A Hybrid Method for Dissimilarity Analysis between Short Text Documents

Ramitha Abeyratne#1, Cassim Farook#2

#Department of Computing, Informatics Institute of Technology
57, Ramakrishna road, Colombo 06, Sri Lanka
1ramithaabeyratne@rocketmail.com
2cassim.f@iit.ac.lk

Abstract— **Similarity analysis is an extremely popular aspect of Natural Language Processing (NLP). Most of the existing works focuses on analysing content of large documents. There are comparatively a smaller number of researches available which focuses on analysing similarity between short unstructured documents. This work proposes a hybrid approach which uses WordNet Path vector cosine angle analysis and Dice co-efficient overlap level analysis to determine the similarity levels of short texts. A regression model is used to dynamically weight and combine the calculated two individual scores into a single score. This hybrid approach was found to have significantly higher accuracy rates against Term Frequency–Inverse Document Frequency (TF-IDF) and Dice co-efficient techniques.**

Keywords— **Similarity Analysis, WordNet Path Vector Cosine Angle, Dice Co-efficient, Linear Regression.**

I. Introduction

We live in a modern world where information plays a major role in defining the lifestyles of people. Therefore, a constant need for seeking information is commonly found among them. People use various different methods and techniques to gather the required information. One popular way is by using forums, which are also known as online discussion boards. Users of these tools usually create topics, ask questions, discuss related issues and post opinions to threads. Forums consist of threads which holds related posts based on the topics. Therefore, knowledge is stored within a sequence of posts [1].

Due to the rapid growth of the internet user base, a severe increase of forum users were identified. This led to a number of issues where the liberties granted to forum users were misused. Detecting duplicate threads, off-topic posts and fake user profiles are three of the most commonly occurred concerns. Out of them, this research targets on managing off-topic posts in forums. It is identified as one of the most complex tasks of forum management [2].

Off-topic posts are posts which break the flow of knowledge contained within threads. They contain content which is irrelevant to the containing thread. Ones such example would be posting a “bank loan rates” post inside a “vehicle fuel efficiency” related thread. Moreover, concurrently discussing two or more topics within a single thread can also considerably decrease the readability of forums [3].

Detection of off-topic posts are currently done manually by iterating over and analysing every individual thread. The interactive nature of forums makes it very difficult to assess relevancy every time a new post is added to the discussion board. Moreover, manual detection of off-topic posts becomes more complex when the number of threads or number of posts increases. It is also common to have access control which limits the off-topic marking privilege to users with escalated permission [2]. Therefore, an alternative is required for manual off-topic post detection.

A good amount of researches are available for similarity analysis between documents. However, most of them fail to perform steadily in the context of short, unstructured documents such as forum posts [3]. The fragmented nature of posts makes it even challenging to detect similarity using traditional NLP methods [3].

A novel approach is presented in this research work to accurately detect a target post’s relevancy level based on context of the containing thread. Four main steps are included in this approach. Initially, keywords which represent threads and posts are extracted based on their significance level. Next, vectors are created using the WordNet Path similarity scores obtained between word synsets [4]. Cosine angle is calculated between the vector representations of the target post against the thread and the target post against the followed post. Dice co-efficient [5] analysis is performed to capture the overlap level between documents. The dissimilarity score of the target post is calculated by combining and normalizing the generated two individual similarity scores using a regression model based on the amount of context available for detection. Finally, posts are evaluated based on the dissimilarity score.

The succeeding sections of this paper is demonstrated as below. Section II contains a brief review of related work in this domain. Section III contains an outline of the proposed approach. Section IV comprises of details regarding accuracy and performance testing of the proposed method while section V benchmarks it with two existing methods. Section VI concludes the paper by discussing the verdicts.

II. Short Text Similarity Analysis Techniques

Several techniques are available for similarity analysis of short documents. They belong to lexical-based and semantic-based evaluation. This section briefly describes the two similarity analysis methods.

A. Lexical Similarity

Lexical similarity analysis method evaluates the surface level similarity between documents [6]. Sequences of text

978-1-5386-7352-2/18/$31.00 ©2018 IEEE