

Andrea Tagarelli  
Hanghang Tong (Eds.)

LNCS 11917

# Computational Data and Social Networks

8th International Conference, CSoNet 2019  
Ho Chi Minh City, Vietnam, November 18–20, 2019  
Proceedings

Andrea Tagarelli · Hanghang Tong (Eds.)

# Computational Data and Social Networks

8th International Conference, CSoNet 2019  
Ho Chi Minh City, Vietnam, November 18–20, 2019  
Proceedings

*Editors*

Andrea Tagarelli  
University of Calabria  
Rende, Italy

Hanghang Tong  
University of Illinois  
at Urbana-Champaign  
Urbana, IL, USA

ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-030-34979-0

ISBN 978-3-030-34980-6 (eBook)

<https://doi.org/10.1007/978-3-030-34980-6>

LNCS Sublibrary: SL1 – Theoretical Computer Science and General Issues

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Contents

## Combinatorial Optimization and Learning

A Probabilistic Divide and Conquer Algorithm for the Minimum Tollbooth Problem . . . . .	3
<i>Julian Nickerl</i>	
Distributed Core Decomposition in Probabilistic Graphs . . . . .	16
<i>Qi Luo, Dongxiao Yu, Feng Li, Zhenhao Dou, Zhipeng Cai, Jiguo Yu, and Xiuzhen Cheng</i>	
Sampled Fictitious Play on Networks. . . . .	33
<i>Alexander Nikolaev, Alexander Semenov, and Eduardo L. Pasiliao</i>	
Outlier Detection Forest for Large-Scale Categorical Data Sets . . . . .	45
<i>Zhipeng Sun, Hongwei Du, Qiang Ye, Chuang Liu, Patricia Lilian Kibenge, Hui Huang, and Yuying Li</i>	
Neural Networks with Multidimensional Cross-Entropy Loss Functions . . . . .	57
<i>Alexander Semenov, Vladimir Boginski, and Eduardo L. Pasiliao</i>	
A RVND+ILS Metaheuristic to Solve the Delivery Man Problem with Time Windows . . . . .	63
<i>Ha-Bang Ban</i>	
Star2vec: From Subspace Embedding to Whole-Space Embedding for Intelligent Recommendation System ( <i>Extended Abstract</i> ). . . . .	70
<i>Quang Nguyen, Vu Nguyen, Doi Tran, Trung Mai, and Tho Quan</i>	
Approximation Algorithm for the Squared Metric Soft Capacitated Facility Location Problem ( <i>Extended Abstract</i> ) . . . . .	72
<i>Lu Han, Dachuan Xu, Yicheng Xu, and Dongmei Zhang</i>	
An Efficient Algorithm for the $k$ -Dominating Set Problem on Very Large-Scale Networks ( <i>Extended Abstract</i> ) . . . . .	74
<i>Minh Hai Nguyen, Minh Hoàng Hà, Dinh Thai Hoang, Diep N. Nguyen, Eryk Dutkiewicz, and The Trung Tran</i>	
Strategy-Proof Cost-Sharing Mechanism Design for a Generalized Cover-Sets Problem ( <i>Extended Abstract</i> ) . . . . .	77
<i>Longkun Guo, Chenchen Wu, Huahui Yu, and Xiaoyan Zhang</i>	

## **Influence Modeling, Propagation, and Maximization**

Hybrid Centrality Measures for Service Coverage Problem. . . . .	81
<i>Anuj Singh, Rishi Ranjan Singh, and S. R. S. Iyengar</i>	
Hop-Based Sketch for Large-Scale Influence Analysis . . . . .	95
<i>Phuc D. Thai and Thang N. Dinh</i>	
Profit Maximization Under Group Influence Model in Social Networks . . . . .	108
<i>Jianming Zhu, Smita Ghosh, Weili Wu, and Chuangen Gao</i>	
Importance Sample-Based Approximation Algorithm for Cost-Aware Targeted Viral Marketing . . . . .	120
<i>Canh V. Pham, Hieu V. Duong, and My T. Thai</i>	
Location-Based Competitive Influence Maximization in Social Networks . . . . .	133
<i>Manh M. Vu and Huan X. Hoang</i>	
Cascade of Edge Activation in Networks . . . . .	141
<i>Gabriel Lopez Zenarosa, Alexander Veremyev, and Eduardo L. Pasiliao</i>	
Reinforcement Learning in Information Cascades Based on Dynamic User Behavior . . . . .	148
<i>Mengnan Chen, Qipeng P. Zheng, Vladimir Boginski, and Eduardo L. Pasiliao</i>	
Predicting New Adopters via Socially-Aware Neural Graph Collaborative Filtering . . . . .	155
<i>Yu-Che Tsai, Muzhi Guan, Cheng-Te Li, Meeyoung Cha, Yong Li, and Yue Wang</i>	
Who Watches What: Forecasting Viewership for the Top 100 TV Networks . . . . .	163
<i>Denis Khryashchev, Alexandru Papiu, Jiamin Xuan, Olivia Dinica, Kyle Hubert, and Huy Vo</i>	

