後の表現に注目して語の特徴を抽出する手法 [6] や，語の共起情報を

用いてキーワードを抽出する手法 [7] などがある。語の出現間隔に着目した手法としては，語の出現間隔の分散を用いてキーワードを抽出する手法 [8] があり，本手法のアイデアに近い。

2. 提案手法

この章では，語の出現間隔を用いてキーワードらしさを定義する提案手法について述べる。

2.1. 語の出現間隔の分布

図 1 のように，あるドキュメントにおいて同一の語が複数回出現するとき，それらの出現間隔がそれぞれ何文字分であるかを全て調べる。ここで語とは，単語もしくは単なる部分文字列である。語の出現回数を tf としたとき，語の出現間隔の数 rf は $\mathrm{tf}-1$ に等しくなる。語のi番目の出現から$i+1$番目の出現までの間隔を $k_i\ (i=1,2,\ldots,\mathrm{rf})$ とする。ただし，出現間隔は文字単位もしくはバイト単位とする。

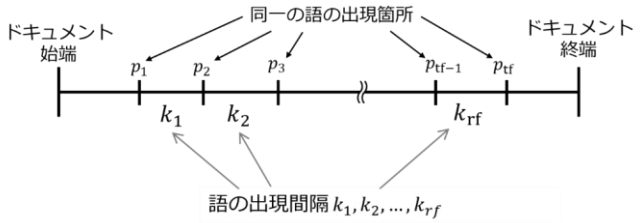

図 1　語の出現間隔

例として，*Alice in Wonderland*（Lewis Carroll 著）の全文における語の出現間隔を計算する。図 2 にいくつかの語の出現間隔の累積グラフを示す。語が文脈に関係なく出現すると仮定した場合のグラフも重ねて示している。グラフの横軸は語の出現間隔 d，縦軸は長さが d 以下の出現間隔の数を表す。語が文脈に関係なく出現するというのは，ドキュメントの始端から終端にかけて語の出現確率が独立でかつ一定であるということを表し，この場合の出現間隔の確率分布は幾何分布となる[1]。図 2 に注目すると，語によって出現間隔の分布が異なっていることが分かる。例えば，語“with”の出現間隔の分布は，語が文脈に関係なく出現すると仮定した場合の分布に近い形をしている。対して，語“Hatter”の出現間隔の分布は，語が文脈に関係なく出現すると仮定した場合の分布とは異なる形をしている。よって，キーワードは文脈に応じて出現間隔が変化するが，冠詞や接続詞などそれ単体では意味を成さない単語は文脈に関係なく出現するという予想が立てられる。

