Extracting Knowledge from Noise Information

Tawanda Chowuraya, Sivaramalingam Kirushanth and Boniface Kabaso
Cape Peninsula University of Technology, Cape Town, South Africa
tawandachowuraya@yahoo.co.uk
sivaramalingamk@cput.ac.za
kabasob@cput.ac.za

Abstract: Information is an important part of creating knowledge. We live at a time when so much information is created. Unfortunately, much of the information is redundant. There is a huge amount of online information in the form of news articles that cover similar stories. The number of articles is projected to grow. The growth makes it difficult for a person to process all that information to extract knowledge. This affects the quality of knowledge. There is need for a solution that can organize this similar information into specific themes. The solution is a branch of Artificial intelligence (AI) called machine learning (ML) using clustering algorithms. Clustering will group information that is similar into containers. When the information is clustered people can be presented with information about their interest, grouped together. The information in a group can even be summarized for better processing. One of the most widely used and studied clustering algorithm is K-Means. K-Means is chosen because of its simplicity and easiness to implement. However, many variations of K-Means have been produced. This makes it difficult to pick a variation of the k-Means to use for the clustering problem. This paper presents the systematic literature review conducted with the aim of finding the application of K-Means and other clustering algorithms using the hypothesis. Studies using clustering algorithms in different contexts, the techniques used, and a summary of the outcome is discussed. The result of the systematic literature review is presented in a tabular and textual format.

Keywords: Knowledge, Clustering, Machine Learning, Artificial Intelligence

1. Introduction

Knowledge is obtained from information. However, there is an overwhelming amount of information that is added to the internet daily. This is because barriers to online communication and cost of data capable devices have been reduced. Anyone with data capable device can participate in the creation and distribution of content online (Vishwakarma et al, 2017). Some publishers cover developing stories to an event (Moerchen et al, 2007). A huge amount of this information is similar. The amount of information is expected to grow yet the reading capacity of a person has not increased. This makes it difficult for a person to process all that information (Fiscus & Doddington, 2002). The result is gaps in knowledge.

2. Background

There is need for a solution that can organize this similar information into specific themes, to get continuous flow of knowledge. The solution is a branch of Artificial intelligence (AI) called machine learning (ML) using clustering algorithms. Clustering will group information that is similar into containers, this will reduce workload for readers. They can get stories pertaining to an event without reading huge number of stories. There is cost of reading a story, the cost is increased for reading a wrong story and the time spent on an irrelevant story is wasted (Fiscus and Doddington, 2002). The clustered information can further be summarized for better processing.

3. Literature review

Weiler et al (2016) have done an experimental analysis and survey on twitter event detection techniques. Their focus is on comparative study of the different techniques. For their study they use qualitative and quantitative comparison. Most techniques use qualitative evidence for motivating benefits of the technique.

Very few use quantitative evaluation. The other techniques also do not compare the results of their performance with competing approaches. The survey discusses techniques that are not discussed in other evaluation surveys. It also include enBlogue(ENB) which is a technique used to analyze evaluation measures.

The survey makes an evaluation of run-time and task based performance of the techniques. They evaluate each technique’s memory requirements, and measure the run-time performance on a separate set of hardware.