Dr. Ali A. Amer

New similarity

Hypothesis

Our combined similarity is the closest to cosine similarity and in some cases behaves better

Our combined similarity is much better than SMTP in almost all cases, but in some cases SMTP behaves better

Part (1): for binary representation

$$Sim\ (Doc1, Doc2) = \frac{\left(1 - \frac{F}{N}\right) + \left(\frac{2N_{ab}}{N_a + N_b}\right)}{2}$$

Nab is the number of common words in both documents

N is the number of all words under consideration

Na or Nb is the length of doc a and doc b respectively

F is the number of differences between them

This measure just take the presence and absence of words irrespective of TF/IDF values.

Ex1:

	W1	w2	w3	w4	w5	w6
Doc1	2	5	7	8	0	9
Doc2	9	0	0	6	5	1

Suppose this TF/IDF

To convert this TF/IDF table into binary representation (0/1):

	W1	w2	w3	w4	w5	w6
Doc1	1	1	1	1	0	1
Doc2	1	0	0	1	1	1

N12 = 3, common words

N= 6

F=3

Na=5

Nb=4

Now,

$$Sim\ (Doc1, Doc2) = \frac{\left(1 - \frac{3}{6}\right) + \left(\frac{6}{5+4}\right)}{2} = \ 0.58$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM of binary representation only (0.58)

0.5 * Cosine + 0.5 * CSM of binary representation only (0.58)

0.5 * Euclidean + 0.5 * CSM of binary representation only (0.58)

The same goes for all examples in binary representation

Ex2:

	W1	w2	w3	w4	w5	w6
Doc1	2	5	7	8	0	9
Doc2	9	0	3	6	5	1

Suppose this TF/IDF

0/1 conversion is

	W1	w2	w3	w4	w5	w6
Doc1	1	1	1	1	0	1
Doc2	1	0	1	1	1	1

N12 = 4, common words

N= 6

F=2

Na=5

Nb=5

Now,

$$Sim\ (Doc1, Doc2) = \frac{\left(1 - \frac{2}{6}\right) + \left(\frac{8}{5+5}\right)}{2} = 0.73$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM

0.5 * Cosine + 0.5 * CSM

0.5 * Euclidean + 0.5 * CSM

Ex3:

	W1	w2	w3	w4	w5	w6
Doc1	2	5	7	8	0	9
Doc2	9	6	3	6	5	1

Suppose this TF/IDF

0/1 conversion is

	W1	w2	w3	w4	w5	w6
Doc1	1	1	1	1	0	1
Doc2	1	1	1	1	1	1

N12 = 5, common words

N= 6

F=1

Na=5

Nb=6

Now,

$$Sim\ (Doc1, Doc2) = \frac{\left(1 - \frac{1}{6}\right) + \left(\frac{10}{5+6}\right)}{2} = \ 0.87$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM

0.5 * Cosine + 0.5 * CSM

0.5 * Euclidean + 0.5 * CSM

Ex4:

	W1	w2	w3	w4	w5	w6
Doc1	2	5	7	8	4	9
Doc2	9	6	3	6	5	1

Suppose this TF/IDF

0/1 conversion is

	W1	w2	w3	w4	w5	w6
Doc1	1	1	1	1	1	1
Doc2	1	1	1	1	1	1

N12 = 6, common words

N= 6

F=0

Na=6

Nb=6

Now,

$$Sim\ (Doc1, Doc2) = \frac{\left(1 - \frac{0}{6}\right) + \left(\frac{12}{6+6}\right)}{2} = 1$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM

0.5 * Cosine + 0.5 * CSM

0.5 * Euclidean + 0.5 * CSM

Ex 5:

0, 2, 1, 1, 0, 1

3, 1, 1, 1, 1, 0

0/1 conversion is

0, 1, 1, 1, 0, 1

1, 1, 1, 1, 1, 0

N12 = 3, common words

N= 6

F=3

Na=4

Nb=5

Now,

$$Sim\ (Doc1, Doc2) = \frac{\left(1 - \frac{3}{6}\right) + \left(\frac{6}{4+5}\right)}{2} = \ 0.58$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM

0.5 * Cosine + 0.5 * CSM

0.5 * Euclidean + 0.5 * CSM

The larger the length of document, the better results s obtained using our similarity

Combined this value with SMTP the similarity is (0.6 + .4908) / 2= 0.54 which is better than SMTP

Unlike SMTP, In our similarity when word is absent in both document, word is neglected (excluded out of equation)

In other words, our measure take the presence of word either in both document or in either one of them.

Ex 6:

1; 1; 3

1; 0; 2

Conversion is

1; 1; 1

1; 0; 1

N12 = 2, common words

N= 3

F=1

Na=3

Nb=2

Now,

$$Sim\ (Doc1, Doc2) = \frac{\left(1 - \frac{1}{3}\right) + \left(\frac{4}{3+2}\right)}{2} = 0.73$$

Now I need the values of the next measures for the same example

SMTP =

Cosine =

Euclidean =

0.5 * SMTP + 0.5 * CSM =

0.5 * Cosine + 0.5 * CSM =

0.5 * Euclidean + 0.5 * CSM =

Part (2): Numerical TF/IDF representation

$$Sim\ (Doc1, Doc2) = \left(1 - \frac{(F+1)}{N}\right)$$

$$F = X * Y$$

$$X = \sum D_{ik} \quad D_{ik} > 0\ and\ D_{jk} = 0$$

$$Y = \sum D_{jk}\ D_{\ jk} > 0\ and\ D_{ik} = 0$$

$$N = \sum D_{ik} * \sum D_{jk} \quad D_{ik} > 0\ and\ D_{jk} > 0$$

Also, as axiom taken for granted, similarity (Doc1, Doc2) =1, if D_{ik}= D_{jk} and both D_{ik}, D_{jk}>0

