A Comment on "A Similarity Measure for Text Classification and Clustering"

Naresh Kumar Nagwani, *Member, IEEE*

1 Introduction

A similarity measure namely, similarity measure for text processing (SMTP) is proposed by Lin et al. [1] for knowledge discovery on text collection. The proposed measure considered the three cases for similarity measurements between the pairs of documents. These cases are based on absence and presence of features in the pair of text documents. The first case covers the features appearing in both of the documents, second case covers the features appears in only one document and the third case covers the features appears in none of the documents. The proposed similarity measure considered to be ideal for finding similarity between the pair of text documents on the basis of presence or absence of features available in text documents, however, while exploring the SMTP similarity measurement it is found that the case of measuring similarity between the pair of similar documents is not covered. The objective of this work is to highlight this gap and propose a minor change to make the SMTP a complete similarity measurement technique for knowledge discovery in line with the other standard similarity techniques.

2 Discussions of Published Methodology

The SMTP similarity measurement technique satisfies a number of properties mentioned in the work [1]. For example the proposed technique satisfies the symmetry property as $S_{SMTP}(d_i, d_j) = S_{SMTP}(d_j, d_i)$. However the case of similar documents (documents with similar features) is not covered in the proposed similarity technique. In other words the case where the standard deviation for a particular feature tending to (or equal to) zero is not covered in the proposed similarity measure SMTP. This work highlights this missing case to make SMTP similarity measurement complete so that the similarity between the similar pair of documents can be measured more accurately just like other existing similarity techniques.

Given two documents $d_i = \langle d_{i1}, d_{i2}, \ldots, d_{im}\rangle$ and $d_j = \langle d_{j1}, d_{j2}, \ldots, d_{jm}\rangle$, the S_{SMTP} similarity measure is given by the Eq. (1), where the function F is given by Eq. (2):

$$S_{SMTP}(d_i, d_j) = \frac{F(d_i, d_j) + \lambda}{1 + \lambda}, \tag{1}$$

$$F(d_i, d_j) = \frac{\sum_{k=1}^{m} N_*(d_i, d_j)}{\sum_{k=1}^{m} N_\cup(d_i, d_j)}. \tag{2}$$

Where

$$N_*(d_{ik}, d_{jk}) = \begin{cases} 0.5 \times \left(1 + \exp\left\{-\left(\frac{d_{ik}-d_{jk}}{\sigma_k}\right)^2\right\}\right) \\ \quad if\ d_{ik}d_{jk} > 0 \\ 0, \quad if\ d_{ik} = 0\ and\ d_{jk} = 0 \\ -\lambda,\ otherwise. \end{cases} \tag{3}$$

Where σ_k is the standard deviation of all the non-zero values for the feature w_k

$$N_\cup(d_{ik}, d_{jk}) = \begin{cases} 0, & if\ d_{ik} = 0\ and\ d_{jk} = 0 \\ 1, & otherwise. \end{cases} \tag{4}$$

Suggestions for setting the values of λ are also given in [2]. It is suggested that $\lambda = 1$, $\lambda = m$ (the number of features) and $\lambda = AL$ (Average Length of training documents) can be considered, but still this will not suffice the case of measurement of similar documents. The case of similar document measurement is explained now using an illustrative example. Let us consider there are two similar documents $d_1 = \langle 1, 1, 1\rangle$ and $d_2 = \langle 1, 1, 1\rangle$ with three similar features. Then process of similarity between the pair of the documents $d^{(1)}$ and $d^{(2)}$ using SMTP technique is given by Eq. (5), Eq. (6) and Eq. (7):

$$F(d_1, d_2) = \frac{\begin{array}{c} 0.5 \times \left(1 + \exp^{\left\{-\left(\frac{1-1}{0}\right)^2\right\}}\right) + 0.5 \times \left(1 + \exp^{\left\{-\left(\frac{1-1}{0}\right)^2\right\}}\right) \\ + 0.5 \times \left(1 + \exp^{\left\{-\left(\frac{1-1}{0}\right)^2\right\}}\right) \end{array}}{1 + 1 + 1}. \tag{5}$$

Since $\exp^{\{-(\frac{1-1}{0})^2\}}$ is not a number (or not defined) considering this part as 0 will give the

$$F(d_1, d_2) = \frac{0.5 \times (1) + 0.5 \times (1) + 0.5 \times (1)}{1 + 1 + 1} = \frac{1.5}{3} = 0.5, \tag{6}$$

$$S_{SMTP}(d_1, d_2) = \frac{0.5 + 1}{1 + 1} = 0.75. \tag{7}$$

However the other similarity technique like Cosine and Jaccard similarity [3] for the similar pair of documents is given in Eq. (8) and Eq. (9):

$$S_{Cos}(d_1, d_2) = \frac{1 \times 1 + 1 \times 1 + 1 \times 1}{\sqrt{1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 1^2 + 1^2}} = 1, \tag{8}$$

$$S_{Jac}(d_1, d_2) = \frac{1 + 1 + 1}{1 + 1 + 1} = 1. \tag{9}$$

So it seems that the condition where the feature is exactly the same (or in other words standard deviation is zero for a feature) is not covered in the SMTP similarity technique. In order to overcome this limitation one supplementary condition is suggested in $N_*(d_{ik}, d_{jk})$ for the SMTP similarity technique as given in the Eq. (10). However the equations for $N_\cup(d_{ik}, d_{jk})$, $F(d^{(1)}, d^{(2)})$ and $S_{SMTP}(d_i, d_j)$ remains the same:

$$N_*(d_{ik}, d_{jk}) = \begin{cases} 1,\ if\ d_{ik} = d_{jk}\ and\ d_{ik}, d_{jk} > 0 \\ 0.5 \times \left(1 + \exp\left\{-\left(\frac{d_{ik}-d_{jk}}{\sigma_k}\right)^2\right\}\right) if\ d_{ik}d_{jk} > 0 \\ 0,\ if\ d_{ik} = 0\ and\ d_{jk} = 0 \\ -\lambda,\ otherwise. \end{cases} \tag{10}$$

