Grenze International Journal of Engineering and Technology, June Issue

New Similarity Measure for Web Documents Retrieval

Dr. Ramya C[1], Dr. Hamsha K[2] and Navya V K[3]
[1]Department of ISE , CMR Institute of Technology, Bengaluru, Karnataka, India
Email: cramyac@gmail.com
[2]Department of Computer Science, NMKRV College for Women, Bengaluru, Karnataka, India
Email: hamshak@gmail.com
[3]Department of CSE , MVJ College of Engineering, Bengaluru, Karnataka, India
Email: navya.vk@mvjce.edu.in

***Abstract*—Web information retrieval (WIR) faces the challenges in finding more relevant information due to unstructured characteristic of the data and the tremendous growth in the size of the information. As a result, the users in search of information face difficulties in obtaining specific information on the web. The similarity measures play significant role in serving the machines to deal with the natural language. Various similarity functions used in the context of WIR to match between query and the documents. In this study, a new version of similarity measure for documents retrieval (SMDR) has been proposed. The experiments are conducted on various queries over CACM and RCV1 document collections to evaluate and analyse the performance of SMDR. The standard particle swarm optimization algorithm is used to optimize the entire WIR process. The comparative study is carried out considering the three important similarity measures in the context of WIR such as Cosine, Dice and Jaccord. The results show that significant improvement in the performance of the proposed similarity measure in retrieving the set of relevant documents.**

***Index Terms*— Swarm Optimization, Web Information Retrieval, Similarity Measure, Cosine, Dice and Jaccard.**

I. INTRODUCTION

In recent times, due to tremendous accumulation of web documents over the Internet, there is a demand to an increasing extent to have effective and efficient techniques to retrieve documents from the web. With reference to this problem, similarity measures play a major role in serving the machines to deal with the natural language. The similarity measure is the measure of how much alike two data objects are. Similarity measure in a data mining context is a distance with dimensions representing features of the objects. The similarity is highly dependent and is subjective on the application and domain. Precaution should be engaged when scheming distance over the unrelated features/dimensions. The values of relation of every element must be normalized, or one dimension could result in dominating the calculation on distance. Similarity is measured in the range 0 to 1 i.e., [0, 1]. There is no similarity measure which is perpetually finest for all kind of documents retrieval problems [1]. Furthermore, it is crucial to choose suitable similarity measure for problems related to web documents retrieval. Hence, it is highly important to study the effectiveness of various similarity measures so that it helps to select the best one.

A great challenge for similarity measure is high dimensionality induced by large documents and sparsity caused

Grenze ID: 01.GIJET.8.2.32

by most of the values in the document vector being zero [2]. The existing measures such as Cosine, Dice and Jaccard measure would normally work as per the two intuitions of similarity measures such as commonality and difference between documents [3]. Although, the above mentioned similarity measures lose the information of word distribution as they don't consider the document structure information. Hence, there is a need for the novel similarity measure to gauge the similarity between query and documents (with low as well as high dimensionality), to take care of resulting vector being sparse, and to help in ranking the documents with high relevance, thereby contributing to the effectiveness of WIR process.

II. Background of Similarity Functions used in WIR

In this section, a brief review is presented as a recap of the researches carried out using similarity measures in the context of WIR. Yung-Shen Lin et al. did catch sight of a new similarity measure for text processing (SMTP) to evaluate the similarity between two text documents [2]. It takes three cases into consideration, the feature appearing in (a) both the documents (b) in only one document and (c) in none of the documents. The preferable properties embedded in SMTP are its symmetric nature, considering the presence or absence of a feature more significant, increased similarity when the number of presence-absence features pairs decrease. Naresh Kumar Nagwani highlighted the drawbacks of the proposed SMTP by Yung-Shen Lin et al. [4]. The case of similar documents i.e. documents with similar features and the condition where the feature is exactly the same are not taken care while satisfying the symmetry property. He suggested a slight change to make the SMTP a complete similarity measurement technique by suggesting a supplementary condition in $N_*(d_{ik}, d_{jk})$ for the SMTP and for setting the values for λ. This inspired Komal Maher et al. to use the same SMTP for text classification and clustering problems [5]. A comparative study based on the performances given by Euclidean distance, Cosine similarity and SMTP is presented. These three measures are applied in k-NN based classification, Naïve Bayes classification and k-means clustering methods to investigate their effectiveness. SMTP leads the others two measures in terms of accuracy, specificity and sensitivity. Experiments are carried out on WebKB, Reuters-8 and RCV1.

