コルモゴロフ複雑性に基づく IDF の単語 N-gram への適用

白川　真澄† 　原　 隆浩† 　西尾章治郎†

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1-5
E-mail: †{shirakawa.masumi,hara,nishio}@ist.osaka-u.ac.jp

あらまし　本稿では，語の大域的重み付け手法である IDF（Inverse Document Frequency）と，コルモゴロフ複雑性に基づく情報距離との間にある関係性を解明する．具体的には，文字列が出現する Web ページをシャノン・ファノ符号によって表現したとき，空文字からの情報距離が対象の文字列の IDF となることを発見した．また，上記の発見に基づき，単語 N-gram に対する大域的重み付け手法を提案する．提案手法は，語の大域的重み付けと単語 N-gram の単語間結合度の測定を一つの理論的枠組みの中で処理するため，単語と複合語との間で重みを比較できる．したがって，単語 N-gram の重みのみを用いてテキストから任意の長さの特徴語を抽出できる．さらに，拡張接尾辞配列とウェーブレット木を用いた提案手法の効率的な実装方法を示す．特徴語抽出およびクエリ分割タスクにおいて，追加の技術や情報を必要とする既存手法と同等の精度を，単語 N-gram の重みのみを用いて達成した．

キーワード　IDF，コルモゴロフ複雑性，情報距離，特徴語抽出，複合語抽出

1. はじめに

TF-IDF（Term Frequency-Inverse Document Frequency）[35] に代表される語の重み付け手法（term weighting scheme）はテキスト解析において重要な基盤技術である．TF-IDF は文書検索における索引語の重みとして導入されて以来，情報検索やテキストマイニングなどの分野においてテキストを単語集合（bag-of-words）として表現するときの重みとして利用されてきた．また，形態素解析（あるいは何らかのフレーズ抽出手法）と組み合わせることで，明示的に特徴語を抽出するときの重みとしても利用できる [15].

TF-IDF をはじめとする語の重み付け手法は一般に，局所的重み付けと大域的重み付けの二つの要素からなる．局所的重みは語の出現頻度や共起頻度などの文書中の情報を利用し，対象とする文書によって語の重みが変化する．一方，大域的重みは語の文書頻度や総出現頻度などの文書集合全体の情報を利用し，語の重みは対象とする文書によらず一定である．具体例として，TF-IDF，Okapi BM25 [31]，TW-IDF [33] ではそれぞれ以下の式により語 t の文書 $d \in D$ における重みを算出する．

$$TF\text{-}IDF(t,d) = tf(t,d) \cdot \log \frac{|D|}{df(t)}$$

$$BM25(t,d) = \frac{(k_1+1) \cdot tf(t,d)}{k_1 \cdot (1-b+b \cdot \frac{|d|}{avdl}) + tf(t,d)} \cdot \log \frac{|D|}{df(t)}$$

$$TW\text{-}IDF(t,d) = \frac{tw(t,d)}{1-b+b \cdot \frac{|d|}{avdl}} \cdot \log \frac{|D|}{df(t)}$$

ここで $tf(t,d)$ は文書 d における語 t の出現頻度，$df(t)$ は文書集合 D における t の文書頻度，$|D|$ は D の文書数，$|d|$ は d の単語数，$avdl$ は D 中の文書の平均単語数，$tw(t,d)$ は tf の代わりとなるグラフに基づく頻度，k_1 および b は定数である (注1)．各式を比較してみると，局所的重み（d の影響を受ける部分）は大きく異なっているのに対し，大域的重み（d の影響を受けない部分）は共通して同じもの，すなわち IDF [17] が採用されていることがわかる．

$$IDF(t) = \log \frac{|D|}{df(t)} \tag{1}$$

IDF が代表的な語の重み付け手法に採用されている理由として，簡潔さとロバスト性が挙げられる．式 (1) のような簡潔さにも関わらず，IDF は様々な文献 [2], [22], [27], [30] において理論的にロバストであることが示されている．

IDF の欠点として，単語 N-gram に対して適用できないことが挙げられる．IDF は文書頻度の小さい語に対してより大きい重みを与えるが，単語 N-gram の場合は連続する単語のつながりが不自然であるほど文書頻度が小さくなる．その結果，単語間の結合が不自然な単語 N-gram に対して IDF は大きい重みを付与してしまう．たとえば，「Osaka University」と「Osaka be」の文書頻度の推定値 (注2) はそれぞれ 1,890,000 と 6,160 であり，「Osaka be」という不自然な単語 N-gram がより大きな重みを得る．このように，IDF は単語 N-gram の単語間のつながりの強さに反した重みを付与する．単語間結合度の測定に関する研究（multiword expression, MWE）は多数存在するが，語の重み付けとは別の問題として研究されてきた．そのため，語の重み付けと単語間結合度の測定を同時に考える場合，ヒューリスティックな方法により手法を組み合わせる必要がある．

そこで本研究では，語の重み付けと単語間結合度の測定の間にある理論的なギャップを埋めることを目指す．具体的には，語の大域的重み付け手法である IDF を，コルモゴロフ複雑性 [18], [20] に基づく情報距離 [3] として定義するための理論的説明を与える．これにより，同様に情報距離として定義された

(注1)：$k_1 = [1.2, 2]$，$b = 0.75$ とすると経験的に性能がよい [21].

(注2)：2014 年 11 月 9 日に Google 検索を用いて取得した．

単語間結合度の測定手法である MED（Multiword Expression Distance）[5] を，IDF と同時に取り扱うことが可能となる．また，IDF と MED を情報距離空間において組み合わせた手法を提案する．提案手法により，単語と複合語に対する重みが比較可能となる．すなわち，テキスト中の全ての単語 N-gram に対して付与された重みを比べることで，任意の長さの特徴語を，言語処理技術を用いることなく抽出できる．さらに，提案手法の効率的な実装を，文字列解析手法である拡張接尾辞配列とウェーブレット木を用いて実現する．特徴語抽出およびクエリ分割のアプリケーションにおいて，提案手法が有効に機能することを示す．

2. 関 連 研 究

語の重み付けはテキスト解析において最も重要な技術の一つであるため，これまで数多くの研究が行われてきた．TF-IDF [35] は代表的な語の重み付け手法であり，TF-IDF よりも一般に精度の高い Okapi BM25 [31] もよく使用される．TF-IDF は情報検索 [39]，文書クラスタリング [12]，特徴語抽出 [15]，映像中のオブジェクトマッチング [37] など，幅広いアプリケーションで利用可能である上に，複数の観点から理論的説明 [2], [16], [30], [32] が与えられており，多くの人が TF-IDF を利用する根拠となっている．また最近では，グラフを用いて TF-IDF をヒューリスティックに改善した TW-IDF [33] が提案されている．

