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

DEIM2015 A3:情報検索(Ph.Dセッション), 2015年3月2日

本研究の貢献

① IDFに新たな理論的説明を付与

② ①に基づきIDFを単語N-gramに適用した手法を提案

③ ②を計算する手法を実装

接尾辞木とウェーブレット木を用いた
計算方法を実装・公開した
http://github.com/iwnsew/ngweight
（計算済みのデータ，デモシステムを含む）

今日は話しませんので詳しくは論文で

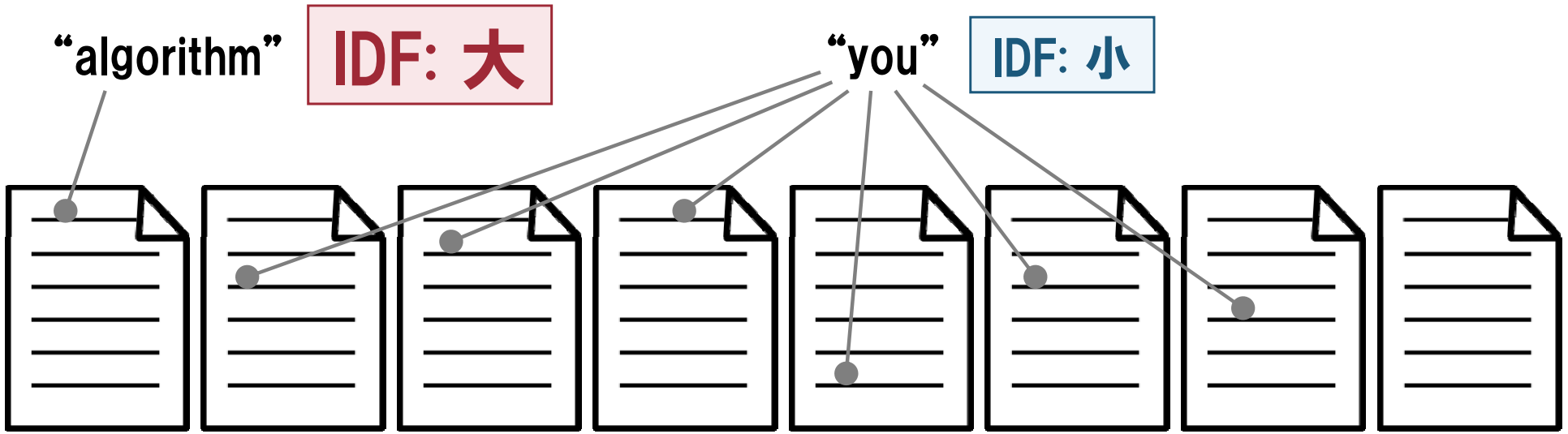
④ ②の頑健性を評価実験にて確認

IDF（Inverse Document Frequency，逆文書頻度）

より少数の文書にしか出現しない語に大きい重みを付与

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

t：語
$df(t)$：語 t の文書頻度
$|D|$：文書集合 D の文書数



IDFの欠点

単語N-gram(複合語)ではうまく機能しない

何故か？

文書頻度が小さい単語N-gram　→ 特徴的！
つながりが不自然な単語N-gram → 文書頻度が小さい！
つながりが不自然な単語N-gram → 特徴的！！！？！？？

Google検索による文書頻度の推定値

$$df(\text{"Osaka University"}) = 2{,}490{,}000$$
$$df(\text{"Osaka be"}) = 7{,}790$$
$$df(\text{"Osaka be to"}) = 4$$

単語間結合度の測定

「語の重み付け」とは別に研究されてきた
Multiword Expression（MWE）とか言われる

代表的な手法
- ✓ PMI（Pointwise Mutual Information）
- ✓ SCP［Silva+, MOL99］
- ✓ EMI［Zhang+, Expert Systems with Applications, 2009］
- ✓ MED［Bu+, COLING10］
- ✓ などなど…

本研究は，IDFに単語間結合度の測定を
「理論的」に組み込むことを目指す

コルモゴロフ複雑性と情報距離

コルモゴロフ複雑性 (Kolmogorov Complexity) (1/2)

[Kolmogorov, Sankhya, 63]

有限長のデータ列(文字列など)の複雑さを表す指標

$K(x)$: **データ列 x の複雑性**

Q1: x_1 と x_2 はどちらが複雑でしょうか？

x_1 =“0101010101010101” → “01”×7+”0”

x_2 =“0111011000101l0” → “0111011000101l0”

A1: たぶん $K(x_1) < K(x_2)$

コルモゴロフ複雑性 (Kolmogorov Complexity) (2/2)

