The advances of Stemming Algorithms in Text Analysis from 2013 to 2018

Liu, Yi-Yu[1], van Deventer, Philip[1], Kruger, Rendani[1]

[1]Department of Informatics, School of Information Technology, University of Pretoria, Pretoria, South Africa

E-mail: bruce.liu@up.ac.za
Telephone: (012) 420-3372

Stemming Algorithm is an activity within the pre-processing step of text analysis which plays a role in the text analysis results. Stemming algorithm falls part of text analysis which drives data mining in business information systems. 80% of existing organisational data that contributes to Big Data is in an unstructured format. One of the focus areas of concerns within the concept of *"Big Data"* is the complexity of processing the data and then being able to represent the results in such a way that it is easily understood. This has been a challenge taken up by researchers over time.

To determine the advances in stemming algorithm research, a systematic review was performed on articles on stemming algorithms published in journals from 2013–2018. Data collection was achieved through accessible scholarly databases followed by filtering the articles by year and topic. The remaining articles were processed through a set of methodological quality criteria. The final articles were put through a bi-gram text analysis process to answer the research questions.

The results concluded the research focus for stemming has started to decrease as it reaches the plateau of productivity. The results show an evident drop in the collected articles from 58 in 2017 to 19 in 2018. Results show that Information retrieval is still a common field of application for stemming algorithms. A major unexpected set of themes revolves around Artificial Intelligence, looking at the increase in interest thereof. Results show that a focus for stemming algorithms have shifted away from the development of stemming algorithms and moved towards the application thereof. There is also a high interest in social media as an application of stemming algorithms. Future research suggestions includes designing a stemming algorithm that would automatically and responsively adapt to the historical and morphological changes of language text.

Keywords: Stemming Algorithm, Text Analysis, Big Data, Systematic Review