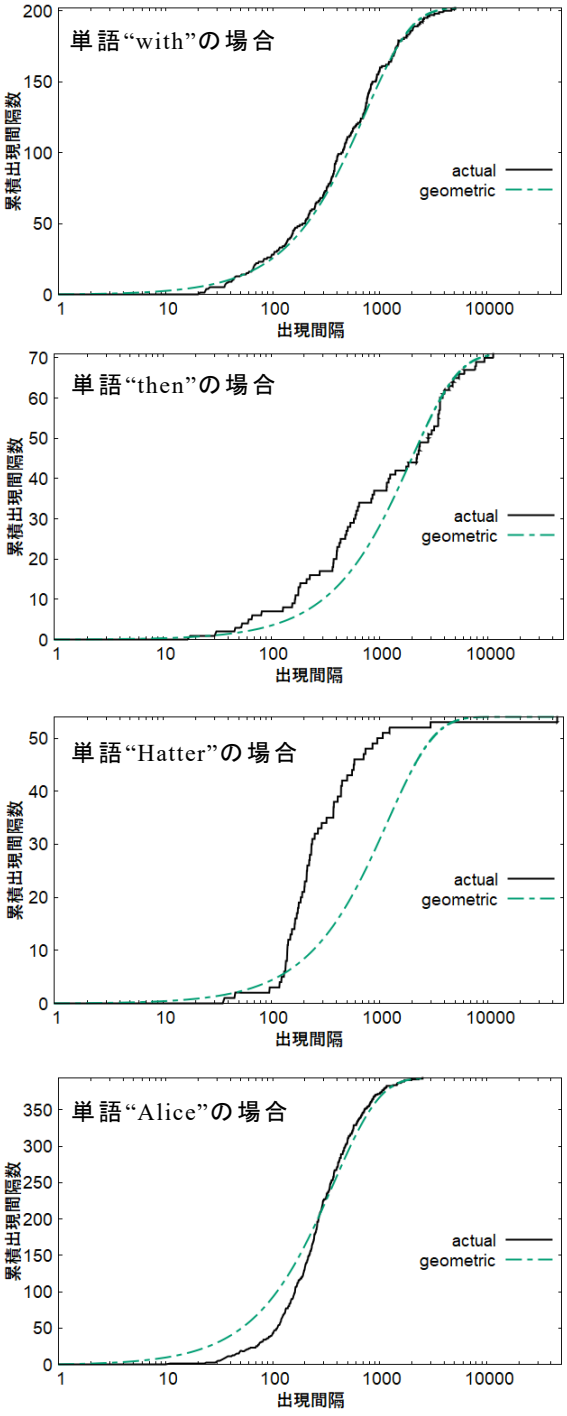

図 2　*Alice in Wonderland* における単語“with”，“Hatter”，“then”および“Alice”の出現間隔の分布（実線）と，語が文脈に関係なく出現すると仮定した場合の出現間隔の分布（鎖線）

出現間隔の分布の特徴的なところに注目するため，四分位数を考える。q_1を語の全ての出現間隔のうち特に短い 25%を除き最短の出現間隔，同様に，q_3を特に長い 25%を除き最長の出現間隔とする。厳密には，四

[1] 語が文脈に関係なく出現すると仮定した場合の語の出現間隔の累積グラフは，$p=1/\bar{k}$ の幾何分布の累積分布関数に rf を掛けて出現間隔数の関数としたものである。

分位数の計算方法についてはいくつかの定義が知られているが，本論文においては値を内分して計算する定義を用いる。

図 3 に，*Alice in Wonderland* における単語“then”と“Hatter”の出現間隔の分布とその四分位数 q_1 と q_3 を示す。単語“Hatter”の出現間隔の分布は，単語“then”の出現間隔の分布よりも対数グラフにおける中央付近の傾きが大きく，特に 100~300 字程度の出現間隔が多い。単語“Hatter”の出現間隔の分布の四分位数を見ると，q_1 の値がおよそ 150 であり，特に多い出現間隔の値を取り出すことができている。

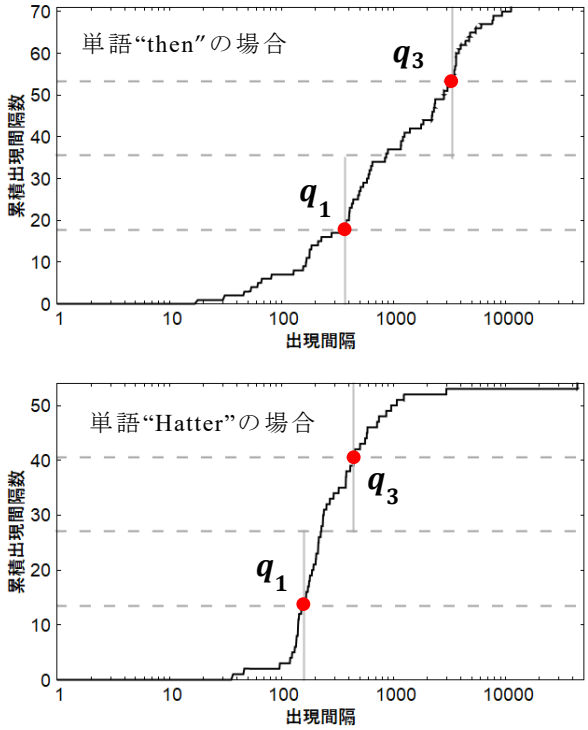

図 3 *Alice in Wonderland* における単語“then”と“Hatter”の出現間隔の分布とその四分位

また，語の出現間隔のうち極端に短いもの，および極端に長いものは，語の出現間隔の分布の主な特徴から外れた値であると考えることができる。例えば，物語文においてある人物が登場し，場面が変わってその人物が言及されなくなった後，しばらくしてから再登場する場合を考える。この人物が言及されていない間の出現間隔は，言及されている間の出現間隔よりも極端に長くなるが，この長さに語の特徴が含まれているとは考えにくい。また，極端に短い出現間隔についても，例えばある登場人物名を連呼するような表現であることが考えられるが，そのような例は基本的な文型に当てはまらない特殊な表現である。よって，語の出現間隔のうち，特に短い出現間隔と特に長い出現間隔は考慮する必要のない外れ値であり，四分位を計算することでこの影響を排除することができる。分布の平均や分散は極端な値の影響を受けてしまうため，本手法では平均や分散を用いるアプローチをとらない。