[Kolmogorov, Sankhya, 63]

有限長のデータ列(文字列など)の複雑さを表す指標

$K(x|y)$: **データ列 y が与えられた条件での x の複雑性**

Q2: x_1 と x_2 はどちらが複雑でしょうか？(ただし y を用いてもよい)

$y =$“0111011000101 1”

$x_1 =$“0101010101010 10” → “01”×7+”0”

$x_2 =$“0111011000101 10” → y+“0”

A2: たぶん $K(x_1|y) > K(x_2|y)$

$K(x,y) = K(x|y) + K(y) = K(y|x) + K(x)$ **が成り立つ**

(一方の文字列を，他方を表すのに再利用できる)

情報距離（Information Distance）

[Bennett+, TOIT, 98]

コルモゴロフ複雑性に基づき定義される距離

$E(x,y)$: **データ列** x **と** y **の情報距離**

$$E(x,y) = \max(K(x|y), K(y|x)) = K(x,y) - \min(K(y), K(x))$$

情報距離とは，一方を他方に変換するためのエネルギー

1ビット変換するのにかかるエネルギー $= 1kT\ln(2)$

[Landauer, IBM Journal of Research and Development, 61]

絶対温度

ボルツマン定数

MED: Multiword Expression Distance

[Bu+, COLING10]

情報距離に基づく普遍的な単語間結合度の測定手法

$$MED(g) = \log \frac{df(w_1, \cdots, w_N)}{df(g)}$$

g: $w_1 \cdots w_N$ で表される N-gramの語

語のコンテキストとセマンティックの間の情報距離

g のコンテキスト: g を含む文書集合
g のセマンティック: ユニグラム $w_1, \cdots, w_N$ を全て含む文書集合

Inspired by [Cilibrasi+, TKDE, 2007]（Normalized Google Distanceの論文）

MEDの導出（1/2）

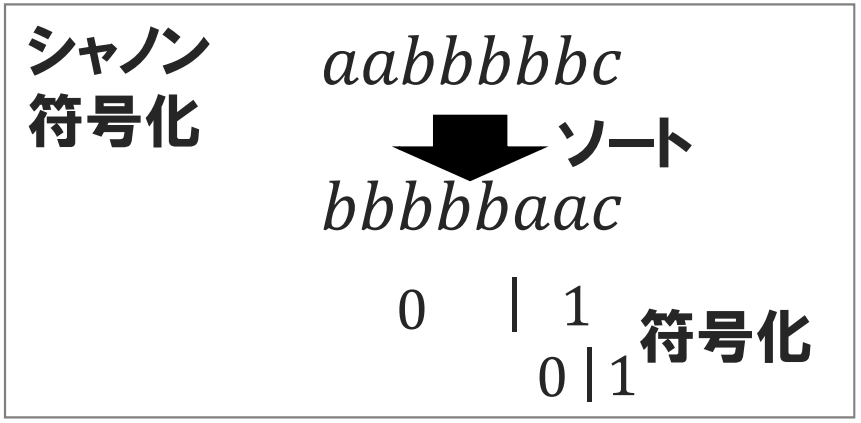
コンテキスト（文書集合）x の確率を，
文書頻度 $|x|$ に比例すると考える

$$\mathrm{P}(x) = \frac{|x|}{\sum_{x_i \in X} |x_i|}$$

X: **全てのコンテキストの集合**

すると，シャノン符号化[*1]を用いて，以下のように
コルモゴロフ複雑性を近似できる [Li+2008]

$$K(x) \approx -\log P(x)$$
$$K(x,y) \approx -\log P(x,y)$$



*1 http://en.wikipedia.org/wiki/Shannon%E2%80%93Fano_coding

MEDの導出（2/2）

$\phi(g)$ と $\mu(g)$ の情報距離は，

g のコンテキスト　　g のセマンティック

$$
\begin{aligned}
&E\big(\phi(g),\mu(g)\big)\\
&= K\big(\phi(g),\mu(g)\big) - \min\Big(K\big(\phi(g)\big),K\big(\mu(g)\big)\Big)\\
&= -\log P\big(\phi(g),\mu(g)\big) + \min\Big(\log P\big(\phi(g)\big),\log P\big(\mu(g)\big)\Big)\\
&= -\log|\phi(g)\cap\mu(g)| + \max(\log|\phi(g)|,\log|\mu(g)|)\\
&= -\log|\phi(g)| + \log|\mu(g)|
\end{aligned}
$$

