

## **Application of Text Mining in Developing Standardized Descriptions of Taxa in Paleontology: A Framework**

Wen-Bin Yu, Department of Information Systems & Technology and Department of Computer Science, University of Missouri – Rolla, Rolla, MO 65409, yuwen@umr.edu, 573-341-7502

Francisca E. Oboh-Ikuenobe, Department of Geological Science and Engineering, University of Missouri – Rolla, Rolla, MO 65409, ikuenobe@umr.edu, 573-341-6946

Bih-Ru Lea, Department of Business Administration, University of Missouri – Rolla, Rolla, MO 65409, leabi@umr.edu, 573-341-6436

### **Abstract:**

Like other disciplines of science, the finding of new information and modification of existing knowledge advance paleontology. The process of discovery of new information generates large volumes of data that can be overwhelming if not properly stored and/or utilized. For example, the treatise on invertebrate macrofossils edited by Raymond in 1959 blazed the trail for similar works that came later. Many paleontological volumes provide information of fossil specimens that have been formally named. In palynology, problems can arise with palynomorph classifications and interpretations because of subjective nature due to human judgments and different levels of training. As a result, the same palynomorph can be interpreted or classified differently, resulting in junior synonyms and emended descriptions that can potentially confuse students and new researchers. It is important to provide a framework to compose a standardized description of each taxon utilizing diverse observations from various taxonomists.

The main objective of this study is to propose a framework that utilizes text mining techniques in developing a taxon description recommendation system. Text mining can apply intelligent methods/algorithms to extract or mine knowledge and meaningful data patterns from a large amount of unstructured texts or documents for decision-making. Therefore, it is expected that common characteristics and features from interpretations done by different scholars can be captured and used for clustering and description to minimize the issue of subjective human judgment.

The proposed framework will be illustrated using a sample database and a tutorial example. This study will provide insights on (1) how text mining can be used to develop a descriptive model, and (2) how descriptive terms generated during the text mining process can be utilized to provide a basic set of standard lexicon to develop a standardized taxon description recommendation. Furthermore, advantages and drawbacks of the proposed framework will be discussed and future research directions will be proposed.