N-gram IDF: A Global Term Weighting Scheme Based on Information Distance

Masumi Shirakawa, Takahiro Hara, Shojiro Nishio

Osaka University, JAPAN

24th International World Wide Web Conference (WWW 2015)
May 18-22, 2015, Florence, Italy.

Contributions

① Give a new explanation of IDF.

② Propose a new IDF scheme that can handle N-grams of any N.

③ Propose an implementation of ②.

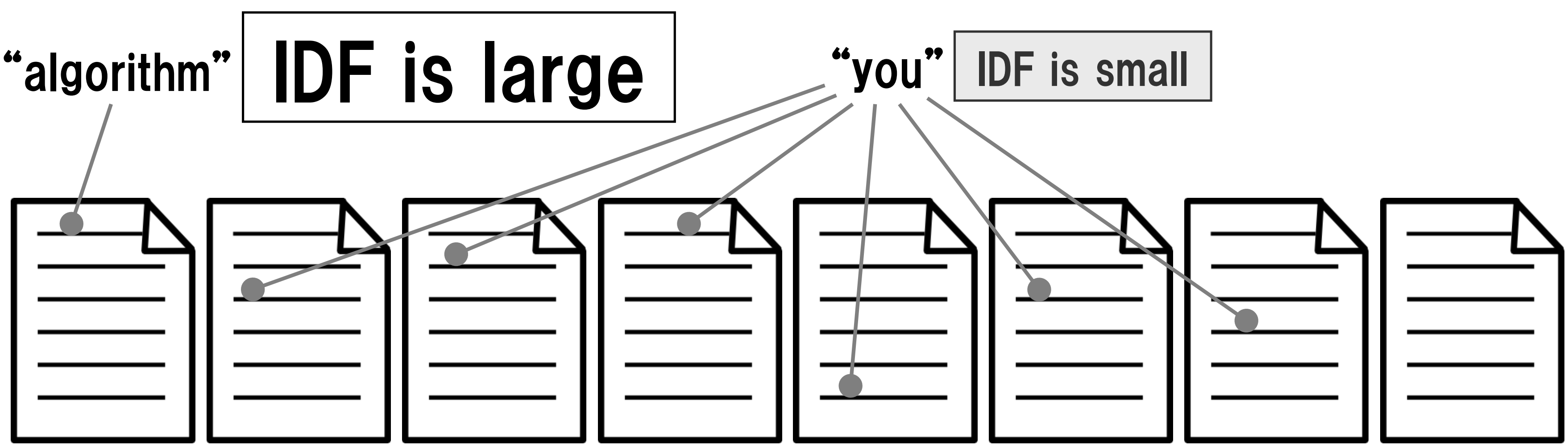
④ Exemplify the potential of ②.

Inverse Document Frequency (IDF)

Give more weight to a term occurring in less documents

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

t : Term
$df(t)$: Document frequency of t
$|D|$: Number of documents in D



Weak point of IDF

IDF does not work well for N-grams（phrases）.

WHY?

N-gram occurring in less documents is more likely to be a key term.
N-gram of unnatural collocation occurs in less documents.

⬇

N-gram of unnatural collocation is more likely to be a key term.

Estimated DF using Web Search

$$df(\text{"Leonardo da Vinci"}) = 31{,}700{,}000$$

$$df(\text{"Leonardo da is"}) = 15$$

The definition of IDF totally contradicts the definition of good phrases.

Multiword Expression（MWE）

MWE is a major research topic in Natural Language Processing（NLP）.

IDF has been developed in Information Retrieval（IR）.

Representative measures of MWE:

- ✓ PMI
- ✓ SCP [Silva+, MOL99]
- ✓ EMI [Zhang+, Expert Systems with Applications, 2009]
- ✓ MED [Bu+, COLING10]
- ✓ …

There was no theoretical explanation to connect term weighting with MWE ⟶ we have done.

Key Theories

Kolmogorov Complexity
and
Information Distance

Kolmogorov Complexity (1/2)

[Kolmogorov, Sankhya, 63]

Measure of the randomness of a (bit) string

$K(x)$: Kolmogorov complexity of x

Q$_1$: Which one has larger complexity?