シャノン符号化

確率は
文書頻度に比例

$\phi(g)\subseteq\mu(g)$ なので

g を含む文書の頻度

$w_1,\cdots,w_N$ を全て含む文書の頻度

すなわち，$MED(g) = \log\dfrac{df(w_1,\cdots,w_N)}{df(g)}$

IDFと情報距離の関係

IDF

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

g のコンテキストと空文字 ε のコンテキストとの情報距離

ε はすべての文書に含まれる

$$E\big(\phi(g), \phi(\varepsilon)\big)$$
$$= -\log|\phi(g) \cap \phi(\varepsilon)| + \max(\log|\phi(g)|, \log|\phi(\varepsilon)|)$$

シャノン符号化＆確率は文書頻度に比例

$$= -\log|\phi(g)| + \log|\phi(\varepsilon)|$$

← $\phi(g) \subseteq \phi(\varepsilon)$ なので

$$= \log \frac{df(\varepsilon)}{df(g)} = \log \frac{|D|}{df(g)} = IDF(g)$$

IDFと一致

本研究の貢献

① IDFに新たな理論的説明を付与

語を，それが出現する文書の集合で表現し，シャノン符号化により文書集合のコルモゴロフ複雑性を近似したとき，語と空文字との情報距離がIDFとなる

② ①に基づきIDFを単語N-gramに適用した手法を提案

③ ②を計算する手法を実装

接尾辞木とウェーブレット木を用いた計算方法を実装・公開した
http://github.com/iwnsew/ngweight
（計算済みのデータ，デモシステムを含む）