2.2. キーワードの性質についての仮定と特徴量

キーワードらしさを定義するために，キーワードの性質について 3 つの仮定をおき，それぞれについて特徴量を考えた。

- 仮定 1　キーワードは出現しやすい：$\log \mathrm{rf}$
- 仮定 2　キーワードの出現は離れすぎない：$\log \frac{3l}{\mathrm{rf} \cdot q_3}$
- 仮定 3　キーワードの出現は近すぎない：$\log q_1$

以下でそれぞれの仮定と特徴量について説明する。

2.2.1. 仮定 1「キーワードは出現しやすい」について

一般に，重要な語は出現頻度が高く，語の出現頻度 tf はキーワードらしさの指標としてよく用いられる。よって，語の出現頻度 tf，または語の出現間隔の数 $\mathrm{rf} = \mathrm{tf} - 1$ が大きいほどキーワードらしいといえる。また，語の出現頻度 tf をそのままキーワードらしさとして用いると，出現頻度の高い語に対して過大な重みを与える傾向があることが知られており，出現頻度の高い語の影響を軽減するために出現頻度の対数による重みづけがよく用いられる [1]。よって，仮定 1 に基づくキーワードらしさの特徴量を，語の出現間隔の数の対数 $\log \mathrm{rf}$ とした。

2.2.2. 仮定 2「キーワードの出現は離れすぎない」について

仮定 2 を言い換えると，キーワードは一度出現するとその付近で再出現しやすいということである。例えば，物語文における一場面である人物が登場しているとき，その登場人物の名前は連続して何度も出現することが予想できる。よって，ある語の出現間隔がまんべんなく小さければキーワードらしいと言えるのではないかと考えた。ただし，物語文の例における人物が言及されていない間の出現間隔は除外する。語の出現間隔の四分位数を用いると，語の全ての出現間隔のうち特に長い 25%を除き最長の出現間隔 q_3 が小さいほどキーワードらしいと考えることができる。

しかし，q_3 の値は出現間隔の数およびドキュメントの文字数によってバイアスがかかる。図 4 (a) に実際の文章における全単語の rf と q_3 を計算した散布図を示す。q_3 の値は，語の出現間隔の数 rf が大きくな

るほど最大値が小さくなる傾向にあり，散布図の点は，図中に示した $\mathrm{rf}\cdot q_3 = 3l$ の直線より下に位置することが予想できる。実際にいくつかの文書において全単語の $\mathrm{rf}\cdot q_3$ を計測したところ，$\mathrm{rf}\cdot q_3$ の値は $3l$ を上回らなかった。よって，$\mathrm{rf}\cdot q_3$ の値は高々 $3l$ であるという予測を立てた。

以上より，q_3の値が小さいほど大きくなり，かつ出現間隔の数およびドキュメントの文字数の影響が小さく，0 よりも小さくならないキーワードらしさの特徴量として，$\log\frac{3l}{\mathrm{rf}\cdot q_3}$ という式を考えた。図 4(a)と図 4(b)を見比べると，この特徴量は $\mathrm{rf}\cdot q_3$ の値が $3l$ に近いほど小さくなることが分かる。

2.2.3. 仮定 3「キーワードの出現は近すぎない」について

例えば，ある登場人物名が文中に出現し，直後にその人物が再び言及されるとき，その登場人物名の代わりに代名詞が用いられることが多いはずである。しかし，“then”や“with”など，名詞以外の語は代名詞に変化することはない。よって，ある語の特に短い出現間隔に注目したとき，その長さがある程度長ければ登場人物名などのキーワードである可能性が高いと考えた。ただし，ある登場人物名を連呼するような例外的な表現については除外する。語の出現間隔の四分位数を用いると，語の全ての出現間隔のうち特に短い 25%を除き最長の出現間隔 q_1 が大きいほどキーワードらしいと考えることができる。よって，仮定 3 に基づくキーワードらしさの特徴量を$\log q_1$ とした。

2.3. キーワードらしさの定義

まず，仮定 1 と仮定 2 から考えた特徴量の積をキーワードらしさのスコア $Score_2$ として定義する。

$$Score_2 = (\log \mathrm{rf})\left(\log\frac{3l}{rf\cdot q_3}\right) \quad (1)$$