$x_1 =$"010101010101010" → "01"×7+"0"

$x_2 =$"011101100010110" → "011101100010110"

A$_1$: Probably $K(x_1) < K(x_2)$

Kolmogorov Complexity (2/2)

[Kolmogorov, Sankhya, 63]

Measure of the randomness of a (bit) string

$K(x|y)$: **conditional Kolmogorov complexity of x given y**

Q$_2$: Which one has larger complexity? (y can be used)

$y =$"01110110001011"

$x_1 =$"010101010101010" → "01"×7+"0"

$x_2 =$"011101100010110" → y+"0"

A$_2$: Probably $K(x_1|y) > K(x_2|y)$

$K(x, y) = K(x|y) + K(y) = K(y|x) + K(x)$

(One string can be reused to describe the other)

Information Distance

[Bennett+, IEEE ToIT, 98]

Universal distance defined by Kolmogorov complexity

$E(x,y)$: information distance between x and y

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

It is equal to energy to convert one string to the other.

Energy cost to convert one bit $= 1kT\ln(2)$

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

Absolute temperature in Kelvin

Boltzmann constant

Indispensable Work

Multiword Expression Distance

Multiword Expression Distance (MED)

[Bu+, COLING10]

Measure of MWE based on information distance

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

g: N-gram $w_1 \cdots w_N$ (w: word)

Information distance between *context* and *semantic*

Context of g: set of documents containing g
Semantic of g: set of documents containing $w_1, \cdots, w_N$

Inspired by Normalized Google Distance [Cilibrasi+, TKDE, 2007]

Derivation of MED (1/2)

We assume that the probability of context x is proportional to its cardinality $|x|$.

set of documents

Num of documents

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

X: set of all contexts

Then we can approximate the Kolmogorov complexity using Shannon-Fano coding*1. [Li+2008]

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

Shannon -Fano coding

$aabbbbbc$

Sort

$bbbbbaac$

0 | 1

0 | 1

Encode

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

Derivation of MED (2/2)

Information distance between $\underbrace{\phi(g)}_{\text{Context of } g}$ and $\underbrace{\mu(g)}_{\text{Semantic of } g}$ is

$$
\begin{aligned}
&E\big(\phi(g), \mu(g)\big) \\
&= K\big(\phi(g), \mu(g)\big) - \min\Big(K\big(\phi(g)\big), K\big(\mu(g)\big)\Big) \\
&= -\log P\big(\phi(g), \mu(g)\big) + \min\Big(\log P\big(\phi(g)\big), \log P\big(\mu(g)\big)\Big) \\
&= -\log|\phi(g) \cap \mu(g)| + \max(\log|\phi(g)|, \log|\mu(g)|) \\
&= -\log\underbrace{|\phi(g)|}_{\text{DF of } g\ (= w_1 \cdots w_N)} + \log\underbrace{|\mu(g)|}_{\text{DF of } w_1, \cdots, w_N}
\end{aligned}
$$

Shannon –Fano coding

Proportional to the cardinality

$\phi(g) \subseteq \mu(g)$

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

IDF and information distance

IDF

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

Information distance between contexts of g and <u>empty string ε</u>

ε is contained in all documents

$$\begin{aligned}
&E\big(\phi(g), \phi(\varepsilon)\big) \\
&= -\log|\phi(g) \cap \phi(\varepsilon)| + \max(\log|\phi(g)|, \log|\phi(\varepsilon)|) \\
&= -\log|\phi(g)| + \log|\phi(\varepsilon)| \\
&= \log \frac{df(\varepsilon)}{df(g)} = \log \frac{|D|}{df(g)} = \boldsymbol{IDF(g)}
\end{aligned}$$

Shannon-Fano coding & proportional to the cardinality

$\phi(g) \subseteq \phi(\varepsilon)$

Contributions

① Give a new explanation of IDF.

IDF is equal to the distance between the term and the empty string in the information distance space.

② Propose a new IDF scheme that can handle N-grams of any N.