TF-IDF や BM25，TW-IDF で共通して採用されている大域的重み付け手法が IDF [17] である．IDF 以外の選択肢として，残差 IDF [6] や gain [27] などがあるが，実際には多くの場合に IDF が用いられている．これは他の手法に対し，IDF が簡潔さとロバスト性に優れているためである．IDF は $\log \frac{|D|}{df(t)}$ という簡潔さにも関わらず，そのロバスト性を支える種々の理論的説明 [2], [22], [27], [30] が与えられている．このことはまた，IDF を理論的に改善することの難しさを意味している．残差 IDF は IDF のヒューリスティックな拡張であり，いくつかの想定環境において IDF よりも優れていることが報告されている [26], [29] が，異なるアプリケーションやデータセットにおいてもうまく機能するかどうかは保証されていない．

IDF の大きな弱点の一つとして，単語 N-gram を扱えないことが挙げられる．単語 N-gram を扱うためには，重み付けだけでなく単語間のつながりの強さを測る必要がある．単語間結合度の測定は，語の重み付けとは別の問題として研究されており，自己相互情報量（Pointwise Mutual Information, PMI）[7] やそのヒューリスティックな拡張手法 [10], [11], [36], [41] がこれまで提案されてきた．二単語のつながりの強さを測る手法としては，PMI が 84 の手法の中で最も有効な手法であることが報告されている [28]．PMI は二単語に対してのみ適用可能であるため，$N > 2$ の単語 N-gram の場合は何らかの拡張が必要となる．文献 [10], [11] では単語 N-gram を各区切りで二つに分けたときの PMI の算術平均を，文献 [36] ではその最大値を用いている．EMI（Enhanced Mutual Information）[41] は，単語 N-gram を構成する各単語の文書頻度を用いた PMI の拡張手法である．SCP（Symmetric Conditional Probability）[10] は PMI と似た手法であり，$N > 2$ の単語 N-gram に対しては算術平均をとる方法が性能がよい．

MED（Multiword Expression Distance）[5] は二単語以上の単語間結合度を測定する理論的な手法であり，コルモゴロフ複雑性に基づく情報距離として定義される．MED は，PMI のヒューリスティックな拡張手法や SCP と比べて性能がよい．MED は本研究と深く関連しているため，3 章で詳細に説明する．筆者らの知る限り，理論的な正当性を持つ単語間結合度の測定手法としては，MED が最も優れている．

このように語の重み付けと単語間結合度の測定はそれぞれ一つの研究対象として研究されてきた．しかし，これら二つの問題を一つの理論的枠組みで捉えようとする研究はこれまで行われてこなかった．本研究では，語の大域的重み付け手法である IDF と単語間結合度の測定手法である MED を，コルモゴロフ複雑性に基づく情報距離として同じ空間の中で定義できることを示す．

3. IDF と情報距離

本章では，IDF と情報距離の関係性を，コルモゴロフ複雑性，情報距離，および MED について説明しつつ解明する．

3.1 コルモゴロフ複雑性

コルモゴロフ複雑性（Kolmogorov complexity）[18], [20] はデータ列あるいは文字列の複雑さを表す指標であり，普遍的なチューリングマシン（universal Turing machine）において，対象の文字列を出力する最小のプログラムの長さとして定義される．$K(x)$ を文字列 x のコルモゴロフ複雑性として定義する．コルモゴロフ複雑性の直観的な例として，以下の文字列

x_1 = “010101010101010100”

x_2 = “011101100010110010”

について，x_1 は “01” ×8+ “00” として短く記述できるのに対し，x_2 は短く記述することが難しいため，$K(x_1)$ は $K(x_2)$ より小さいと推測できる．実際は $K(x)$ は計算不可能であるため，圧縮アルゴリズム [8] や Web ページ [9] を用いて近似される．

コルモゴロフ複雑性は条件付きとして定義することができる．$K(x|y)$ を文字列 y が入力として与えられたときの文字列 x の条件付きコルモゴロフ複雑性として定義する．このとき，$K(x)$ は空文字 ϵ を用いて $K(x|\epsilon)$ と表現できる．また，上記の文字列 x_1，x_2，および

y = “01110110001011001”

について，x_2 は y を用いると y+ “0” と短く記述できる一方，x_1 は y を用いても “01” ×8+ “00” 以上に短く記述することが困難である．したがって，$K(x_2|y)$ は $K(x_1|y)$ よりもわずかに小さいと考えられる．さらに $K(x, y)$ を，x と y を連結した文字列のコルモゴロフ複雑性とすると，対数誤差の範囲で以下の等式が成り立つ．

$$K(x, y) = K(x|y) + K(y) = K(y|x) + K(x) \quad (2)$$

式 (2) は一方の文字列が他方の文字列を記述するのに再利用できることを意味している．

3.2 情報距離

情報距離（information distance）[3] はコルモゴロフ複雑性に基づく普遍的な距離関数（metric）であり，実世界における距離と同様，アプリケーションに依存しない唯一の客観的な距離である．情報距離は，一方の文字列を他方に変換する際のエネルギーとして定義される．ランダウアーの原理 [19] によると，非可逆的に 1 ビットの情報を書き換えるのに要するエネルギーは $1k\mathcal{T}\cdot ln(2)$（k はボルツマン定数，$\mathcal{T}$ は絶対温度）となる．コルモゴロフ複雑性を用いると，文字列 x，y の情報距離 $E(x,y)$ は対数誤差の範囲で以下の式により表される．

$$E(x,y) = \max\{K(x|y), K(y|x)\} \tag{3}$$

式 (2) より，式 (3) は以下のように変換できる．

$$E(x,y) = K(x,y) - \min\{K(y), K(x)\} \tag{4}$$

情報距離は距離関数であることが証明されている．すなわち，情報距離は非負性，同一律，対称律，三角不等式を満たす．さらに，情報距離は普遍的あるいは最適であることが明らかにされている．$\mathcal{D}(x,y)$ を許容距離（admissible distance），すなわち密度条件（density condition）

$$\sum_{y:y\neq x} 2^{-\mathcal{D}(x,y)} \leqq 1, \quad \sum_{x:x\neq y} 2^{-\mathcal{D}(x,y)} \leqq 1$$

を満たす正規化された距離関数とすると，全ての x，y に対して以下の式が成り立つような非負定数 C が存在する．

$$E(x,y) \leqq \mathcal{D}(x,y) + C \tag{5}$$

式 (5) を満たす情報距離 E は，全ての許容距離の中で最適であることが保証される．直感的に説明すると，これは x と y の間にある類似した特徴を，他の許容距離よりも必ずうまく捕捉できることを意味している．