Now the SMTP similarity between the pair of documents is given by the Eq. (11) and Eq. (12):

$$F(d_1, d_2) = \frac{1 + 1 + 1}{1 + 1 + 1} = \frac{3}{3} = 1.0, \tag{11}$$

$$S_{SMTP}(d_1, d_2) = \frac{1.0 + 1}{1 + 1} = 1.0. \tag{12}$$

• *The author is with the National Institute of Technology Raipur, G E Road, Raipur 492010, India. E-mail: nknagwani.cs@nitrr.ac.in.*

Manuscript received 18 Apr. 2015; revised 21 May 2015; accepted 25 June 2015; date of current version 3 Aug. 2015.
Recommended for acceptance by L. Chen.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2451616

Another example favoring the proposed changes is, let us consider two similar documents $d_1 = \langle 1, 1, 3 \rangle$ and $d_2 = \langle 1, 0, 2 \rangle$ with three features. Then similarity between the pair of the documents $d^{(1)}$ and $d^{(2)}$ using SMTP technique is 0.7725, whereas Cosine similarity is 0.944 and Jaccard similarity is 0.67. The SMTP similarity between the pair of documents using the suggested changes is calculated as 0.856, which is higher than the existing similarity measure and close to the Cosine and Jaccord similarity measures.

The work presented by Lin et al. [1] also discussed about the similarity between the document sets G_1 and G_2, $S_{SMTP}\,(G_1, G_2)$, as given in the Eq. (13), where $F(G_1, G_2)$ is given by Eq. (14):

$$S_{SMTP}\,(G_1, G_2) = \frac{F(G_1, G_2) + \lambda}{1 + \lambda}, \tag{13}$$

$$F\,(G_1, G_2) = \frac{\sum_{k=1}^{m} \sum_{i=1}^{q_1} \sum_{j=1}^{q_2} N_*(d_{ik}^1, d_{jk}^2)}{\sum_{k=1}^{m} \sum_{i=1}^{q_1} \sum_{j=1}^{q_2} N_\cup(d_{ik}^1, d_{jk}^2)}. \tag{14}$$

The numerator part of $F(G_1, G_2)$ can be simplified as given in the Eq. (15):

$$\begin{aligned} &\sum_{i=1}^{q_1} \sum_{j=1}^{q_2} N_*\big(d_{ik}^1, d_{jk}^2\big) \\ &\quad = \sum_{d_{ik}^1 > 0} \sum_{d_{jk}^2 > 0} N_*\left(d_{ik}^1, d_{jk}^2\right) + \sum_{d_{ik}^1 = 0} \sum_{d_{jk}^2 > 0} N_*\left(d_{ik}^1, d_{jk}^2\right) \\ &\qquad + \sum_{d_{ik}^1 > 0} \sum_{d_{jk}^2 = 0} N_*\left(d_{ik}^1, d_{jk}^2\right) + \sum_{d_{ik}^1 = 0} \sum_{d_{jk}^2 = 0} N_*\left(d_{ik}^1, d_{jk}^2\right). \end{aligned} \tag{15}$$

The proposed modification is also valid for measuring the similarity between the pair of document sets since again the similar document similarity case is missing in the SMTP similarity measure. After incorporating the suggested changes the modified measure for $N_*(d_{ik}^1, d_{jk}^2)$ will be given by Eq. (16), and will cover the case of measuring the similarity between the pairs of document sets also:

$$\begin{aligned} \sum_{i=1}^{q_1} \sum_{j=1}^{q_2} N_*\left(d_{ik}^1, d_{jk}^2\right) &= \sum_{d_{ik}^1 > 0} \sum_{d_{jk}^2 > 0} \sum_{d_{ik}^1 \neq d_{jk}^2} N_*\left(d_{ik}^1, d_{jk}^2\right) \\ &\quad + \sum_{d_{ik}^1 = 0} \sum_{d_{jk}^2 > 0} N_*\left(d_{ik}^1, d_{jk}^2\right) + \sum_{d_{ik}^1 > 0} \sum_{d_{jk}^2 = 0} N_*\left(d_{ik}^1, d_{jk}^2\right) \\ &\quad + \sum_{d_{ik}^1 = 0} \sum_{d_{jk}^2 = 0} N_*\left(d_{ik}^1, d_{jk}^2\right) + \sum_{d_{ik}^1 = d_{jk}^2} \sum_{d_{ik}^1 = d_{jk}^2} N_*\left(d_{ik}^1, d_{jk}^2\right). \end{aligned} \tag{16}$$

3 Conclusion

The case of measuring the similarity between the document pairs consisting of similar features was not described in SMTP measure. The proposed suggestion covers this case and justification for the suggestion is demonstrated with the help of illustrative examples. The suggested change is also compared in line with the two standard similarity measurement techniques namely Cosine and Jaccard similarity measures. From the examples it is demonstrated that the suggested change can be a definite improvement over the existing proposed SMTP technique by Lin et al. [1].

References

[1] Y.-S. Lin, J.-Y. Jiang, and S.-J. Lee, "A similarity measure for text classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1575–1590, Jul. 2014.

[2] J.-Y. Jiang, W.-H. Cheng, Y.-S. Chiou, and S.-J. Lee, "A similarity measure for text processing," in *Proc. Int. Conf. Mach. Learn. Cybernetics*, Guilin, China, Jul. 10–13, 2011, pp. 1460–1465.

[3] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 2006.