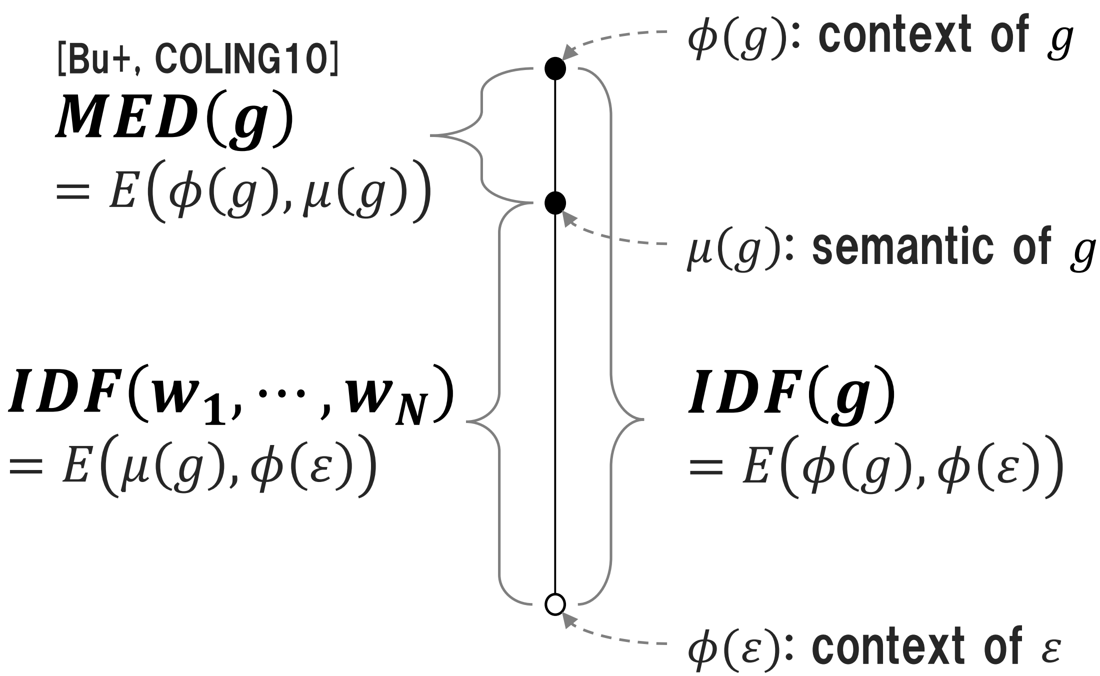
③ Propose an implementation of ②.

④ Exemplify the potential of ②.

IDF and MED in information distance space

IDF of N-gram g is the sum of IDF of $w_1, \cdots, w_N$ and MED of g

However, large MED means g' s collocation is unnatural.

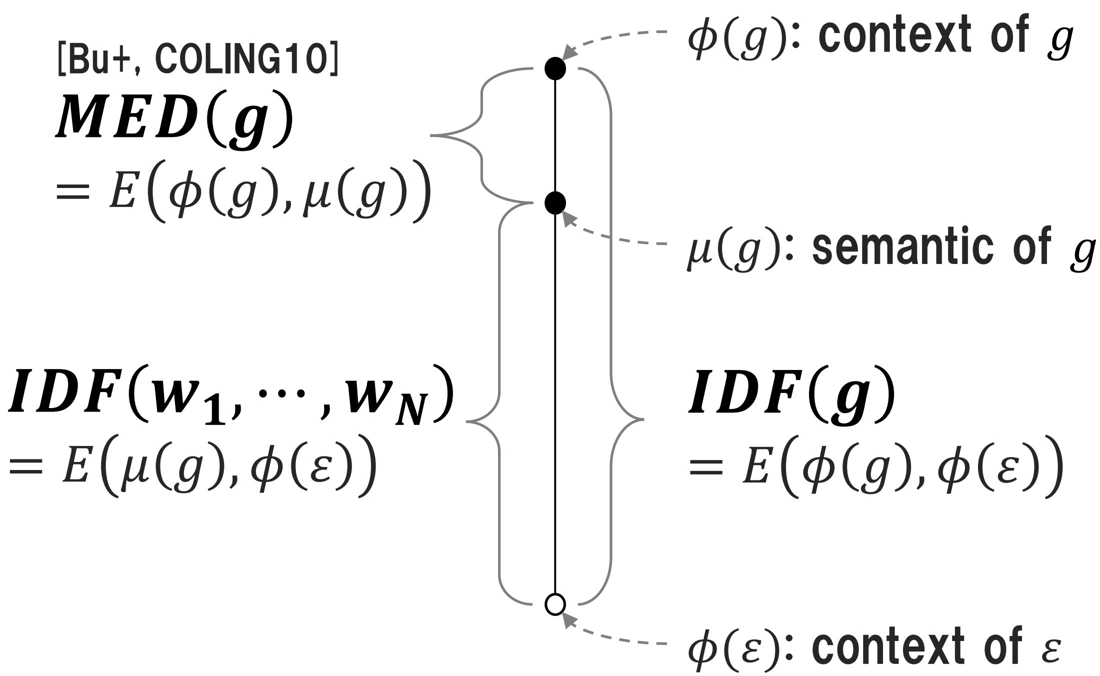


N-gram IDF

We redesign IDF for N-grams.

Larger IDF and smaller MED is better.

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



Key Term Extraction using N-gram IDF

Input: "Alice's Adventures in Wonderland – Kindle edition by Lewis Carroll"

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

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

Contributions

① Give a new explanation of IDF.