N is the number of all words under consideration

F is the summation of all TF/IDF as one of both document have a zero value

Ex 1:

	W1	w2	w3	w4	w5	w6
Doc1	2	5	7	8	0	9
Doc2	9	0	0	6	5	1

Suppose this TF/IDF

Now,

X=5 + 7 =12

Y= 5

F= 12 * 5 = 60

N= (2+5+7+8+9) * (9+6+5+1) = 31 * 21 = 651

$$Sim\ (Doc1, Doc2) = \left(1 - \frac{(60 + 1)}{651}\right) = 0.91$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM

0.5 * Cosine + 0.5 * CSM

0.5 * Euclidean + 0.5 * CSM

Combined Similarity Measure (CSM) of TF/IDF dependent and binary dependent: (0.91 + 0.58)/2 = 0.75

Ex2:

	W1	w2	w3	w4	w5	w6
Doc1	2	5	7	8	0	9
Doc2	9	0	3	6	5	1

Suppose this TF/IDF

Now,

X=5

Y= 5

F= 5 * 5 = 25

N= (2+5+7+8+9) * (9+3+6+5+1) = 31 * 24 =744

$$Sim\ (Doc1, Doc2) = \left(1 - \frac{(25+1)}{744}\right) = 0.96$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM

0.5 * Cosine + 0.5 * CSM

0.5 * Euclidean + 0.5 * CSM

Combined similarity of TF/IDF dependent and binary dependent: (0.96 + 0.73)/2 = 0.85

Ex3:

	W1	w2	w3	w4	w5	w6
Doc1	2	5	7	8	0	9
Doc2	9	6	3	6	5	1

Suppose this TF/IDF

X=0

Y= 5

F= 12 * 5 = 0

N= (2+5+7+8+9) * (9+6+3+6+5+1) = 31 * 30 =930

$$Sim\ (Doc1, Doc2) = \left(1 - \frac{(0+1)}{930}\right) = 0.99$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM

0.5 * Cosine + 0.5 * CSM

0.5 * Euclidean + 0.5 * CSM

Combined similarity of TF/IDF dependent and binary dependent: (0.99 + 0.87)/2 = 0.93

Ex4:

	W1	w2	w3	w4	w5	w6
Doc1	2	5	7	8	4	9
Doc2	9	6	3	6	5	1

Suppose this TF/IDF

Now,

X=0

Y= 0

F= 0

N= (2+5+7+8+9) * (9+6+5+1) = 35 * 30 =1050

$$Sim\ (Doc1, Doc2) = \left(1 - \frac{(0+1)}{1050}\right) = 0.99$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM

0.5 * Cosine + 0.5 * CSM

0.5 * Euclidean + 0.5 * CSM

Combined similarity of TF/IDF dependent and binary dependent: (0.99 + 0.1)/2 = 0.99

Ex 5:

0, 2, 1, 1, 0, 1

3, 1, 1, 1, 1, 0

Now,

X=1

Y= 4

F= 4

N= (2+1+1+1) * (3+1+1+1+1) = 5 * 7 =35

$$Sim\ (Doc1, Doc2) = \left(1 - \frac{(4+1)}{35}\right) = 0.86$$

Now I need the values of the next measures for the same example

SMTP

Cosine

Euclidean

0.5 * SMTP + 0.5 * CSM

0.5 * Cosine + 0.5 * CSM

0.5 * Euclidean + 0.5 * CSM

Combined similarity of TF/IDF dependent and binary dependent: (0.86 + 0.58)/2 = 0.72

Ex 6:

1; 1; 3

1; 0; 2

Now,

X=1

Y= 0

F= 0

N= 5* 3 = 15

$$Sim\ (Doc1, Doc2) = \left(1 - \frac{(0+1)}{15}\right) = 0.93$$

Now I need the values of the next measures for the same example

SMTP =

Cosine =

Euclidean =

0.5 * SMTP + 0.5 * CSM =

0.5 * Cosine + 0.5 * CSM =

0.5 * Euclidean + 0.5 * CSM =

Combined similarity of binary and TF/IDF dependent: (0.93 + 0.73)/2 = 0.83

Finally

similarity (Doc1, Doc2) =1, if D_{ik}= D_{jk} and both D_{ik}, D_{jk}>0

ex

Doc1 2 2 2

Doc2 2 2 2

Combined similarity = (1 for binary representation + 1 for TF/IDF representation)/2=1

Optimal case satisfaction

This case happened when N =Na=Nb=Nab and F is zero value, therefore the similarity equation would as follows (replacing all symbols by N)

For equation 1 (Binary Dependent)

$$Sim\ (Doc1, Doc2) = \frac{\left(1 - \frac{0}{N}\right) + \left(\frac{2N}{N+N}\right)}{2} = \frac{1 + \left(\frac{2N}{2N}\right)}{2} = \frac{2}{2} = 1$$

Similarly for equation 2 (TF/IDF Dependent): because of the fact that N =Na=Nb=Nab and F is zero value which means that the axiom presented in general similarity equation as $D_{ik}=D_{jk}$, is met, the similarity is "1" in this case.

Finally the combined similarity is (1+1)/2=1