さらに，仮定 1，仮定 2，仮定 3 から考えた特徴量の積をキーワードらしさのスコア $Score_3$ として定義する。

$$Score_3 = (\log \mathrm{rf})\left(\log\frac{3l}{\mathrm{rf}\cdot q_3}\right)(\log q_1) \quad (2)$$

Alice in Wonderland において，全単語に対して 3 つの特徴量 $\log \mathrm{rf}$, $\log\frac{3l}{\mathrm{rf}\cdot q_3}$, $\log q_1$ を計算した散布図を図 5 に示す。キーワードの典型例である登場人物名が中央付近に集まっていることから，3 つの特徴量がそろって大きい語はキーワードらしいと言ってよいと考える。

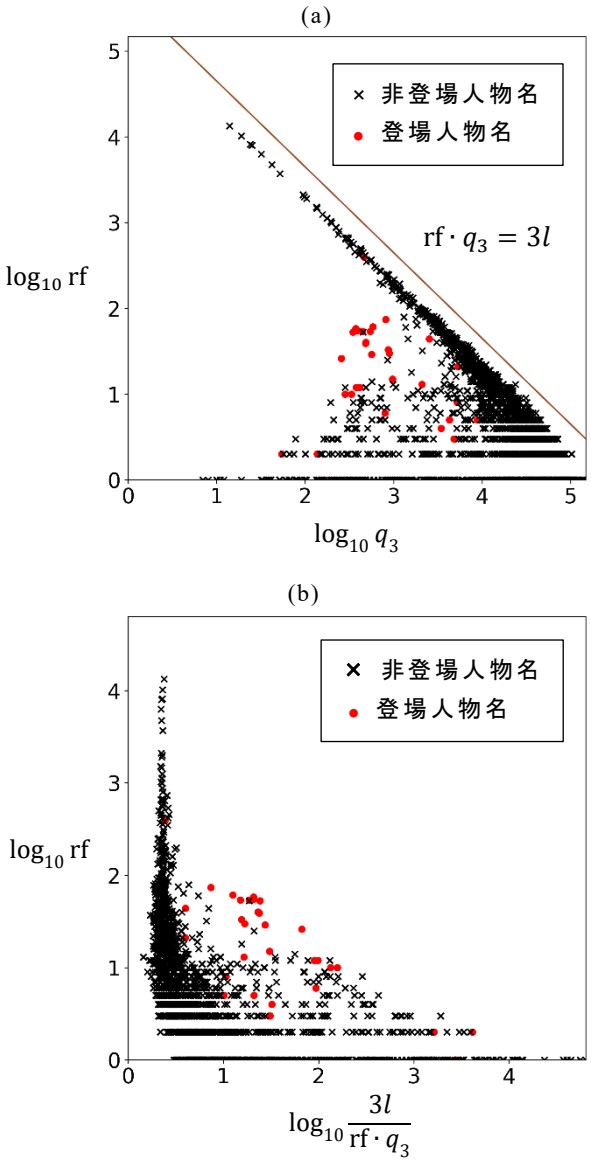

図 4 *Alice in Wonderland* における全単語に対して $\log \mathrm{rf}$ と $\log q_3$ を計算した散布図（a）と，$\log \mathrm{rf}$ と $\log\frac{3l}{\mathrm{rf}\cdot q_3}$ を計算した散布図（b）

また，*Alice in Wonderland* において，全部分文字列に対して 3 つの特徴量を計算した散布図を図 6 に示す。いずれかの特徴量が 0 に近くなる部分文字列が存在することが分かる。こういった部分文字列はキーワードらしくないと言えるが，それぞれの仮定よりなぜキーワードらしくないかを図中に示したように解釈することができる。$\log \mathrm{rf}$ が小さい部分文字列は出現頻度が低すぎるため，$\log\frac{3l}{\mathrm{rf}\cdot q_3}$ が小さい部分文字列は出現間隔が長すぎるため，$\log q_1$ が小さい部分文字列は出現間隔が短すぎるため，キーワードらしくないと言える。