Mubarak Himmat et. al. imparted a new Ligand-based virtual screening similarity measure ASMTP (Adapted Similarity Measure of Text Processing) which is derived from document retrieval areas, to quantify the similarity of molecules [6]. The Effectiveness of the ASMTP is investigated on the two datasets Maximum Unbiased Validation (MUV) and the MDL Drug Data Report (MDDR). ASMTP's performance is compared with Tanimoto and the recent Standard Quantum-Based(SQB) measures and authors claim superior performance of the proposed measure than the existing measures. D Gunawan et. al. demonstrated the use of cosine similarity function to calculate the text relevance between the documents [7]. The results indicate the performance of Cosine is significant than Dice and Euclidian for text documents retrieval.

Maedeh Afzali et. al. analysed and compared the performances of the four dissimilarity measures such as Manhattan dissimilarity, Euclidean, Braycurtis and Canberra distance, and also similarity measures such as Cosine, extended Dice, extended Jaccord, and Pearson measure on four text document datasets such as BBCSport, BBC-fulltext, Classic and WebKB using k-means clustering method [8]. Naveen Kumar et. al. [9] studied on partitioned clustering for text documents retrieval. They analysed and evaluated the effectiveness of the various measures using standard K-means Algorithm for clustering. It has been observed in their studies that Euclidean distance measure is more effective for document clustering where as Pearson and cosine measures are more unbiased to obtain the clusters. Apart this, to find the rational clusters, Pearson and Jaccard measures are more advisable. Dan Goldwasser and Xiao Wang [10] proposed a novel measure Smooth Cosine Similarity (SCS) which is able to avoid gradients explodes and enforces the model to be stable for cross-lingual information retrieval process.

To compute the similarity of two documents, R. Lakshmi and S. Baskar [11] proposed two new similarity measures, namely distance of term frequency-based similarity measure (DTFSM) and presence of common terms-based similarity measure (PCTSM) on reuters-21578 and WebKB datasets for clustering the documents using K-means and K-means++ clustering algorithms. DTFSM and PCTSM are tend to yield better results with respect to F- measure, recall, accuracy and entropy. S. Sumathi and G. Hannah Grace [12], worked on the BBC Sports dataset using k-medoids clustering algorithm for text document clustering. The newly proposed measure was compared with Manhattan, Euclidean distance measures and was experimentally proved to work well for sparse matrix. However, this proposed measure can also be elongated to work for weighted terms and to be compatible with other clustering algorithms.

A. Existing Similarity functions used in WIR

This Section discusses the variety of similarity measures used for web documents retrieval. Although, the three most popular and quite commonly used similarity measurement functions are Cosine, Dice and Jaccord. A brief description of them is given as follows:

Cosine similarity: Cosine similarity measure has a great impact on the relevancy of the information to be provided to the user [3]. Equation (1) gives the formula for Cosine Similarity.

$$\text{Cosine (q, d}_j) = \frac{\sum_{i=1}^{m} w(t_i,q) \times w(t_i,d_j)}{\sqrt{\sum_{i=1}^{m} w(t_i,q)^2} \times \sqrt{\sum_{i=1}^{m} w(t_i,d_j)^2}} \qquad (1)$$

where, $w(t_i,q)$ and $w(t_i,d_j)$ are the weights of the term t_i in query q and document dj respectively.

Jaccard similarity: This would gauge the relevance between documents [13] and is as shown in the equation (2).

$$\text{Jaccord (q, d}_j) = \frac{\sum_{i=1}^{m} w(t_i,q) \times w(t_i,d_j)}{\sum_{i=1}^{m} w(t_i,q)^2 + \sum_{i=1}^{m} w(t_i,d_j)^2 - \sum_{i=1}^{m} w(t_i,q) \times w(t_i,d_j)} \qquad (2)$$

where, $w(t_i,q)$ and $w(t_i,d_j)$ are the weights of the term t_i in query q and document dj respectively.