情報距離は一般に，二つのオブジェクト間の類似度を計算するのに用いられる．情報距離を定義するコルモゴロフ複雑性は計算不可能であるため，情報距離は近似的に計算される．たとえば，NCD（Normalized Compression Distance）[8] は，文字列 x，y を別々に圧縮したときのデータサイズと x と y をまとめて圧縮したときのデータサイズを用いて情報距離を推定する．NGD（Normalized Google Distance）[9] はテキストデータに特化した手法で，文字列 x，y をそれぞれ含む Web ページの頻度と x と y を両方含む Web ページの頻度を用いて情報距離を推定する．Cilibrasi と Vitányi [9] らは，NGD が Web ページ集合の大きさに対して安定していることを例証している．

3.3 Multiword Expression Distance

MED（Multiword Expression Distance）[5] は，情報距離に基づく普遍的な単語間結合度の測定手法である．具体的には，MED は単語 N-gram のコンテキスト（context）とセマンティック（semantic）の情報距離として定義される．文献 [9] に着想を得て，Bu ら [5] は，単語 N-gram のコンテキストを，それ自体を含む Web ページの集合，単語 N-gram のセマンティックを，その構成要素である単語を全て含む Web ページの集合と定義している．定義より，単語 N-gram のセマンティックは単語 N-gram のコンテキストを包含する．たとえば，「football player」のセマンティックには，「football player」が直接出現する Web ページに加え，「football」と「player」が両方出現する Web ページも含まれる．

w を単語，W を単語集合，g を単語 N-gram，$G \equiv W^+$ を単語 N-gram 集合，D を Web ページ集合，t を検索語（単語 N-gram，あるいは検索語の論理積），T を検索語集合（$G \subset T$）とする．$\phi : T \to 2^D$ を，t から t を含む Web ページ集合に写像するコンテキスト関数とし，$\phi(t)$ で表す．$\theta : G \to T$ を，単語 N-gram $g = w_1 ... w_N$ から単語の論理積 $\bigwedge_{i=1}^{N} w_i$ に分解する写像関数とし，$\theta(g)$ で表す．$\mu : G \to 2^D$ をセマンティック関数とし，合成関数 $\phi \circ \theta$ によって定義する．すなわち，$\mu(g) = \phi(\theta(g))$ である．定義より，$\phi(g) \subseteqq \mu(g)$ となる．式 (4) より，単語 N-gram g が与えられたとき，MED は g のコンテキスト $\phi(g)$ とセマンティック $\mu(g)$ を用いて以下の式により計算される．

$$\begin{aligned} MED(g) &= E(\phi(g), \mu(g)) \\ &= K(\phi(g), \mu(g)) - \min\{K(\phi(g)), K(\mu(g))\} \end{aligned} \tag{6}$$

コルモゴロフ複雑性 $K(x)$ を近似するため，シャノン・ファノ符号を用いて文字列 x の確率を算出する．まず，全ての Web ページが等確率 $\frac{1}{|D|}$ で選ばれると仮定する．$p : \phi(T) \to [0,1]$ を，コンテキストを確率に写像する関数とする．コンテキストは Web ページの集合として表されるため，コンテキスト c の確率は以下の式により計算される．

$$p(c) = \frac{|c|}{M} \tag{7}$$

なお，$M = \sum_{c\in\phi(T)} p(c)$ である．式 (7) は，c が選ばれる確率が c の Web ページ数 $|c|$ に比例することを意味している．p についてのシャノン・ファノ符号の長さは，コルモゴロフ複雑性 K の近似として利用できる [20].

$$K(x) \approx -\log p(x) \tag{8}$$

$$K(x,y) \approx -\log p(x,y) \tag{9}$$

式 (7)，(8)，(9) を用いると，式 (6) は以下のように近似できる．

$$\begin{aligned} &MED(g) \\ &\approx \max\{\log p(\phi(g)), \log p(\mu(g))\} - \log p(\phi(g), \mu(g)) \\ &= \max\{\log|\phi(g)|, \log|\mu(g)|\} - \log M \\ &\quad - \log|\phi(g) \cap \mu(g)| + \log M \\ &= \max\{\log|\phi(g)|, \log|\mu(g)|\} - \log|\phi(g) \cap \mu(g)| \end{aligned} \tag{10}$$

ここで，$\phi(g) \subseteqq \mu(g)$ であるから，式 (10) は，

$$MED(g) \approx \log|\mu(g)| - \log|\phi(g)| = \log\frac{|\mu(g)|}{|\phi(g)|} \tag{11}$$

となる．式 (11) はコンテキスト $\phi(g)$ とセマンティック $\mu(g)$ の Web ページ数から計算できる．Bu ら [5] は，Web 検索エンジンを用いて $|\phi(g)|$ と $|\mu(g)|$ を計算している．$|\mu(g)|$ は単語 N-gram g を構成する各単語の AND 検索により推定できる．

3.4 Inverse Document Frequency

前節までの理論的背景を基に，IDF と情報距離との間にある関係性を解明する．式 (1) より，単語 N-gram g の IDF は以下のように記述できる．

$$IDF(g) = \log \frac{|D|}{df(g)} = \log \frac{|D|}{|\phi(g)|} \tag{12}$$

ここで，式 (12) が式 (11) と似た形をしていることがわかる．異なる点として，対数項の中の分子が，式 (12) では $|D|$ であるのに対し，式 (11) では $|\mu(g)|$ となっている．ここで仮説として，IDF と情報距離との間に何らかの関係性があると考える．以下ではこの仮説が正しいことを示す．

まず，空文字（あるいは $N = 0$ のときの単語 N-gram）を ϵ とし，$\epsilon \in G$，$G \equiv W^*$ と定義する．ϵ を含む Web ページ集合を $\phi(\epsilon)$ とすると，ϵ は全ての Web ページに出現しているとみなせるため，$\phi(\epsilon) = D$ となる．次に，単語 N-gram g と空文字 ϵ の情報距離を，MED の場合と同様に近似計算する．具体的には，情報距離 $E(\phi(g), \phi(\epsilon))$ は以下のように定義できる．

$$E(\phi(g), \phi(\epsilon)) = K(\phi(g), \phi(\epsilon)) - \min\{K(\phi(g)), K(\phi(\epsilon))\} \tag{13}$$

MED の場合と同様に，式 (13) は以下のように展開できる．

$$\begin{aligned} & E(\phi(g), \phi(\epsilon)) \\ & \approx \max\{\log p(\phi(g)), \log p(\phi(\epsilon))\} - \log p(\phi(g), \phi(\epsilon)) \\ & = \max\{\log|\phi(g)|, \log|\phi(\epsilon)|\} - \log|\phi(g) \cap \phi(\epsilon)| \end{aligned} \tag{14}$$