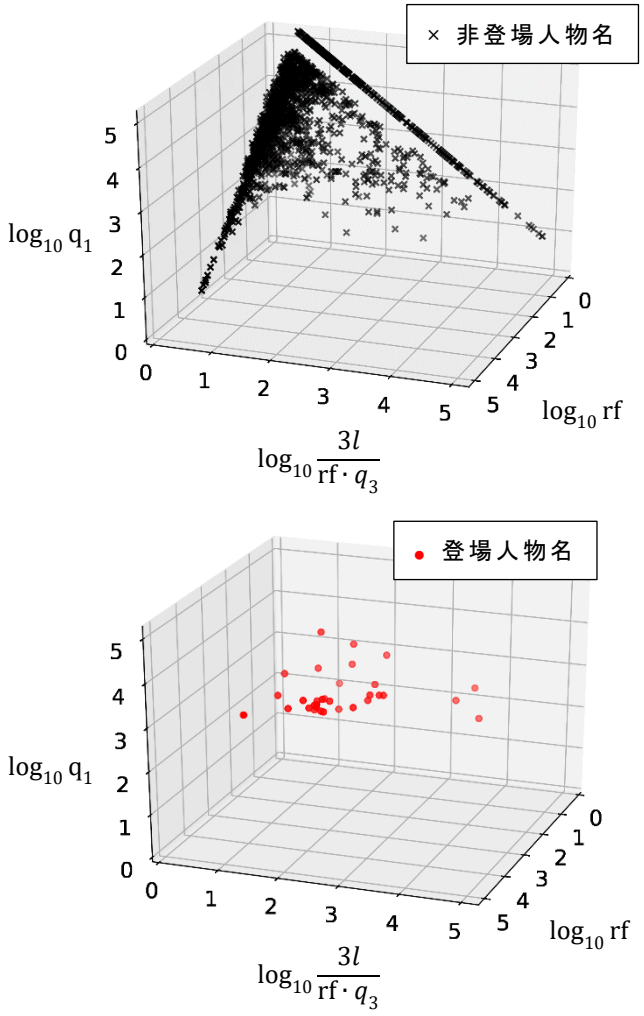

図 5　*Alice in Wonderland* における全単語に対して 3 つの特徴量 $\log \mathrm{rf}$ と $\log \frac{3l}{\mathrm{rf} \cdot q_3}$ と $\log q_1$ を計算した散布図

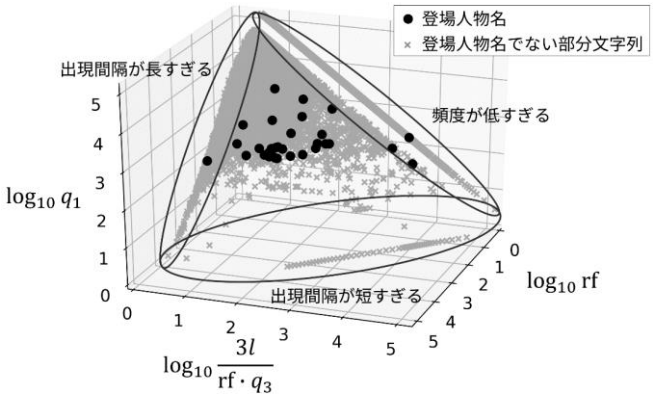

図 6　*Alice in Wonderland* における全部分文字列に対して 3 つの特徴量 $\log \mathrm{rf}$ と $\log \frac{3l}{\mathrm{rf} \cdot q_3}$ と $\log q_1$ を計算した散布図とその解釈

3. 実験

この章では，提案手法を用いてキーワード抽出ができることを確認するために行ったキーワード抽出実験について述べる。

3.1. 実験方法

提案手法の章で定義したキーワードらしさのスコア $Score_2$ および $Score_3$ を用いて，実際の文書に含まれる語に対してスコア付けを行い，スコアの降順でソートする。

まず，文書内の全ての英単語をスコア付けの対象として実験を行う。使用データは以下の通りである。

- *Alice in Wonderland*（Lewis Carroll 著）
  全文　英語　148,547 字
- 欽定訳聖書（The King James Bible）
  全文　英語　4,332,557 字

英単語の抽出方法は，その文書中において非アルファベット文字に囲まれた部分文字列を全て抽出した。ただし，語の出現間隔の分析においては，非アルファベット文字に囲まれているかに関係なく出現箇所を検索した。

続いて，文書内の全ての部分文字列をスコア付けの対象として実験を行う。使用データは以下の通りである。

- *Alice in Wonderland*（Lewis Carroll 著）
  全文　英語　148,547 字
- 『吾輩は猫である』（夏目漱石著）
  全文　日本語　321,976 字　960,158 バイト
- 『人間失格』（太宰治著）
  全文　日本語　73,899 字　219,981 バイト