Dice similarity: In literature, Dice similarity measure is proved to be effective in retrieving the relevant documents in the context of WIR [8]. The following equation (3) shows the Dice Coefficient measure:

$$\text{Dice (q, d}_j) = \frac{2\sum_{i=1}^{m} w(t_i,q) \times w(t_i,d_j)}{\sum_{i=1}^{m} w(t_i,q)^2 \times \sum_{i=1}^{m} w(t_i,d_j)^2} \qquad (3)$$

where, $w(t_i,q)$ and $w(t_i,d_j)$ are the weights of the term t_i in query q and document dj respectively.

III. Proposed Similarity Measure SMDR

Based on the properties of a similarity measure, a new similarity measure named as Similarity Measure for Documents Retrieval (SMDR) has been proposed and is defined in equation (4). While a WIR system accepted the query and has to process it against a documents repository with t distinct terms of the entire collection, firstly, the WIR system formulates a vector D $(d_{i1}, d_{i2}, \ldots, d_{it})$ of size t for each document. Similarly, the system has to construct a vector Q $(w_{q1}, w_{q2}, \ldots, w_{qt})$ for the query terms. A similarity measure SMDR (Q, D_i) for the query Q and a document D_i is calculated by using equation (4).

$$SMDR(Q, D_i) = \frac{\sum_{j=1}^{t} (w_{qj} d_{ij})}{\sqrt{\sum_{j=1}^{t} (w_{qj} d_{ij})^2} + \sqrt{\sum_{j=1}^{t} (w_{qj})^2} \times \sqrt{\sum_{j=1}^{t} (d_{ij})^2}} \qquad (4)$$

where, $\sum_{j=1}^{t}(w_{qj})$ and $\sum_{j=1}^{t}(d_{ij})$ are the weights of the terms in query Q and document D_i respectively.

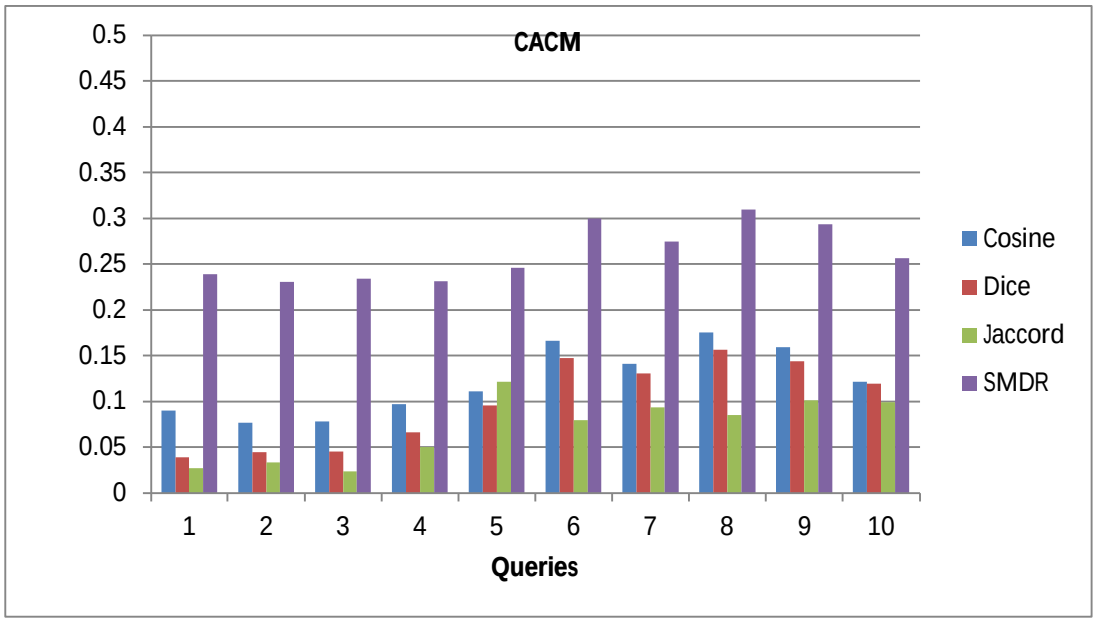
When the query Q is given by the user into the WIR system, the query Q is term weighted and the corresponding vector model is generated. Similarly, the terms in the documents are indexed and their corresponding document vectors are also formulated. For term weighting, $tf_{ij} \times idf_j$ scheme is applied. That is to compute the value of the j^{th} entry in the vector corresponding to document i, where, tf_{ij} = number of occurrences of term t_j in document Di, known as the *term frequency*. And $idf_j = log(\frac{d}{df_j})$, where d is the total number of documents and df_j is the number of documents which contain t_j (document frequency), known as *inverse document frequency*.

