Minh-Le Nguyen
School of Information Science,
Japan Advance Institute of Science and Technology.
1-1 Asahidai, Nomi-city, Ishikawa, 923-1292, JAPAN.
☎ *+81 (0761) 51 1221*
FAX *+81 (0761) 51 1149*
✉ *nguyenml@jaist.ac.jp*

Prof. H. Fujita and Prof. J. Lu April 17, 2017
Editors in Chief of Knowledge-Based Systems.

Dear Prof. Fujita and Prof. Lu,

I am writing to submit our manuscript entitled: Social Context Summarization using User-generated Content and Third-party Sources[1], which is an improved and extended version of the paper: Summarizing Web Documents using Sequence Labeling with User-generated Content and Third-party Sources, presented at 22^{nd} *International Conference on Natural Language & Information Systems* (NLDB) 2017, for the consideration of publication in Knowledge-Based Systems.

Traditional summarization methods only use inherent information of a Web document to generate summarization. They, however, usually ignores two beneficial aspects of a Web document: (i) the user-generated content (comments or tweets) and (ii) content tent reflection from variety of relevant articles (third-party sources) retrieved from a search engine. Such kind of information can enrich the summarization towards a special event. This paper proposes a framework named *SoSVMRankSum* to take the advantage of user-generated content and third-party sources to extract important sentences and comments (or tweets) as the summarization. In order to to that, the summarization was formulated as a learning to rank task (L2R), in which sentences, user-generated content, and third-party sources were modeled in a unified framework, which exploits the support from additional information in a mutual reinforcement fashion. To model the relation of a sentence-comment (or tweet) pair, 13 new features are presented. Top *m* ranked sentences and social messages are selected as the summarization using score-based or voting-based methods.

Experimental results on three datasets in two languages show that our model significantly outperforms baselines and obtains competitive results with state-of-the-art methods. We believe our findings are likely to be of great interests to information retrieval and data mining scientists who read your journal.

Compared to the original paper, this manuscript makes eight new and significant improvements.
- It investigates a literature review, which makes a story of text as well as social context summarization.
- It annotates and releases a dataset which contains news articles and their tweets collected from Twitter. The dataset is annotated by the human with Cohen's Kappa is 0.617.
- It presents the summarization as a L2R instead of sequence labeling problem and utilizes Ranking SVM instead of Conditional Random Fields for sentence extraction.
- It also adds a voting-based method based on L2R methods for sentence selection.
- It validates our methods along with deep analyses on three datasets in English and Vietnamese.
- It clearly describes features which are not sufficiently mentioned in the original paper.
- It also compares our model to simple Ranking SVM for social context summarization. Experimental results indicate that our model sufficiently outperforms this method.
- It also compares our approach to state-of-the-art methods in social context summarization. Experimental results illustrate that SoSVMRankSum obtains very competitive results.

All authors approved the manuscript and this submission.
Thank you very much for receiving our manuscript and considering it for review. We appreciate your time and look forward to your response.

Sincerely,

Minh-Le Nguyen

[1]All the necessary documents can be accessed at: https://github.com/nguyenlab/KBS-Submission