ただし，実験結果においては，全ての出現位置が同じ部分文字列を同一視し，そのうち最も長い部分文字列（極大部分文字列）のみを表示する。全ての出現位置が同じ部分文字列は，その出現間隔の分布が完全に一致するため，全ての極大部分文字列に対してスコアを計算するだけで実験を行うことができる。接尾辞配列と呼ばれるデータ構造を用いると，全ての極大部分文字列に対して効率よく計算を行うことができる [9] [10]。

全ての実験において，使用データに対する前処理は基本的に行わない。ただし，元の文献に含まれるルビなどの特殊記法は適宜削除している。大文字小文字を区別し，改行は一文字として扱う。語の出現間隔はバイト単位で計算する。

3.2. 実験結果

3.2.1. *Alice in Wonderland* 全単語

表 1 に，*Alice in Wonderland* における全単語を対象としたキーワード抽出結果を示す。$Score_2$ を用いた結果と $Score_3$ を用いた結果のどちらにおいても，登場人物名が上位にきていることが分かる。また，“Soup”など，登場人物名ではないが重要と考えられる語が上位にきていることが分かる。

3.2.2. 欽定訳聖書 全単語

表 2 に，欽定訳聖書における全単語を対象としたキーワード抽出結果を示す。$Score_2$ を用いた結果と $Score_3$ を用いた結果のどちらにおいても，登場人物名が上位にきていることが分かる。

3.2.3. *Alice in Wonderland* 全部分文字列

表 3 に，*Alice in Wonderland* における全部分文字列を対象としたキーワード抽出結果を示す。対象が文字列であるので,空白文字,改行文字も分析対象になり，結果に出ることがある。これらはそれぞれ下線付きの空白，矢印で表示する（表 4，表 5 も同様)。$Score_2$ を用いた結果では，空白文字のみの部分文字列が上位を占めていることが分かる。対して， $Score_3$ を用いた結果では，登場人物名を含む部分文字列がいくつか上位にきていることが分かる。

3.2.4. 『吾輩は猫である』全部分文字列

表 4 に,『吾輩は猫である』における全部分文字列を対象としたキーワード抽出結果を示す。$Score_2$ を用いた結果と $Score_3$ を用いた結果のどちらにおいても,登場人物名を含む部分文字列がいくつか上位にきていることが分かる。また,「ヴァイオリン」など，登場人物名ではないが重要と考えられる語が上位にきていることが分かる。

3.2.5. 『人間失格』全部分文字列

表 5 に,『人間失格』における全部分文字列を対象としたキーワード抽出結果を示す。$Score_2$ を用いた結果と $Score_3$ を用いた結果のどちらにおいても，登場人物名を含む部分文字列がいくつか上位にきていることが分かる。また,「薬」など，登場人物名ではないが重要と考えられる語が上位にきていることが分かる。

表 1 *Alice in Wonderland* における全単語を対象としたキーワード抽出結果（登場人物名に下線）

rank	$Score_2$	$Score_3$
1	Caterpillar	Caterpillar
2	Cat	Mock
3	Turtle	Turtle
4	Mock	Duchess
5	Soup	Pigeon
6	Beau	Cat
7	Hatter	Pig
8	Pigeon	Hatter
9	Dormouse	Footman
10	mouse	Mouse
11	Duchess	serpent
12	Pig	King
13	Footman	Dodo
14	Dodo	moral
15	Mouse	hedgehog

表 2 欽定訳聖書における全単語を対象としたキーワード抽出結果（登場人物名に下線）

rank	$Score_2$	$Score_3$
1	duke	charger
2	suburbs	Abimelech
3	begat	precepts
4	curtain	Samson
5	Abram	Hezekiah
6	board	Gideon
7	repaired	Absalom
8	Absalom	Hezeki
9	Saul	Pau
10	Abimelech	Amaziah
11	Haman	Paul
12	cubit	Haman
13	Esther	Saul
14	Esau	Abram
15	boar	repaired

表 3 *Alice in Wonderland* における全部分文字列を対象としたキーワード抽出結果（登場人物名に下線）

rank	$Score_2$	$Score_3$
1	␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣	↲␣␣␣␣
2	␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣	↲␣␣␣
3	␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣	␣the␣Cat
4	␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣	the␣Cat
5	␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣	aterpillar
6	␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣	he␣Caterpillar
7	␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣	llar
8	␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣	lla