Subsequently, the proposed SMDR measure is applied to compute the similarity between the query and each document in collection. Similarity values are computed for every query and document vectors and *SMDR(Q, Di)* has to be in the range *[0,1]*.If SMDR value for a document is more towards 1, then it is considered to be more relevant to the user issued query and if more towards 0 then it is considered to be less relevant. Based on the values obtained by documents, the ranks will be assigned. The set of documents which obtain highest ranks will be collectively retrieved and given by WIRS to the user as results. So the results would be most relevant documents which match the user query and satisfy the user need.

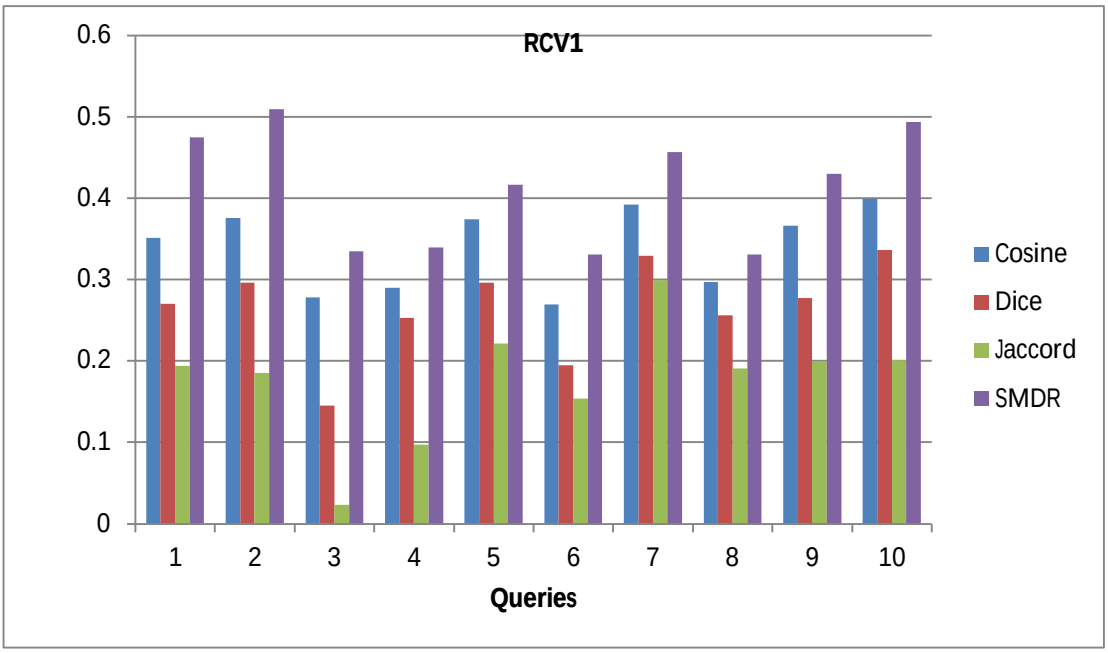
IV. Experimental Results and Discussion

The proposed SMDR similarity measure is used to compute query-documents similarity in the WIR system. The effectiveness of SMDR is tested over 50 different queries of different term frequency. The input datasets CACM and RCV1 are initially pre-processed and the process of indexing is accomplished with the LUCENE indexing

procedure. The query and documents are represented using vector space model and all the terms are weighted. The standard PSO algorithm is used to optimise the WIR process. More details on how to develop PSO based WIR system can be referred from the articles [14, 15]. SMDR is used as a fitness function in the standard PSO algorithm.