ここで $\phi(g) \subseteqq \phi(\epsilon)$ であるから，式 (14) は，

$$\begin{aligned} E(\phi(g), \phi(\epsilon)) &\approx \log|\phi(\epsilon)| - \log|\phi(g)| \\ &= \log \frac{|\phi(\epsilon)|}{|\phi(g)|} = \log \frac{|D|}{|\phi(g)|} = IDF(g) \end{aligned}$$

となり，$IDF(g)$ に一致する．

IDF とコルモゴロフ複雑性に基づく情報距離との関係性を以下にまとめる．文字列を，それが出現する Web ページの集合に写像し，シャノン・ファノ符号により Web ページ集合を表現したとき，文字列のコルモゴロフ複雑性はシャノン・ファノ符号の長さによって近似できる．このとき，ある文字列と空文字との情報距離が，その文字列の IDF となる．情報距離はアプリケーションに依存しない普遍的な距離であることから，この発見により今後，情報距離の大きさを語の重みとして利用できる．

4. 提案手法: $IDF_{N\text{-}gram}$

3 章の発見に基づき，任意の長さの単語 N-gram に対して IDF を適用する手法を提案する．単語 N-gram の場合，単語間のつながりが不自然なほど IDF が大きくなることが問題であった．そこで，情報距離により二点間の距離が表される空間（情報距離空間）において，IDF を単語間結合度の測定手法である MED と組み合わせる．図 1 に情報距離空間における IDF と MED の関係を示す．文字列をそれが出現する Web ページ集合として表現し，情報距離空間に写像したとき，単語 N-gram

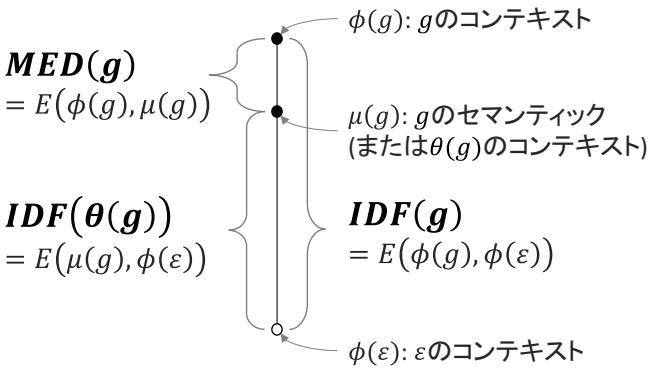

図 1　情報距離空間における IDF と MED の関係．

g の IDF は空文字 ϵ との距離，g の MED は g を構成する各単語の論理積 $\theta(g)$ との距離として表せる．

単語 N-gram の重み付けに求められる要件を整理する．$N = 1$，すなわち単語の場合，特徴的な語ほど IDF が大きくなる．また，$N > 1$ の場合，特徴的な語ほど IDF が大きくなるが，単語間のつながりが不自然な語は IDF，MED ともに大きくなる．これらを考慮すると，IDF が大きく，かつ MED が小さくなる場合に，より大きい重みを与えるような重み付け手法が妥当であると考えられる．図 1 より，そのような要件を満たす重み付け手法が二種類存在することがわかる．一つは $IDF(\theta(g))$ をそのまま用いる手法，もう一つは $IDF(\theta(g))$ と $MED(g)$ を両方用いる手法である．

$IDF(\theta(g))$ は単語 N-gram g を構成する各単語の論理積 $\theta(g)$ の IDF である．これは，g のコンテキストの代わりに g のセマンティックを用いるという解釈である．$IDF(\theta(g))$ は，情報距離空間においては g の各単語の論理積 $\theta(g)$ と空文字 ϵ との距離であり，$IDF(\theta(g)) = IDF(g) - MED(g)$ である．具体的には，$IDF(\theta(g))$ は以下の式により表される．

$$IDF(\theta(g)) = \log \frac{|D|}{|\mu(g)|} = \log \frac{|D|}{df(\theta(g))} \tag{15}$$

式 (15) は，$IDF(g)$ と同様に N の増加に対して重みが単調増加するため，異なる長さの単語 N-gram 間で大域的重みを比較することが困難である．

もう一つの手法は，$IDF(\theta(g)) - MED(g)$ である．この手法も $IDF(\theta(g))$ を基準とするが，同時に $MED(g)$ を使用する．これは，「g のコンテキストを与えることによって，g のセマンティックまでの情報距離をどの程度縮められるか」という解釈である．すなわち，g のコンテキストがより重要な意味を持つほど，g のセマンティックまでの距離を短縮できることを表している．$IDF(\theta(g)) - MED(g)$ を $IDF_{N\text{-}gram}(g)$ とし，式を展開すると以下のようになる．

$$\begin{aligned} IDF_{N\text{-}gram}(g) &= IDF(\theta(g)) - MED(g) \\ &= \log \frac{|D|}{|\mu(g)|} - \log \frac{|\mu(g)|}{|\phi(g)|} \\ &= \log \frac{|D|}{|\mu(g)|} + \log \frac{|\phi(g)|}{|\mu(g)|} \\ &= \log \frac{|D| \cdot |\phi(g)|}{|\mu(g)|^2} = \log \frac{|D| \cdot df(g)}{df(\theta(g))^2} \end{aligned} \tag{16}$$

表 1　$IDF_{N\text{-}gram}$ の例．アスタリスク付きの太字は，自身より大きい重みを持つ単語 N-gram の集合に包含されていないことを表す．

入力：new york times		入力：new york is	
***new york times**	4.241	***new york**	3.529
york times	4.205	york	3.524
times	3.531	new	1.907
new york	3.529	***is**	0.335
york	3.524	york is	-2.215
new	1.907	new york is	-2.237

$N = 1$ のとき，$df(\theta(g)) = df(g)$ であるため，$IDF_{N\text{-}gram}(g)$ と $IDF(\theta(g))$ は $IDF(g)$ に一致する．