9	␣ ␣	he␣Cat
10	␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣	␣the␣Caterpillar
11	␣ ␣	␣Caterpillar
12	␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣	↲␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣
13	␣ ␣	␣ ␣ said ␣ the ␣ Caterpillar
14	␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣	ootm
15	␣ ␣ ␣ ␣ ␣ ␣ ␣	↲␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣ ␣

表 4 『吾輩は猫である』における全部分文字列を対象としたキーワード抽出結果（登場人物名に下線）

rank	$Score_2$	$Score_3$
1	i	陰士
2	に心を	蝉
3	␣	個性
4	フフ	地蔵
5	大和	多々良
6	液	禿
7	大和魂	多々良君
8	候	保険
9	舐め	陰士は
10	冊	ァイ
11	蝉	オリ
12	の松	ヴァイオリン
13	陰士	鼻子は
14	崎	鼻子
15	ダム	崎

表 5 『人間失格』における全部分文字列を対象としたキーワード抽出結果（登場人物名に下線）

rank	$Score_2$	$Score_3$
1	↲↲	薬
2	名詞	ツネ子
3	詞	の絵
4	アント	竹一
5	ツネ子	アント
6	アン	私は
7	トラ	ツ
8	…	ヒラメ
9	薬	、竹一
10	鮨	である。
11	ツ	奥
12	に訴え	の運動
13	個人	あなたの
14	のアント	鮨
15	ヒラメ	、シヅ子

3.3. 考察

前節に示した実験結果について考察する。

3.3.1. 英語の文章における全単語からのキーワード抽出結果について

Alice in Wonderland および欽定訳聖書における全単語からのキーワード抽出結果をまとめると，これらの文章に含まれる単語から登場人物名を抽出することができていると言える。また，登場人物名ではないが重要と考えられる語も少なからず抽出できている。

これらの文章における全単語からのキーワード抽出では，$Score_2$ と $Score_3$ で抽出される語に多少の変化はあるものの，キーワードを抽出する性能においては大きな差はないと考える。

以上より，$Score_2$ または $Score_3$ を用いて，英語の文章の全単語から登場人物名および登場人物名以外の重要な語をある程度抽出することができると考える。

3.3.2. 英語の文章における全部分文字列からのキーワード抽出結果について

Alice in Wonderland における全部分文字列を対象としたキーワード抽出結果では，$Score_2$ による結果と $Score_3$ による結果とで大きな差があることが確認できる。$Score_3$ による結果ではある程度キーワードが含まれる部分文字列を抽出できているのに対し，$Score_2$ による結果は空白文字が連続する意味のない部分文字列が上位を占めており，キーワード抽出ができていない。この結果は，同じ文字が連続するだけの部分文字列において，仮定 1 と仮定 2 に基づく特徴量である $\log \mathrm{rf}$ と $\log \frac{3l}{\mathrm{rf} \cdot q_3}$ が大きい値を取り，仮定 3 に基づく特徴量である $\log q_1$ が 0 または 0 に近い値を取るためであると考える。実際，文書中に同じ文字が連続するだけの箇所があるとき，その内側の部分文字列は 1 文字分だけずれた再出現が多数存在する状態になることが容易に想像できる。出現間隔の長さが 1 ばかりになれば，q_1 および q_3 の値は 1 になる。よって，仮定 3 に基づく特徴量である $\log q_1$ は，同じ文字が連続する部分文字列を除外する効果がある。もともと $\log q_1$ は，登場人物名などのキーワードは代名詞に変化することが多いという考えに基づいた特徴量であるが，違った意味付けをすることができる可能性が示唆される。

また，$Score_3$ による結果に注目すると，登場人物名を含む部分文字列をある程度抽出できているが，語の区切りで分割されていない部分文字列が多いことから，キーワード抽出の性能としては難があると言える。ただし，似たような部分文字列が多く見られることから，そういった部分文字列をまとめることで性能を改善することができる可能性があるといえる。

以上より，*Alice in Wonderland* における全部分文字列を対象としたキーワード抽出においては，$Score_2$ よ

りも $Score_3$ を用いたほうが良い結果になり，$Score_3$ を用いると英文の全部分文字列から登場人物名を抽出することができる可能性があると考える。