(a)

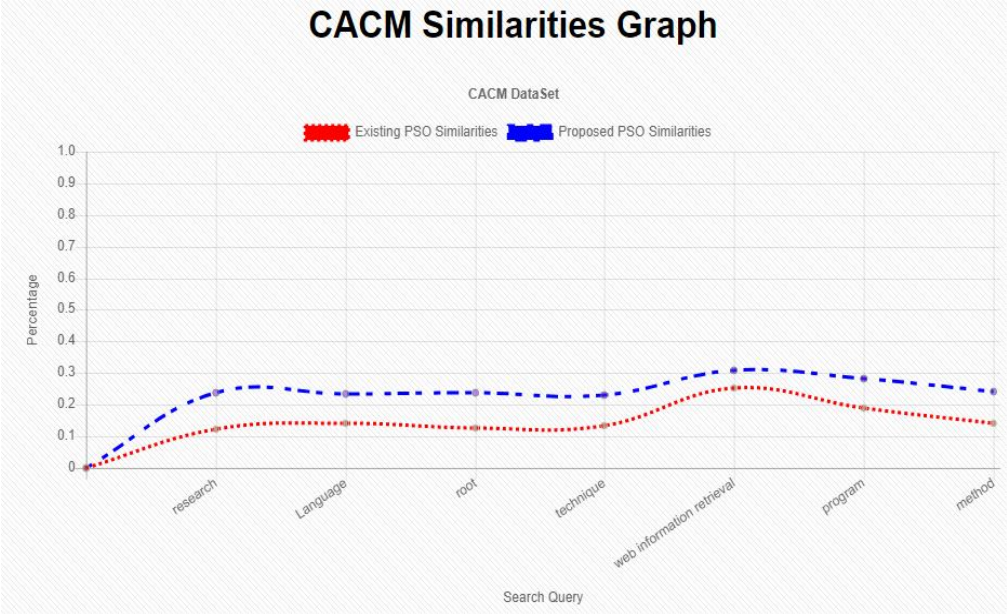


(b)

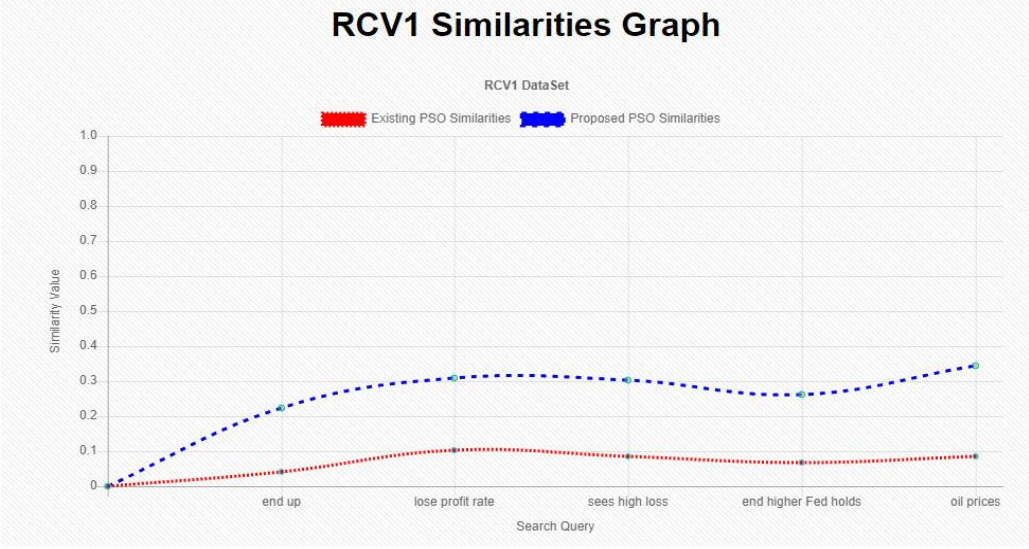
Figure 1. Performance of the similarity measures on (a) CACM and (b) RCV1 collections

The performance of SMDR is compared with the existing similarity measures Cosine, Jaccard, and Dice. The common query set is considered to measure the query-document similarity. The corresponding performance graph for both existing and proposed similarity measures is shown in Fig. 1. From the analysis it is very evident that the proposed SMDR is more effective than the Cosine, Dice and Jaccord similarity measures as the precision value of the SMDR is comparatively better than the existing measures. The performance of the fitness functions

used in WIR system are compared for various queries. Fig. 2 shows the plot of similarity values directly drawn by the system for CACM in (a) and RCV1collection in (b).



