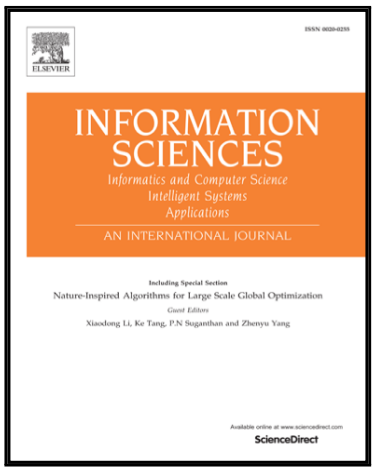

Controversy detection in Wikipedia using semantic dissimilarity

M. Zeeshan Jhandir , Ali Tenvir , Byung-Won On , Ingyu Lee , Gyu Sang Choi

PII: S0020-0255(17)30882-4
DOI: 10.1016/j.ins.2017.08.037
Reference: INS 13041

To appear in: *Information Sciences*

Received date: 15 June 2016
Revised date: 6 August 2017
Accepted date: 9 August 2017

Please cite this article as: M. Zeeshan Jhandir , Ali Tenvir , Byung-Won On , Ingyu Lee , Gyu Sang Choi , Controversy detection in Wikipedia using semantic dissimilarity, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.08.037

Controversy detection in Wikipedia using semantic dissimilarity

M. Zeeshan Jhandir[a], Ali Tenvir[a], Byung-Won On[b1], Ingyu Lee[c], Gyu Sang Choi[a2]

[a]Yeungnam University, 280-Dehakeo-ro, Gyeongsan-si and 38541, Rep. South Korea
email: {zeeshanjhandir, tenvirali, castchoi }@ ynu.ac.kr}
[b]Kunsan National University, Gunsan-si, Rep. of South Korea
email: on.byung.won@gmail.com
[c]Sorrel College of Business, Troy University, Troy, AL 36082, USA
email: inlee@troy.edu

Abstract

The advent of search engines and wikis has made access to information easy and almost free. Wikipedia is the efficacious outcome of an enormous collaboration, and its peer review-like methods of creation, maintenance, and evolution of contents, ensure high quality and reliability. However, the "anyone-can-edit" policy of Wikipedia has created many problems such as trolling, vandalism, controversies, and doubts about the content and reliability of the information provided due to non-expert involvement. People have tried to identify and rank controversies in Wikipedia articles through various techniques that use quantitative data, ignoring the semantic significance of conflicts among authors. In this paper, we have addressed the problem of identifying controversy using natural language processing techniques for the first time. The proposed method spots the impact on existing meanings of the text due to new editing processes along with their relationship to the topic of the article. The experimental results for precision (0.901), recall (0.901), accuracy (0.908), and F-measure (0.901) demonstrate the effectiveness of the proposed method. The technique is deemed useful for automatic identification of conflicts newly introduced into existing article contents, and could prove helpful in making decisions for inclusion or exclusion of controversies under the same topic.

Keywords: Wikipedia; controversy; semantic dissimilarity; sentence similarity; natural language processing; edit similarity

[1] Corresponding author. Tel.: +82-63-469-4612 ; fax: +82-63-469-7423
E-mail address: on.byung.won@gmail.com
[2] Corresponding author. Tel.: +82-53-810-3091; fax: +82-53-810-4742.
E-mail address: castchoi@ynu.ac.kr
Submitted to Information Sciences for initial review