式 (16) の $IDF_{N\text{-}gram}$ は単語 N-gram の N の変化に対して安定しており，異なる長さの単語 N-gram 間で重みを比較できる．したがって，テキストが入力として与えられたとき，$IDF_{N\text{-}gram}$ の重みのみを用いて任意の長さの特徴語を抽出できる．表 1 は「new york times」と「new york is」を入力としたときの $IDF_{N\text{-}gram}$ である (注3)．それぞれ「new york times」と「new york」という特徴語の重みが最も大きくなっていることがわかる．また，出現位置が重複している単語 N-gram 間で重みを比較することにより，不要な単語 N-gram を排除でき，結果として単語，複合語を含めた語句抽出が可能となる．具体的には，自身より大きい重みを持つ単語 N-gram の集合に包含されている場合，その単語 N-gram は不要であると判断できる．なお，一つの不要語（stopword）においてのみ包含されていない場合でも，その単語 N-gram が不要である可能性が高かったため，6 章の評価実験では不要語リストも用いている．表 1 では，左の例では「new york times」，右の例では「new york」と「is」が不要でない単語 N-gram として残っている．このように，形態素解析などの自然言語処理技術を用いずに任意の長さの特徴語を抽出できる．

5. 実　　装

$IDF_{N\text{-}gram}$ あるいは MED は，単語の論理積に対する出現頻度を計算する必要があるため計算量が大きい．Bu ら [5] は単語 N-gram を入力とし，Web 検索エンジンを用いて入力中の各単語の AND 検索を処理しているが，この方法では全ての単語 N-gram に対して事前に $IDF_{N\text{-}gram}$ を計算することは難しい．また，単語 N-gram の種類数は非常に大きいため，$IDF_{N\text{-}gram}$ を計算すべき単語 N-gram を絞ることも必要である．

本研究では，これら二つの問題に対し，それぞれ文字列解析手法を用いて対処する．まず，あらゆる単語 N-gram の中から有効な単語 N-gram を列挙するため，拡張接尾辞配列 [1] を用いた極大部分文字列の抽出 [25] (注4) を行う．極大部分文字列の概念は Blumer ら [4] によって提案され，出現位置が完全に一致する文字列群を一つのクラスとみなしたとき，各クラスにおける最長の文字列が極大部分文字列となる．極大部分文字列の数は，最大でも入力テキスト長の二倍未満となる．

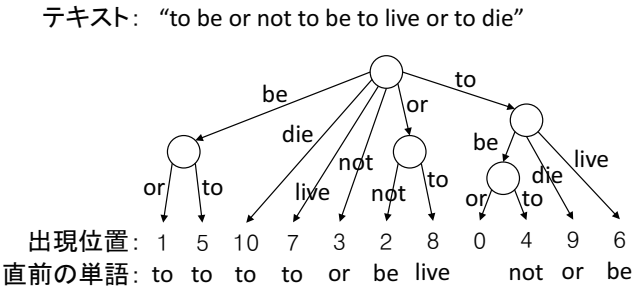

図 2　テキスト「to be or not to be to live or to die」に対する接尾辞木（拡張接尾辞配列でも表現可能）の例．複数の子ノードを持ち，かつ複数のプレフィックス（直前の単語）を持つ文字列が極大部分文字列となる．ここでは「or」「to」「to be」が極大部分文字列となる．

極大部分文字列の候補は，テキストを接尾辞木で表現したとき，複数の子ノードを持つ中間ノードとして表される．図 2 は「to be or not to be to live or to die」というテキストを接尾辞木で表現している．出現位置をユニークに指す単語 N-gram が木構造で表現され，辞書順に出現位置の番号がソートされている．全ての単語 N-gram の中で，「be」「or」「to」「to be」が複数の子ノードを持っており，これらは極大部分文字列の候補である．この中で複数のプレフィックス（直前の単語）を持つ単語 N-gram が極大部分文字列となる．図 2 では，「be」の直前の単語が「to」のみであるから，これを排除する．最終的に「or」「to」「to be」が極大部分文字列であることがわかる．本研究では，極大部分文字列を有効な単語 N-gram とし，拡張接尾辞配列（接尾辞木を表現可能）のライブラリである esaxx(注5) を利用して有効な単語 N-gram を列挙するコードを実装した．

次に，単語の論理積に対して文書頻度を計算するため，ウェーブレット木 [14] を用いる．ウェーブレット木と呼ばれるデータ構造は様々な処理を高速に行えることが最近わかってきたが，Gagie ら [13] は，ウェーブレット木が文書頻度の計算にも利用できることを示した．文書頻度の計算はこれまでも接尾辞配列を用いた高速な手法 [40] が存在していたが，ウェーブレット木を用いることで，論理積を含む場合にも文書頻度を比較的高速に計算できるようになった．時間計算量は，単語 N-gram g の場合は $O(df(g) \cdot \log |D|)$ であるのに対し，g の各単語の論理和 $\theta(g)$ の場合は $O(N \cdot df(\theta(g)) \cdot \log |D|)$ であり，単語 N-gram の長さ N が乗算されるのみである．またこれは，文書頻度の厳密計算において理論的な限界に近づいていることがわかっている [13]．本研究では Gagie らの手法を，ウェーブレット木のライブラリである wat-array(注6) を用いて実装した．

以下では，文書集合 D から全ての有効な単語 N-gram に対して $IDF_{N\text{-}gram}$ を計算する処理について説明する．まず，D 中の文書を一つにつなげたテキストを生成し，このテキストから拡張接尾辞配列を用いて極大部分文字列を列挙し，有効な単語 N-gram とする．このとき，ソートされたテキストの出現位置に対応する文書 ID のリストを同時に作成する．これにより，

(注3)：5 章の実装を用いて算出した．
(注4)：文献 [24] でも同様の手法が提案されている．
(注5)：https://code.google.com/p/esaxx/
(注6)：https://code.google.com/p/wat-array/

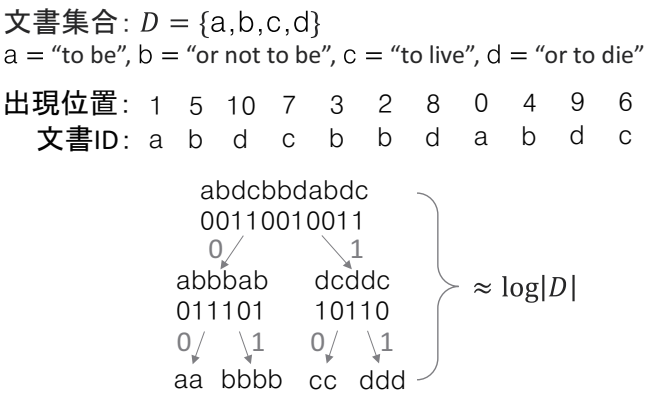

図 3　ウェーブレット木の例.

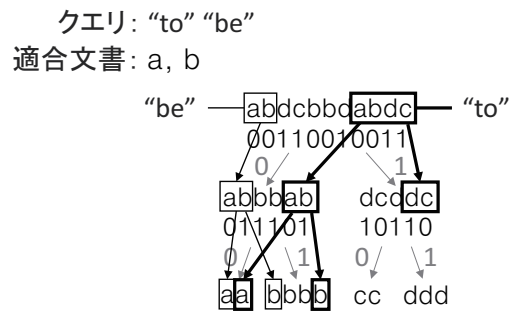

図 4　ウェーブレット木を用いた文書頻度の計算の例.