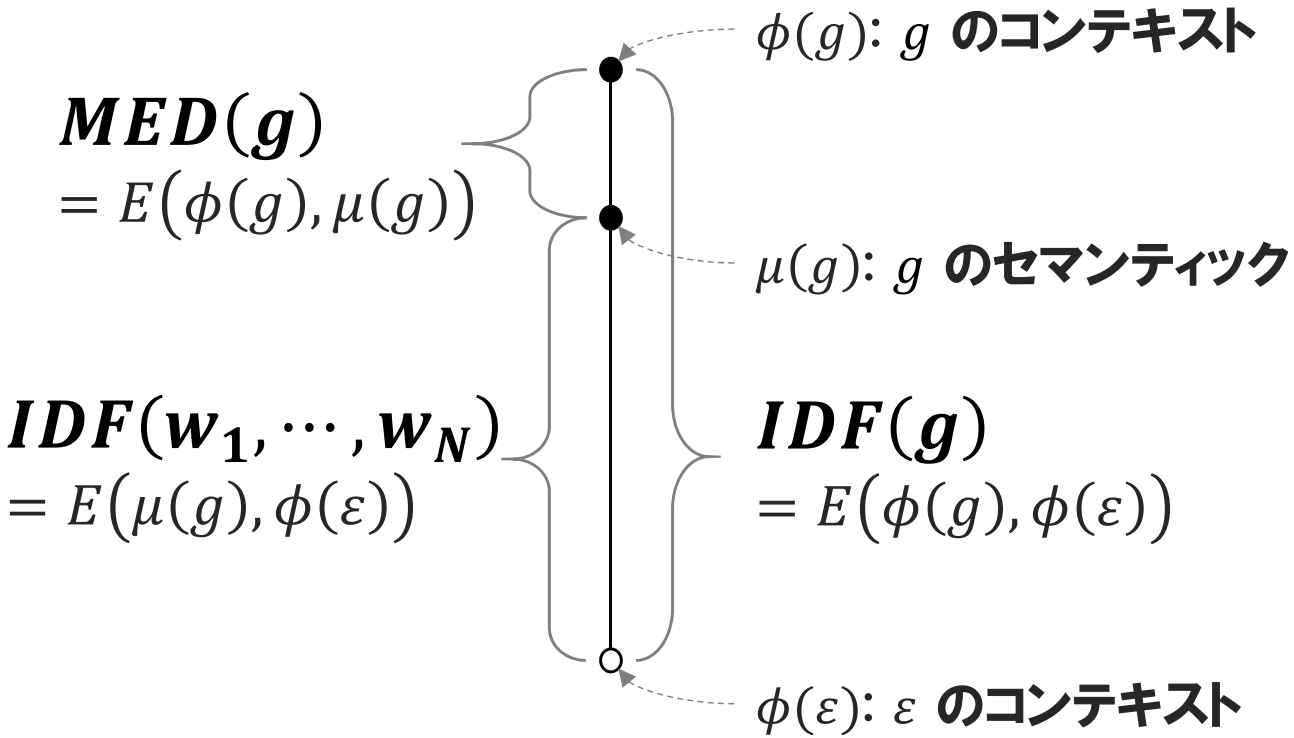
今日は話しませんので詳しくは論文で

④ ②の頑健性を評価実験にて確認

情報距離空間におけるIDFとMED

単語N-gram g のIDFは，g を単語に分解したときのIDFと g のMEDの和である

でもMEDはむしろ小さいほうが，単語間の結合が自然なのでは？

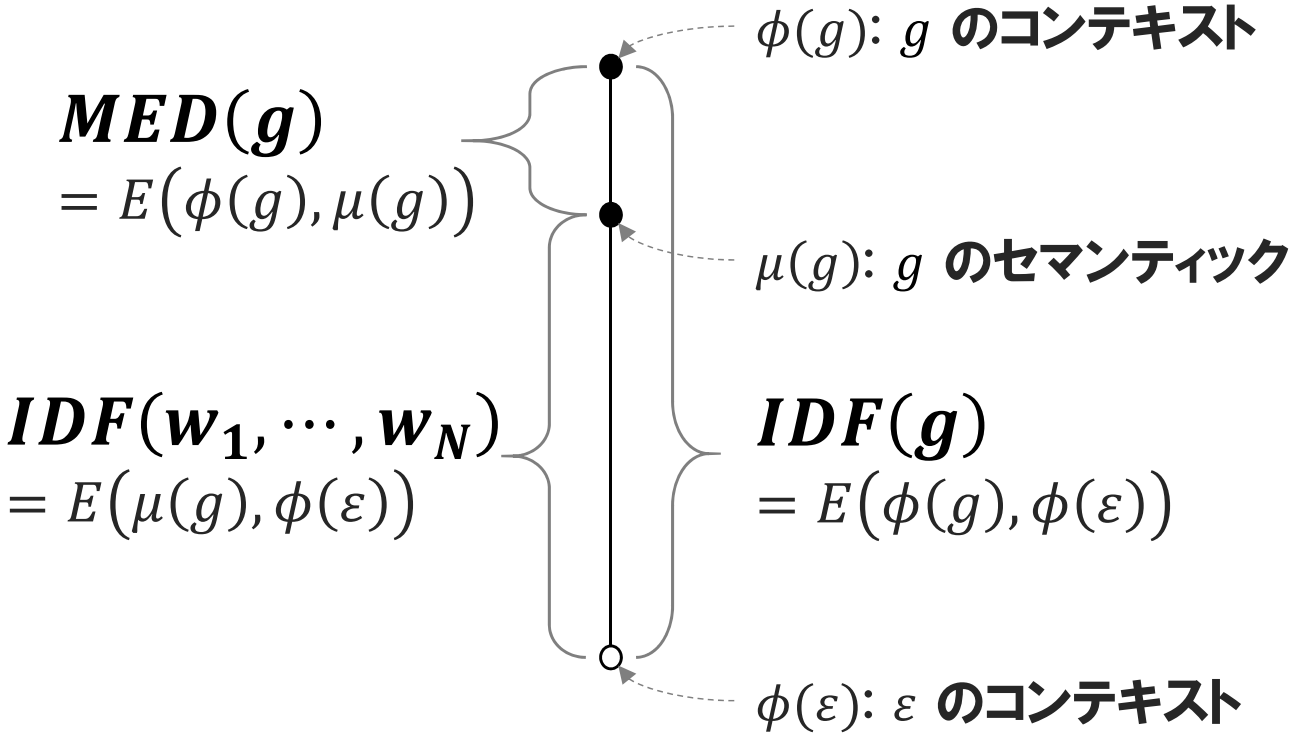


提案手法: IDF_{N-gram}

情報距離に基づいてIDFを再設計

IDFは大きく，且つMEDが小さいときに大きい重みを付与

$$IDF_{N-gram}(g) = IDF(w_1, \cdots, w_N) - MED(g) = log \frac{|D| \cdot df(g)}{df(w_1, \cdots, w_N)^2}$$



重みのみを用いた特徴語抽出

入力：”Alice’s Adventures in Wonderland - Kindle edition by Lewis Carroll”

単語N-gram	IDF_{N-gram}
kindle edition	12.043
kindle	11.653
alice s adventures in wonderland	11.496
adventures in wonderland	10.906
s adventures in wonderland	10.804
wonderland	9.670
lewis carroll	9.498
alice s adventures	9.385
alice s adventures in	9.348
in wonderland	8.762
carroll	8.152
by lewis carroll	7.461
alice	7.234