## **NLP and Affective Computing**

PhonSenticNet: A Cognitive Approach to Microtext Normalization for Concept-Level Sentiment Analysis . . . . .	177
<i>Ranjan Satapathy, Aalind Singh, and Erik Cambria</i>	
One-Document Training for Vietnamese Sentiment Analysis . . . . .	189
<i>Dang-Khoa Nguyen-Nhat and Huu-Thanh Duong</i>	

Assessing the Readability of Vietnamese Texts Through Comparison (Extended Abstract) . . . . .	201
<i>An-Vinh Luong and Phuoc Tran</i>	

Learning Representations for Vietnamese Sentence Classification ( <i>Extended Abstract</i> ) . . . . .	203
<i>Phuc H. Duong and Hien T. Nguyen</i>	
Exploring Machine Translation on the Chinese-Vietnamese Language Pair ( <i>Extended Abstract</i> ) . . . . .	205
<i>Huu-Anh Tran, Phuoc Tran, Phuong-Thuy Dao, and Thi-Mien Pham</i>	
<b>Computational Methods for Social Good</b>	
Towards an Aspect-Based Ranking Model for Clinical Trial Search . . . . .	209
<i>Soumyadeep Roy, Koustav Rudra, Nikhil Agrawal, Shamik Sural, and Niloy Ganguly</i>	
Scalable Self-taught Deep-Embedded Learning Framework for Drug Abuse Spatial Behaviors Detection . . . . .	223
<i>Wuji Liu, Xinyue Ye, Hai Phan, and Han Hu</i>	
Limiting the Neighborhood: De-Small-World Network for Outbreak Prevention . . . . .	229
<i>Ruoming Jin, Yelong Sheng, Lin Liu, Xue-Wen Chen, and NhatHai Phan</i>	
Online Community Conflict Decomposition with Pseudo Spatial Permutation . . . . .	246
<i>Yunmo Chen and Xinyue Ye</i>	
Attribute-Enhanced De-anonymization of Online Social Networks . . . . .	256
<i>Cheng Zhang, Shang Wu, Honglu Jiang, Yawei Wang, Jiguo Yu, and Xiuzhen Cheng</i>	
Subgraph-Based Adversarial Examples Against Graph-Based IoT Malware Detection Systems . . . . .	268
<i>Ahmed Abusnaina, Hisham Alasmay, Mohammed Abuhamad, Saeed Salem, DaeHun Nyang, and Aziz Mohaisen</i>	
Incorporating Content Beyond Text: A High Reliable Twitter-Based Disaster Information System . . . . .	282
<i>Qixuan Hou and Meng Han</i>	
Reduced-Bias Co-trained Ensembles for Weakly Supervised Cyberbullying Detection . . . . .	293
<i>Elaheh Raisi and Bert Huang</i>	
A Novel Privacy-Preserving Socio-technical Platform for Detecting Cyber Abuse ( <i>Extended Abstract</i> ) . . . . .	307
<i>Sriram Chellappan and Nathan Fisk</i>	

Projection-Based Coverage Algorithms in 3D Camera Sensor Networks  
for Indoor Objective Tracking. . . . . 309  
*Yi Hong, Yongcai Wang, Yuqing Zhu, Deying Li, Mengjie Chang,  
and Zhibo Chen*

**User Profiling and Behavior Modeling**

Enhancing Collaborative Filtering with Multi-label Classification . . . . . 323  
*Yang Zhou, Ling Liu, Qi Zhang, Kisung Lee, and Balaji Palanisamy*

Link Prediction on Dynamic Heterogeneous Information Networks . . . . . 339  
*Chao Kong, Hao Li, Liping Zhang, Haibei Zhu, and Tao Liu*

Information Network Cascading and Network Re-construction  
with Bounded Rational User Behaviors . . . . . 351  
*Guanxiang Yun, Qipeng P. Zheng, Vladimir Boginski,  
and Eduardo L. Pasiliao*

Gender Prediction Through Synthetic Resampling of User Profiles  
Using SeqGANs . . . . . 363  
*Munira Syed, Jermaine Marshall, Aastha Nigam, and Nitesh V. Chawla*

**Author Index** . . . . . 371



# Assessing the Readability of Vietnamese Texts Through Comparison (*Extended Abstract*)

An-Vinh Luong<sup>1</sup>  and Phuoc Tran<sup>2</sup>  

<sup>1</sup> University of Science, VNU-HCM, Ho Chi Minh City, Vietnam  
anvinhluong@gmail.com

<sup>2</sup> NPL-KD Lab, Faculty of Information Technology, Ton Duc Thang University,  
Ho Chi Minh City, Vietnam  
tranthanhphuoc@tdtu.edu.vn

**Abstract.** Text readability is a measure of how easy/difficult a text is to read. Studies on readability have been noticed for a long time in English and other popular languages. In Vietnamese, studies on the text readability are still quite limited, mainly because of the lack of corpora with capacity large enough to carry out the examination. In this paper, we present a method for assessing the readability of Vietnamese texts that does not need too much cost to develop a training corpus, based on comparing the correlation difficulty between texts but still achieving positive results.