ある単語 N-gram を与えたとき，それが出現する文書集合をリスト上の一つの範囲として表現できる．次に，文書 ID のリストに対してウェーブレット木を構築する．図 3 にウェーブレット木の例を示す．文書集合 D 中の文書を一つにつなげたテキストが「to be or not to be to live or to die」となるため，出現位置のリストは図 2 と全く同じになる．また，出現位置に対応する文書 ID のリストが abdcbbdabdc となる．構築されたウェーブレット木は高さ約 $\log|D|$ の全二分木となり，全てのノードが単一の文書 ID になるまで文書 ID を 0 または 1 の子ノードに割り振り，子ノード内の文書 ID の順序は保持される．ウェーブレット木を構築した後，全ての有効な単語 N-gram とその各単語の論理積について，それぞれをクエリとしたときの文書頻度を計算し，$IDF_{N\text{-}gram}$ を求める．クエリの範囲についてウェーブレット木をたどっていき，到達できた葉ノード（文書 ID）を取得することで文書頻度が得られる．図 4 の例では，ウェーブレット木を用いて「to」「be」の AND 検索に対する適合文書（文書頻度）を取得している.「to」「be」のそれぞれの範囲についてウェーブレット木をたどっていくと，a と b の二つの文書に適合することがわかる．

上記の処理を行うプログラムを，esaxx と wat-array を用いて実装し，2013 年 10 月 1 日の英語版 Wikipedia のテキストデータ（約 10.5GB）から $IDF_{N\text{-}gram}$ を計算した．なお，アルファベットは全て小文字に変換した．また，処理時間短縮のため，単語 N-gram g の単語の論理積に対して文書頻度を計算する際，単語の最小の文書頻度に対する g の文書頻度の比率が 0.0005 に満たない場合，$IDF_{N\text{-}gram}$ が大きくなる可能性が著しく低いものとして計算を省略した．メモリ 60GB の計算機 2 台を用い，12 日かけて有効な単語 N-gram に対する $IDF_{N\text{-}gram}$ を計算した．実装したプログラム，処理済みのデータ，デモシステムは Web ページで公開している (注7).

6. 評 価 実 験

$IDF_{N\text{-}gram}$ のロバスト性を検証するため，特徴語抽出およびクエリ分割の評価実験を行った．なお，$IDF_{N\text{-}gram}$ を bag-of-words の特徴量として利用する実験についても，情報検索，文書分類，文書クラスタリングにおいて行ったが，重み計算の手法によらず，$N > 1$ の単語 N-gram の情報が性能向上に寄与しなかったため，ここでは割愛する．

6.1 特徴語抽出

Wikipedia データセットを用いて特徴語抽出の実験を行った．Wikipedia では，各ページにおける特徴語がアンカーテキストとして別の Wikipedia のページにリンクされているため，これを正解データとして利用した．具体的には，Wikipedia からランダムに記事を選び，各記事の第一段落を抽出した後，第一段落に含まれるアンカーテキスト及び太字の語（主に記事のタイトルが太字で表記される）を正解の特徴語とした．ここで，第一段落のみを利用しているのは，局所的重みの影響を抑えるためである．短文においては，語の出現頻度が多くの場合 1 となり，局所的重みが機能しないという問題が存在する [38]．つまり，短文における評価を行うことで，大域的重みの性能をより直接的に評価できる．第一段落から作成したデータセットは，平均して 60.2 の単語（最大 291，最小 8），6.7 の特徴語（最大 30，最小 3) (注8) を含む 1,678 のテキストである．

評価を行う手法は，提案手法の $IDF_{N\text{-}gram}$ を TF と組み合わせた手法（$TF\text{-}IDF_{N\text{-}gram}$），形態素解析と TF-IDF を組み合わせた手法（TF-IDF * with NPs），および単語のみを対象とする TF-IDF（TF-IDF uni-gram）とした．形態素解析と TF-IDF を組み合わせた手法は，他の様々な手法と比較してもロバストに機能することが知られている [15]．品詞情報を用いて名詞句を特徴語の候補として抽出した後，各候補について，その語を構成する単語の TF-IDF の和を最終的な重みとする（TF-IDF sum with NPs）．また，文献 [15] では検証されていないが，TF-IDF の平均を用いる場合（TF-IDF average with NPs），語自体の TF-IDF を用いる場合（TF-IDF direct with NPs）についても評価を行った．加えて，形態素解析による名詞句の抽出精度が低下するような状況を想定し，データセットのテキストを全て小文字に変換した場合（TF-IDF * with NPs, No-Cap）についても評価した．各手法を用いて特徴語のランク付きリストを生成した後，正解数 R に対して上位 R の出力の適合率 R-Prec を評価尺度とした．

表 2 に評価結果を示す．TF-IDF sum with NPs が最も高い 0.386 を達成し，提案手法の R-Prec が次いで 0.377 となっている．形態素解析を用いた手法の中では，文献 [15] で用いられているように各単語の TF-IDF の和を用いる手法が最もスコア

(注7)：http://iwnsew.com
(注8)：特徴語が 3 未満のテキストは破棄した.

表 2　Wikipedia 第一段落データセットにおける特徴語抽出の評価.

手法	R-Prec
$TF\text{-}IDF_{N\text{-}gram}$	0.377
TF-IDF sum with NPs	0.386
TF-IDF sum with NPs, No-Cap	0.369
TF-IDF average with NPs	0.367
TF-IDF average with NPs, No-Cap	0.352
TF-IDF direct with NPs	0.369
TF-IDF direct with NPs, No-Cap	0.355
TF-IDF uni-gram	0.229

表 3　Wikipedia 全文データセットにおける特徴語抽出の評価.

