UNIVERSITY OF WESTMINSTER

INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In collaboration with

UNIVERSITY OF WESTMINSTER, UK.

Forum Off-topic Post Detection Using Natural Language Processing

A dissertation by

Mr. Ramitha Ishan Abeyratne

Supervised by

Mr. Cassim Farook

Submitted in partial fulfillment of the requirements for the

BEng (Hons) Software Engineering Degree

Department of Computing

May 2018

Abstract

A constant need to seek information is found among people who live in this fast moving world. One prominent way of meeting this demand is by using forums. People use forums to create topics, post questions, search for answers, discuss and post replies to threads. Due to the extreme growth of internet users, drastic increase of forum users were observed. A number of issues were identified when managing forums. One issue is managing off-topic posts. It is one of the most complex tasks of online forum management. Off-topic posts break the flow of knowledge stored within threads. They significantly reduce the readability of forums. Detection of off-topic posts are currently done manually. It is a very tedious and nearly impossible task when the number of threads or posts increases.

This research illustrates an automated web-based solution which can be used to detect off-topic posts in online forums. Natural Language Processing is used to differentiate off-topic content from relevant content. A modified algorithm is proposed for evaluating similarity. WordNet "path" vector cosine angle semantic analysis and Dice co-efficient overlap level lexical analysis techniques are used to generate two distinct scores for each post. The final dissimilarity score is calculated by dynamically weighting the two individual scores based on the average thread word count using a regression model. The modified algorithm was compared against TF-IDF and Dice. A forum dataset obtained from Stack Exchange was used as the input. Results show that a phenomenal increase in accuracy, as high as 73.34%, was obtained.

Subject Descriptors

1.2: Artificial Intelligence

    1.2.7: Natural Language Processing

H.3 Information Storage and Retrieval

    H.3.3 Information Search and Retrieval

Keywords

Lexical similarity, Semantic similarity, WordNet path, Binomial Log-ratio Test, Vector, Cosine angle, Dice co-efficient, Corpus based, Knowledge based, Regression