単語N-gram	IDF_{N-gram}
adventures	7.101
kindle edition by	6.739
lewis	6.192
edition	4.836
adventures in	4.280
s adventures	3.586
alice s	3.507
s adventures in	2.255
by lewis	1.768
s	1.030
by	0.820
in	0.154
edition by	-0.875

本研究の貢献

① IDFに新たな理論的説明を付与

語を，それが出現する文書の集合で表現し，シャノン符号化により文書集合のコルモゴロフ複雑性を近似したとき，語と空文字との情報距離がIDFとなる

② ①に基づきIDFを単語N-gramに適用した手法を提案

単語N-gramに対応した重み付け手法を提案し，さらに，重みのみを用いて任意の長さの特徴語をテキスト中から抽出できることを発見した

③ ②を計算する手法を実装

接尾辞木とウェーブレット木を用いた計算方法を実装・公開した
http://github.com/iwnsew/ngweight
（計算済みのデータ，デモシステムを含む）

今日は話しませんので詳しくは論文で

④ ②の頑健性を評価実験にて確認

評価1: 特徴語抽出

Wikipediaの第一段落1,678記事分をデータセットとして利用

アンカーテキストと太字の語を正解として適合率(上位R個, Rは正解の数)を算出

品詞タグ不要！ →

品詞タグ必要

手法	適合率
$TF\text{-}IDF_{N-gram}$	0.377
名詞句抽出＋各単語のTF-IDFの和 [Hasan+, COLING10]	0.386
名詞句抽出(全部小文字)＋各単語のTF-IDFの和 [Hasan+, COLING10]	0.369
名詞句抽出＋各単語のTF-IDFの平均	0.367
名詞句抽出(全部小文字)＋各単語のTF-IDFの平均	0.352
名詞句抽出＋単語N-gram自体のTF-IDF	0.369
名詞句抽出(全部小文字)＋単語N-gram自体のTF-IDF	0.355
TF-IDF (単語のみ抽出)	0.229

評価2: クエリ分割

情報検索ベースのクエリ分割データセット [Roy+, SIGIR12] を利用

分割した語句をWeb検索の条件とし，得られた検索結果を評価

nDCG: 正解が上位にあるほど高いスコア

	手法	nDCG（上位5位）	nDCG（上位10位）
外部情報不要！ →	IDF_{N-gram}	0.730	0.742
外部情報必要	Mishra手法（クエリログ使用） [Mishra+, WWW11]	0.706	0.737
	Mishra手法＋Wikipediaタイトル [Roy+, SIGIR12]	0.725	0.750
	PMI（クエリログ使用） [Roy+, SIGIR12]	0.716	0.736
	PMI（Webコーパス使用） [Roy+, SIGIR12]	0.670	0.707
	分割なし	0.655	0.689

本研究の貢献

① IDFに新たな理論的説明を付与

語を，それが出現する文書の集合で表現し，シャノン符号化により文書集合のコルモゴロフ複雑性を近似したとき，語と空文字との情報距離がIDFとなる

② ①に基づきIDFを単語N-gramに適用した手法を提案

単語N-gramに対応した重み付け手法を提案し，さらに，重みのみを用いて任意の長さの特徴語をテキスト中から抽出できることを発見した

③ ②を計算する手法を実装

接尾辞木とウェーブレット木を用いた計算方法を実装・公開した
http://github.com/iwnsew/ngweight
（計算済みのデータ，デモシステムを含む）

今日は話しませんので詳しくは論文で

④ ②の頑健性を評価実験にて確認

特徴語抽出およびクエリ分割において，追加の情報を必要とする既存手法と同等の性能を，IDFの重みのみを用いて達成した

今後の課題

IDF_{N-gram}の計算にかかる処理速度の向上

日本語などの単語間の区切りがない言語への対応

本研究のまとめ

① IDFに新たな理論的説明を付与

語を，それが出現する文書の集合で表現し，シャノン符号化により文書集合のコルモゴロフ複雑性を近似したとき，語と空文字との情報距離がIDFとなる

② ①に基づきIDFを単語N-gramに適用した手法を提案

単語N-gramに対応した重み付け手法を提案し，さらに，重みのみを用いて任意の長さの特徴語をテキスト中から抽出できることを発見した

③ ②を計算する手法を実装

接尾辞木とウェーブレット木を用いた計算方法を実装・公開した
http://github.com/iwnsew/ngweight
（計算済みのデータ，デモシステムを含む）

今日は話しませんので詳しくは論文で

④ ②の頑健性を評価実験にて確認

特徴語抽出およびクエリ分割において，追加の情報を必要とする既存手法と同等の性能を，IDFの重みのみを用いて達成した