手法	R-Prec	Prec@10
$TF\text{-}IDF_{N\text{-}gram}$	0.300	0.358
TF-IDF sum with NPs	0.317	0.427
TF-IDF sum with NPs, No-Cap	0.301	0.409
TF-IDF average with NPs	0.283	0.359
TF-IDF average with NPs, No-Cap	0.269	0.333
TF-IDF direct with NPs	0.282	0.355
TF-IDF direct with NPs, No-Cap	0.270	0.341
TF-IDF uni-gram	0.154	0.200

が高い．TF-IDF uni-gram は，複合語を一切抽出できないため，明らかに他の手法よりも劣っている．結果より，提案手法は単語 N-gram の重みのみを用いているにも関わらず，形態素解析によって名詞句を抽出する手法とほぼ同等の性能を達成できていることがわかる．テキストを全て小文字に変換した場合，主に固有名詞の抽出精度が低下したため，提案手法が TF-IDF sum with NPs をわずかに上回る結果となっている．このことから，特にマイクロブログのようなノイズの多い短文に対して，提案手法が優位性を持つ可能性が高い．

また，Wikipedia のページの全文をデータセットとしたときの評価も行った．全文データセットは 1,747 のテキストを含み，単語数の平均は 805.5（最大 16,904，最小 102），特徴語数の平均は 36.8（最大 999，最小 10）(注9) である．表 3 に結果を示す．全文データセットでは一つのテキストに多くの特徴語が含まれているため，R-Prec だけでなく，出力の上位 10 語についての適合率 Prec@10 も測定した．提案手法は，特に Prec@10 でみたとき，TF-IDF sum with NPs に劣っていることがわかる．これは局所的重みの影響によるものと考えられる．$IDF_{N\text{-}gram}$ は全ての長さの単語 N-gram に対して情報距離に基づく公平な大域的重みを付与する．そのため，テキストが長くなると，$TF\text{-}IDF_{N\text{-}gram}$ は出現回数が多くなりがちな単語に対して大きい重みを付与する傾向があった．一方，TF-IDF sum with NPs は名詞句を構成する各単語の TF-IDF の和をとるため，複数単語からなる名詞句が出力の上位に多く出現した．複数単語からなる名詞句は特徴語であることが多かったため，Prec@10 でみると，TF-IDF sum with NPs が高いスコアを達成できたものと考えられる．この結果から，$IDF_{N\text{-}gram}$ を長いテキストに適用する場合は，局所的重みについても単純な TF 以外の手法を検討する必要がある．

(注9)：特徴語が 10 未満のテキストは破棄した．

6.2　クエリ分割

クエリ分割の実験では，Roy らの情報検索ベースのクエリ分割データセット [34] を利用した．このデータセットは 13,959 の Web ページ，500 のクエリ，クエリに対する Web ページの関連性のスコア（Qrel）を含む．Qrel は 3 人の被験者が 3 段階（2 : 強い関連がある，1 : 関連がある，0 : 関連がない）でクエリ中の全単語を含む Web ページを評価したものであり，その平均を正解データとして用いる．情報検索ベースのクエリ分割では，与えられたクエリ（例 :「larry the lawnmower tv show」）に対し，クエリを語句に分割（例 :「larry the lawnmower」「tv show」）し，それらの語句が出現する Web ページのみを出力に含める（すなわち Web 検索における引用符を利用する）ことによって情報検索の性能がどの程度向上するかを測定する．なお，クエリに対する文書の関連度は通常の TF-IDF を用いて測る．分割された語句のうち，どの語句に引用符を付けるかを決めるのは難しい問題であるため，Roy ら [34] はその中で最適な組合せをとった場合のスコアを測定している．本研究では，Roy らの評価方法に従う．

また，Roy らのデータセットには，複数の比較手法についてのクエリ分割結果が含まれている．Mishra [23] はクエリログを用いた確率的な手法であり，Mishra+Wiki はその手法に Wikipedia のタイトルを組み合わせた手法である．PMI を用いた比較手法としては，クエリログから計算した PMI-Q と Web コーパスから計算した PMI-W がある．PMI-Q，PMI-W の閾値は Roy らのトレーニングセットを用いて調整されている．さらに，3 人の人によるクエリ分割結果（Human）と，引用符の組合せで最大のスコアとなった場合（BQV）もデータセットに含まれている．これらに加え，引用符を用いない場合（Unsegmented）についても評価を行った．本実験では，データセットのドメインに対応するため，提案手法の $IDF_{N\text{-}gram}$ をデータセットの Web ページ集合から計算した．なお，データセットの Web ページ（283MB）を処理するのに 40 分を要した．評価尺度も Roy らと同様に，nDCG，MAP，MRR を上位 5 および 10 の Web ページに対して測定した．MAP と MRR は Qrel を二値（関連がある，関連がない）に変換する必要があるため，MAP では 2 と 1 を，MRR では 2 のみを関連があるとみなした．

表 4 はクエリ分割の評価結果である．なお，比較手法の表 4 における結果が Roy らの結果 [34] と異なっているが，これはデータセットの Web ページ集合を用いて TF-IDF を計算したためである．本研究で得られた結果は，Roy らの結果と手法間のスコアの傾向が一致している．表 4 より，提案手法は，Mishra+Wiki や PMI-Q などのクエリ分割に特化した手法，あるいは人による分割結果を用いた場合と同等程度の性能を達成できていることがわかる．提案手法が Web ページ集合から計算した単語 N-gram の大域的重みのみを用いていることを考慮すると，クエリログや Wikipedia のタイトルなどの追加の情報を必要とする手法と同等の性能であることは，提案手法の簡潔さやロバスト性を表している．

表 4　Roy ら [34] のクエリ分割データセットにおける評価.

手法	nDCG @5	nDCG @10	MAP @5	MAP @10	MRR @5	MRR @10
$IDF_{N\text{-}gram}$	0.730	0.742	0.900	0.893	0.582	0.593
Mishra	0.706	0.737	0.895	0.892	0.529	0.542
Mishra+Wiki	0.725	0.750	0.907	0.902	0.561	0.571
PMI-Q	0.716	0.736	0.898	0.892	0.567	0.577
PMI-W	0.670	0.707	0.860	0.863	0.493	0.506
Unsegmented	0.655	0.689	0.852	0.854	0.465	0.481
Human A	0.728	0.746	0.904	0.899	0.575	0.585
Human B	0.727	0.747	0.903	0.898	0.567	0.577
Human C	0.717	0.744	0.899	0.896	0.543	0.555
BQV	0.765	0.768	0.927	0.914	0.673	0.680

7. おわりに

本研究では，コルモゴロフ複雑性に基づく情報距離と IDF との関係性を明らかにした．IDF は，文字列が出現する Web ページをシャノン・ファノ符号によって表現したとき，空文字からの情報距離に一致する．また，このことを利用して，単語 N-gram に対する大域的重み付け手法を提案した．提案手法は，情報距離に基づく単語 N-gram の単語間結合度の測定手法である MED を，情報距離空間において IDF と組み合わせ，任意の長さの単語 N-gram に対する普遍的な重みを付与する．単語と複合語の重みが比較可能となるため，単語 N-gram の重みのみを用いてテキストから任意の長さの特徴語を抽出できる．提案手法は，特徴語抽出とクエリ分割の評価実験において，追加の技術や情報を必要とする既存手法と同等の精度を達成できた．