**Keywords:** Text readability · Text comparison · Vietnamese language

Text readability—as the definition of Bailin and Grafstein [1]—is a measure of how easy/difficult a text is to read. Studies on readability have been noticed for a long time in English and other popular languages, such as the work of Chen and Meurers [2], Jiang et al. [3]. In Vietnamese, research on the readability of texts is still quite limited, the main reason is the lack of corpora that was graded according to the difficulty level (often with the participation of language experts and at a high cost). In this study, we present a method for assessing the Vietnamese texts readability, using a small number of standard texts (texts that was graded according to the difficulty level), through comparing the correlate readability of the texts.

We collected 100 texts in the field of literature: children's stories, short stories, novels. We randomly selected some texts and gave them to the language experts to read. Each document is given to 2 experts to assess the readability according to 3 levels: Very easy (texts for children at the Elementary School); Easy (texts for people at the Middle School); and Medium (texts for people at the High School). If the evaluation results of both experts are the same, the text will be selected as a standard text. We conduct the such choice and evaluation process until each level has enough 10 texts (the total of 3 levels is 30 documents). These 30 texts are standard texts which will be used to evaluate the remaining texts.

All documents collected are extracted features for training models: Average sentence length calculated on word and syllable; Average word length calculated on syllable; Ratio



of difficult words and Ratio of difficult syllables; Ratio of Sino-Vietnamese words; Ration of local words; Ratio of proper nouns and Ratio of distinct proper nouns. With standard texts, we construct vectors of text pairs by taking all 2-permutations, totally we have 870 text pairs. The feature vectors of each pair is created by vectors subtraction. For example, with 2 texts  $a$  and  $b$ ,  $v_a$  and  $v_b$  are 2 feature vectors of  $a$  and  $b$ ; the feature vector of text pairs  $(a, b)$ , denoted as  $v_{ab}$ ; calculated as  $v_{ab} = v_a - v_b$ . The vector  $v_{ab}$  are labeled as follows: if  $a$  is more difficult than  $b$  then  $v_{ab} = 1$ ; if  $b$  is more difficult than  $a$  then  $v_{ab} = -1$ ; otherwise  $v_{ab} = 0$ . All these vectors will be used for training a text comparative model using SVM. Using k-fold cross validation, we have checked the accuracy of the model achieved 74.94%.

For the remaining texts, we compare the readability with the standard texts: for each remaining text  $x$ , we create all feature vectors of text pairs  $(x, a)$ , with  $a$  is a text in standard texts, by  $v_x - v_a$ . These feature vectors will, in turn, be evaluated by the trained comparison model. Finally, we classify the remaining texts through the comparison results: A text  $x$  will be classified as: (1) Medium if it is more difficult than minimum  $k\%$  Easy and Very easy texts and it is easier than maximum  $t\%$  Medium texts; (2) Very easy if it is easier than minimum  $k\%$  Easy and Medium texts and it is more difficult than maximum  $t\%$  Very easy texts; (3) Easy if it is more difficult than minimum  $k\%$  Very easy texts and easier than minimum  $k\%$  Medium texts and it is more difficult than maximum  $t\%$  Easy texts and it is easier than maximum  $t\%$  Easy texts; (4) otherwise, it will be labeled as unclassifiable.  $k$  and  $t$  are classification thresholds, obtained through experiments: the higher the value of  $k$  and  $t$ , the larger the number of unclassified documents, but the precision of the classification results is higher. In this study, we used  $k = 60\%$  and  $t = 50\%$ .

Finally, we selected 30 newly classified documents for re-testing by experts. Each text will be read by an expert to decide if the classification results are proper. The classification results are presented in Table 1. The Precision, Recall and F1 score are 78.13%, 35.71% and 49.02% accordingly.

**Table 1.** Classification results.

No. of texts	Very easy	Easy	Medium	Unclassifiable
	18	9	5	38

For conclusion, with only a few pre-classified documents, we can build a model to assess the correlate readability between Vietnamese texts with an acceptable accuracy. Based on that assessment model, we have classified the texts by readability with an accuracy of over 78%.

# References

1. Bailin, A., Grafstein, A.: Readability: Text and Context. Palgrave Macmillan, UK (2016)
2. Chen, X., Meurers, D.: Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. J. Res. Reading (2018)
3. Jiang, Z., et al.: Enriching word embeddings with domain knowledge for readability assessment. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA (2018)