EGU21-15361
https://doi.org/10.5194/egusphere-egu21-15361
EGU General Assembly 2021

Developing Efficient Web Crawler for Effective Disaster Management

Lakshmi S Gopal, Rekha Prabha, Divya Pullarkatt, and Maneesha Vinodini Ramesh
Amrita Center for Wireless Networks and Applications, Amrita Vishwa Vidyapeetham, Amritapuri , India

The exponential escalation of disaster loss in our country has led to the awareness that disaster risks are presumably increasing. In the past few years, numerous hazards have been reported in India which has caused severe casualties, infrastructural, agricultural and economic damages. Over the years, researchers have scrutinized social media data for disaster management as it has the advantage of being available in real time and stays relevant in hazard response. But, the authenticity of social media data has been questioned particularly in a disaster management scenario where false information cannot be afforded. Collection of credible disaster statistics during or after a hazard occurrence is a demanding task. Web documents such as a news report are credible when compared to social media data and hence, the proposed work aims in developing a web crawler which is a software that's capable of indexing legitimate news websites from the world wide web which contains news articles related to hazards. The articles are extracted by incorporating the technique of data scraping which includes the use of a developed hazard ontology. The ontology contains hazard relevant keywords at multiple granularities. The developed crawler is able to prioritise websites based on its contents which makes the data collection more accurate. The collected data is analyzed and structured as it may assist in administering hazard emergencies during a hazard, preparedness before a hazard occurrence and other post disaster activities efficiently. The proposed work also focuses on local media as it may provide news reports from regional locations which might not be reported in the mainstream media. News articles are written in natural languages and hence structuring them into a statistical form involves natural language processing methodologies. The proposed work mainly focuses on semantic information extraction from news articles to extract statistical data related to the hazard, its impacts and loss. News illustrations often include less newsworthy content such as advertisements and past studies of the hazard location. Hence, a supervised learning based text classification is performed to classify newsworthy content from the articles and approximately 70% accuracy has been achieved.