今後の課題として，計算量の削減および日本語などの区切り文字のない言語への対応が挙げられる．計算量に関しては，単語の論理積に対する文書頻度のカウント，すなわち，複数の単語が同時に出現する文書を列挙する処理において，理論的にこれ以上の大幅な高速化が期待できないため，何らかの近似手法が必要とされる．また，区切り文字のない言語への対応では，単語分割手法の導入，あるいは文字 N-gram に対応可能な手法を検討する必要がある．

謝　辞

本研究の一部は文部科学省国家課題対応型研究開発推進事業－次世代 IT 基盤構築のための研究開発－「社会システム・サービスの最適化のための IT 統合システムの構築」（2012 年度～2016 年度）の助成による．

文　献

[1] M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch. Replacing Suffix Trees with Enhanced Suffix Arrays. *Journal of Discrete Algorithms*, 2(1):53–86, Mar. 2004.
[2] A. Aizawa. An Information-Theoretic Perspective of TF-IDF Measures. *Information Processing and Management*, 39(1):45–65, Jan. 2003.
[3] C. H. Bennett, P. Gács, M. Li, P. M. Vitányi, and W. H. Zurek. Information Distance. *IEEE ToIT*, 44(4):1407–1423, July 1998.
[4] A. Blumer, J. Blumer, D. Haussler, R. M. McConnell, and A. Ehrenfeucht. Complete Inverted Files for Efficient Text Retrieval and Analysis. *Journal of the ACM*, 34(3):578–595, July 1987.
[5] F. Bu, X. Zhu, and M. Li. Measuring the Non-compositionality of Multiword Expressions. In *COLING*, pp. 116–124, Aug. 2010.
[6] K. W. Church and W. A. Gale. Poisson Mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.
[7] K. W. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29, Mar. 1990.
[8] R. L. Cilibrasi and P. M. Vitányi. Clustering by Compression. *IEEE ToIT*, 51(4):1523–1545, Apr. 2005.
[9] R. L. Cilibrasi and P. M. Vitányi. The Google Similarity Distance. *IEEE TKDE*, 19(3):370–383, Mar. 2007.
[10] J. F. da Silva and J. G. P. Lopes. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multi-word Units from Corpora. In *Meeting on Mathematics of Language (MOL)*, pp. 369–381, July 1999.
[11] G. Dias. Mining Textual Associations in Text Corpora. In *ACM SIGKDD Workshop on Text Mining*, pp. 20–23, Aug. 2000.
[12] B. C. Fung, K. Wang, and M. Ester. Hierarchical Document Clustering Using Frequent Itemsets. In *Proceedings of SIAM International Conference on Data Mining (SDM)*, pp. 59–70, May 2003.
[13] T. Gagie, G. Navarro, and S. J. Puglisi. New Algorithms on Wavelet Trees and Applications to Information Retrieval. *Theoretical Computer Science*, 426–427:25–41, Apr. 2012.
[14] R. Grossi, A. Gupta, and J. S. Vitter. High-Order Entropy-Compressed Text Indexes. In *SODA*, pp. 841–850, 2003.
[15] K. S. Hasan and V. Ng. Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art. In *COLING*, pp. 365–373, Aug. 2010.
[16] D. Hiemstra. A Probabilistic Justification for Using TF×IDF Term Weighting in Information Retrieval. *International Journal on Digital Libraries*, 3(2):131–139, Aug. 2000.
[17] K. S. Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28:11–21, 1972.
[18] A. Kolmogorov. On Tables of Random Numbers. *Sankhyā Ser. A*, 25:369–376, 1963.
[19] R. Landauer. Irreversibility and Heat Generation in the Computing Process. *IBM Journal of Research and Development*, 5(3):183–191, July 1961.
[20] M. Li and P. M. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, Berlin, 3rd edition, 2009.
[21] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.
[22] D. Metzler. Generalized Inverse Document Frequency. In *CIKM*, pp. 399–408, Oct. 2008.
[23] N. Mishra, R. S. Roy, N. Ganguly, S. Laxman, and M. Choudhury. Unsupervised Query Segmentation Using only Query Logs. In *WWW*, pp. 91–92, Mar./Apr. 2011.
[24] K. Narisawa, S. Inenaga, H. Bannai, and M. Takeda. Efficient Computation of Substring Equivalence Classes with Suffix Arrays. In *CPM*, pp. 340–351, July 2007.
[25] D. Okanohara and J. Tsujii. Text Categorization with All Substring Features. In *SDM*, pp. 838–846, Apr./May 2009.
[26] C. Orăsan, V. Pekar, and L. Hasler. A Comparison of Summarisation Methods Based on Term Specificity Estimation. In *LREC*, pp. 1037–1041, May 2004.
[27] K. Papineni. Why Inverse Document Frequency? In *NAACL*, pp. 1–8, June 2001.
[28] P. Pecina. An Extensive Empirical Study of Collocation Extraction Methods. In *ACL Student Research Workshop*, pp. 13–18, June 2005.
[29] J. D. M. Rennie and T. Jaakkola. Using Term Informativeness for Named Entity Detection. In *SIGIR*, pp. 353–360, Aug. 2005.
[30] S. Robertson. Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, 2004.
[31] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *TREC*, pp. 109–126, 1994.
[32] T. Roelleke and J. Wang. TF-IDF Uncovered: A Study of Theories and Probabilities. In *SIGIR*, pp. 435–442, July 2008.
[33] F. Rousseau and M. Vazirgiannis. Graph-of-word and TW-IDF: New Approach to Ad Hoc IR. In *CIKM*, pp. 59–68, Oct./Nov. 2013.
[34] R. S. Roy, N. Ganguly, M. Choudhury, and S. Laxman. An IR-based Evaluation Framework for Web Search Query Segmentation. In *SIGIR*, pp. 881–890, Aug. 2012.
[35] G. Salton, A. Wong, and C.-S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.
[36] P. Schone and D. Jurafsky. Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? In *EMNLP*, pp. 100–108, June 2001.
[37] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, pp. 1470–1477, Oct. 2003.
[38] M. Timonen. *Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion*. PhD thesis, University of Helsinki, 2013.
[39] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok. Interpreting TF-IDF Term Weights as Making Relevance Decisions. *ACM TOIS*, 26(3):13:1–13:37, June 2008.
[40] M. Yamamoto and K. W. Church. Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus. *Computational Linguistics*, 27(1):1–30, Mar. 2001.
[41] W. Zhang, T. Yoshida, X. Tang, and T.-B. Ho. Improving Effectiveness of Mutual Information for Substantival Multiword Expression Extraction. *Expert Systems with Applications*, 36(8):10919–10930, Oct. 2009.