IDF is equal to the distance between the term and the empty string in the information distance space.

② Propose a new IDF scheme that can handle N-grams of any N.

IDF and MED are connected in the proposed scheme. Also, it is capable of extracting key N-grams from texts without using NLP techniques.

③ Propose an implementation of ②.

④ Exemplify the potential of ②.

Calculation of N-gram IDF

Calculation of the document frequency of $w_1, \cdots, w_N$ (set of words) requires much computational cost.

Bu et al. [COLING10] just used Web search engine for their experiments.

➡ **Wavelet tree**
[Gagie+, TCS, 12]

How to determine N is unclear.

➡ **Suffix tree (or enhanced suffix array)**
[Okanohara+, SDM09]

Please refer to our paper for the detail.

Implementation, Data, and Demo

Code to calculate N-gram IDF for all N-grams

https://github.com/iwnsew/ngweight

Processed English Wikipedia (Oct 1, 2013)

http://mljournalism.com/ngw/ngram.bz2

It took 12 days to process whole Wikipedia

Online demo

http://mljournalism.com/ngw/

Or, search “N-gram TF-IDF”.

Contributions

① Give a new explanation of IDF.

IDF is equal to the distance between the term and the empty string in the information distance space.

② Propose a new IDF scheme that can handle N-grams of any N.

IDF and MED are connected in the proposed scheme. Also, it is capable of extracting key N-grams from texts without using NLP techniques.

③ Propose an implementation of ②.

Two cutting-edge string processing algorithms were combined.
https://github.com/iwnsew/ngweight

④ Exemplify the potential of ②.

Evaluation 1: Key Term Extraction

Use 1,678 first paragraphs of English Wikipedia

Use Anchor texts and bold faces as correct labels R

Just use the weight!

Require POS tags

Method	R-Precision
$TF\text{–}IDF_{N-gram}$	0.377
Noun Phrase + TF-IDF sum [Hasan+, COLING10]	0.386
Noun Phrase (no capitalization) + TF-IDF sum [Hasan+, COLING10]	0.369
Noun Phrase + TF-IDF average	0.367
Noun Phrase (no capitalization) + TF-IDF average	0.352
Noun Phrase + TF-IDF	0.369
Noun Phrase (no capitalization) + TF-IDF	0.355
TF-IDF (word only)	0.229

Evaluation 2: Query Segmentation

Use IR-based query segmentation dataset [Roy+, SIGIR12]

Evaluate search results by using segmented phrases

nDCG: normalized Discounted Cumulative Gain

Method	nDCG (Top 5)	nDCG (Top 10)
IDF_{N-gram}	0.730	0.742
Mishra (use query logs) [Mishra+, WWW11]	0.706	0.737
Mishra + Wikipedia titles [Roy+, SIGIR12]	0.725	0.750
PMI (use query logs) [Roy+, SIGIR12]	0.716	0.736
PMI (use Web corpus) [Roy+, SIGIR12]	0.670	0.707
No segmentation	0.655	0.689

Just use the weight!

Require query logs or external knowledge

Contributions and Conclusion

① Give a new explanation of IDF.

IDF is equal to the distance between the term and the empty string in the information distance space.

② Propose a new IDF scheme that can handle N-grams of any N.

IDF and MED are connected in the proposed scheme. Also, it is capable of extracting key N-grams from texts without using NLP techniques.

③ Propose an implementation of ②.

Two cutting-edge string processing algorithms were combined.
https://github.com/iwnsew/ngweight

④ Exemplify the potential of ②.

On key term extraction and query segmentation tasks, N-gram IDF achieved competitive performance with task-oriented methods.

Future work

Efficient computation of N-gram IDF

Supporting languages without spaces between words such as Japanese and Chinese

Theoretical explanation of TF

THANKS!!

N-gram IDF: A Global Term Weighting Scheme Based on Information Distance
Masumi Shirakawa, Takahiro Hara, Shojiro Nishio
Osaka University, JAPAN

24th International World Wide Web Conference (WWW 2015)
May 18-22, 2015, Florence, Italy.

① Give a new explanation of IDF.

IDF is equal to the distance between the term and the empty string in the information distance space.

② Propose a new IDF scheme that can handle N-grams of any N.

IDF and MED are connected in the proposed scheme. Also, it is capable of extracting key N-grams from texts without using NLP techniques.

③ Propose an implementation of ②.

Two cutting-edge string processing algorithms were combined.
https://github.com/iwnsew/ngweight

④ Exemplify the potential of ②.