3.3.3. 日本語の文章における全部分文字列からのキーワード抽出結果について

『吾輩は猫である』および『人間失格』における全部分文字列からのキーワード抽出結果をまとめると，これらの文章に含まれる部分文字列から登場人物名を含む部分文字列を抽出することができていると言える。また，登場人物名ではないが重要と考えられる語を含む部分文字列も少なからず抽出できている。これらの文章における全部分文字列からのキーワード抽出では，$Score_2$と $Score_3$ で抽出される部分文字列に多少の違いが見られる。$Score_3$ による結果に比べて，$Score_2$ による結果は，登場人物名を含む部分文字列が少なく，空白文字や記号など意味のない部分文字列が上位に存在することから，$Score_3$ による抽出結果のほうが優れた結果であると言える。

$Score_2$ と $Score_3$ のどちらの実験結果においても，語の区切りで分割されていない部分文字列も含まれていることから，キーワード抽出の性能としては難があると言える。しかし，一つの単語を抽出できている結果もあることから，日本語の文章の全部分文字列から単語を抽出することができる可能性がある。ただし，この結果は極大部分文字列のみを表示しているものであることに注意して考える必要がある。

以上より，$Score_2$ または $Score_3$ を用いて，日本語の文章の全部分文字列から登場人物名および登場人物名以外の重要な語を含む部分文字列をある程度抽出することができ，$Score_2$ よりも $Score_3$ を用いたほうが良い結果が得られると考える。

3.3.4. 考察のまとめ

以上の考察をまとめると，まず，英語の文章における全単語からのキーワード抽出結果から以下のことが考察できる。

- 英語の文章に含まれる単語から，登場人物名を抽出することができている。
- 英語の文章に含まれる単語から，登場人物名ではないが重要な語を抽出することができている。

よって，キーワード抽出がある程度できており，キーワード抽出に出現間隔を使うことができると考える。

また，日本語および英語の文章における全部分文字列からのキーワード抽出結果から以下のことが考察できる。

- $Score_3$ を用いて，日本語および英語の文章の全部分文字列から登場人物名を含む部分文字列を抽出することができている。
- $Score_2$ を用いると，日本語および英語の文章の全部分文字列から登場人物名を含む部分文字列を抽出できないことがある。
- $Score_3$ を用いると，文章の全部分文字列から登場人物名および単語を抽出することができる可能性が考えられる。

4. 結論

本研究では，大きさ順に並べた出現間隔の四分位数を用いて語のキーワードらしさを定義し，キーワード抽出を行う手法を提案した。提案した手法は先行研究と異なり，ドキュメントの区切りを必要としないことが特徴である。提案手法を用いることで，英語および日本語の文書からキーワード抽出がある程度できることを確認し，キーワード抽出に語の出現間隔を使うことができることを示した。

参　考　文　献

[1] 北研二，津田和彦，獅々堀正幹，情報検索アルゴリズム，共立出版, 2002.

[2] 武田善行，梅村恭司，“キーワード抽出を実現する文書頻度分析,” 計量国語学, Vol. 23, No. 2, pp. 65-90, 2001.

[3] K. W. Church, “Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p^2,” Proceedings of the 18th conference on Computational linguistics, Vol. 1, pp. 180-186, 2000.

[4] Y. Xu, K. Umemura, “Improvements of Katz K Mixture Model,” Information and Media Technologies, Vol. 9, No. 4, pp. 604–629, 2014.

[5] A. Sarkar, P. H. Garthwaite, A. D. Roeck, “A Bayesian mixture model for term re-occurrence and burstiness,” CONLL '05 Proceedings of the Ninth Conference on Computational Natural Language Learning, pp. 48-55, 2005.

[6] 木本晴夫，“日本語新聞記事からのキーワード自動抽出と重要度評価,” 電子情報通信学会誌, Vol. 74-D-I, No. 8, pp.556-266, 1991.

[7] 松尾豊，石塚満，“語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム,” 人工知能学会論文誌, Vol. 17, No. 3, pp.217-223, 2002.

[8] M. J. Berryman, A. Allison, D. Abbott, “Statistical techniques for text classification based on word recurrence intervals,” Fluctuation and Noise Letters, Vol. 3, No. 1, pp. 1-10, 2003.

[9] U. Manber, G. Myers, “Suffix arrays: A new method for on-line string searches,” SIAM Journal on Computing, Vol. 22, No. 5, pp. 935-948, 1993.

[10] M. Yamamoto, K. W. Church, “Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus,” Computational Linguistics, Vol. 27, No. 1, pp. 1–30, 2001.