(a)



(b)

Figure 2. Plot of Similarity values for (a) CACM collection (b) RCV1 collection

V. CONCLUSIONS

The formulation of novel and effective similarity measure SMDR is presented. The SMDR is implemented for estimating the similarity values between the query and documents. It is used as fitness function in standard PSO algorithm. The obtained similarity values showed the better performance of the novel similarity measure over the existing measures. Thus SMDR helps in retrieving the most relevant documents to the user hence contributing to the effectiveness of the WIR system. Hence, this research work contributes to the optimization of WIR process. In future, the proposed SMDR can be further enhanced for its effectiveness with less computations to save the resources of WIR system.

REFERENCES

[1] Anna Huang, "Similarity Measures for Text Document Clustering", *Proceedings of the New Zealand Computer Science Research Student Conference*, pp. 49-56, 2008.

[2] Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", *IEEE Transactions on knowledge and data engineering*, VOL. 26, NO. 7, pp. 1575-1589, 2014. DOI: 10.1109/TKDE.2013.19.

[3] Xiaojun Wan, "A novel document similarity measure based on earth mover's distance", *Information Sciences*, vol. 177, pp. 3718–3730, 2017. DOI:10.1016/j.ins.2007.02.045.

[4] Naresh Kumar Nagwani, "A Comment on a Similarity Measure for Text Classification and Clustering", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 9, pp 2589-2590, 2015. DOI: 10.1109/TKDE.2015.2451616.

[5] Komal Maher and Madhuri S. Joshi, "Effectiveness of Different Similarity Measures for Text Classification and Clustering", *International Journal of Computer Science and Information Technologies*, Vol. 7 (4), pp. 1715-1720, 2016.

[6] Mubarak Himmat, Naomie Salim, Mohammed Mumtaz Al-Dabbagh, Faisal Saeed and Ali Ahmed, "Adapting Document Similarity Measures for Ligand-Based Virtual Screening", *Molecules*, 2016, DOI:10.3390/molecules21040476.

[7] D Gunawan, C A Sembiring and M A Budiman, "The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents", *Journal of Physics: Conference Series: 978 012120*, 2018, DOI :10.1088/1742-6596/978/1/012120.

[8] Maedeh Afzali and Suresh Kumar, " An Extensive Study of Similarity and Dissimilarity Measures Used for Text Document Clustering using K-means Algorithm", *International Journal of Information Technology and Computer Science*, Vol. 9, pp 64-73, 2018. DOI: 10.5815/ijitcs.2018.09.08.

[9] Naveen Kumar , Sanjay Kumar Yadav and Divakar Singh Yadav , "Similarity Measure Approaches Applied in Text Document Clustering for Information Retrieval", *Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pp. 88-92,2020.

[10] Dan Goldwasser and Xiao Wang , "Cross-Lingual Document Retrieval with Smooth Learning", *Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online)*, December 8-13, pp. 3616–3629, 2020.

[11] R. Lakshmi and S. Baskar, "Efficient text document clustering with new similarity Measures", *Int. J. Business Intelligence and Data Mining*, Vol. 18, No. 1, 2021.

[12] S. Sumathi and G. Hannah Grace, "A Similarity Measure for Text Document Using Term Cardinality", *IOP Conf. Series: Materials Science and Engineering*, 2021. DOI:10.1088/1757-899X/1012/1/012059.

[13] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu, "Using of Jaccard Coefficient for Keywords Similarity", *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, Vol I, pp 13 – 15, 2013.

[14] Kaushik, Nainika & Bhatia, Manjot, "Information Retrieval from Search Engine Using Particle Swarm Optimization", 2020. DOI: 10.1007/978-981-15-0222-4_11.

[15] Surya, S and Sumitra, P, "An Innovative Information Retrieval Model Implementing Particle Swarm Optimization Technique", Journal of Computational and Theoretical Nanoscience, Vol. 17, No. 12, pp. 5613-5617, 2020. DOI: https://doi.org/10.1166/jctn.2020.9460.