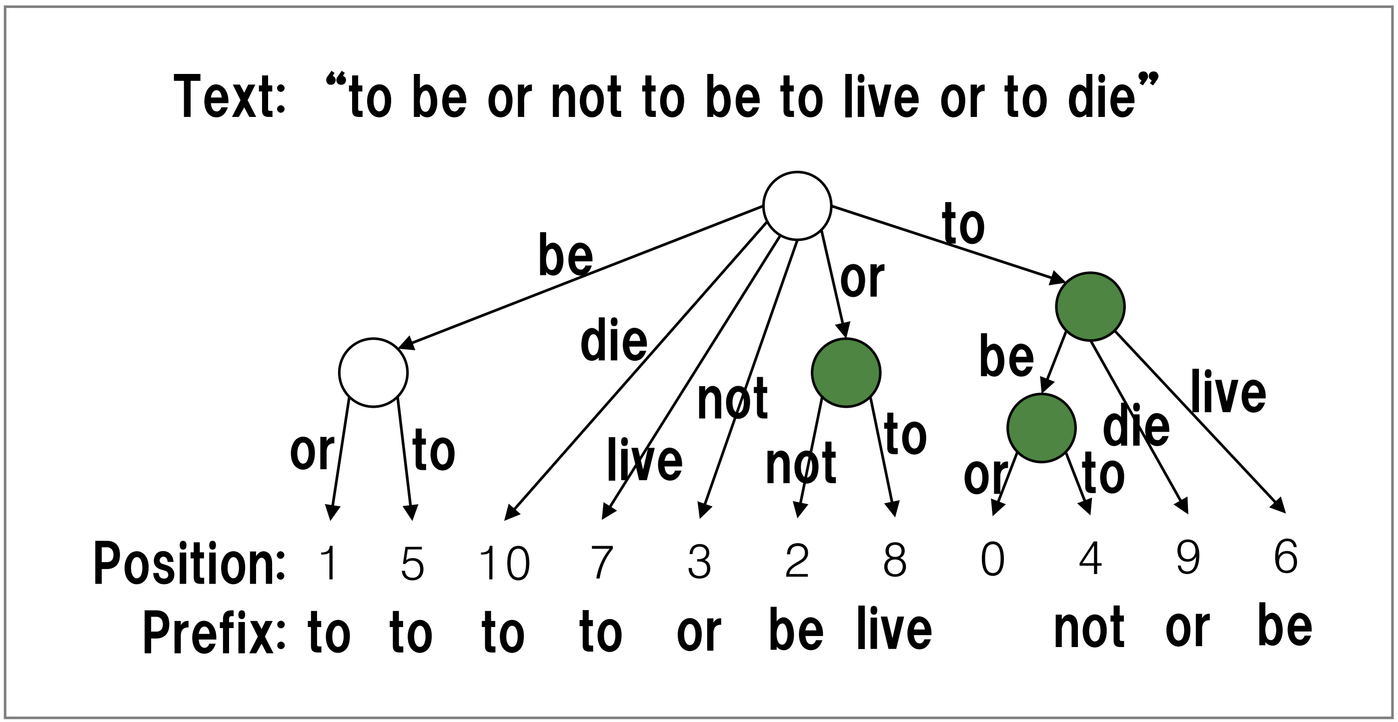
On key term extraction and query segmentation tasks, N-gram IDF achieved competitive performance with task-oriented methods.

Appendix

Suffix tree for enumerating valid N-grams
[Okanohara+, SDM09]

Intermediate nodes having multiple prefixes = valid N-gram

Number of valid N-grams is proved to be linearly proportional to text length.



Wavelet tree for DF counting

[Gagie+, TCS, 12]

The most efficient algorithm for counting DF for a set of words

Document set: $D = \{\text{a,b,c,d}\}$

a = **“to be”**, b = **“or not to be”**, c = **“to live”**, d = **“or to die”**

Position:	1	5	10	7	3	2	8	0	4	9	6
Document ID:	a	b	d	c	b	b	d	a	b	d	c

abdcbbdabdc
00110010011

0 / 1

abbbab   dcddc
011101   10110

0 1   0 1

aa   bbbb   cc   ddd

$\approx \log|D|$

Example of DF counting

Keep beginning and end positions of each word, and traverse the tree toward leaves.

Document set: $D = \{\mathrm{a,b,c,d}\}$
a = **"to be"**, b = **"or not to be"**, c = **"to live"**, d = **"or to die"**

Query: "to" "be"
Results: a, b