Enhancing an Evolving Tree-based Text Document Visualization Model with Fuzzy c-Means Clustering

[1]Wui Lee Chang, [1*]Kai Meng Tay, [2]Chee Peng Lim
[1]Faculty of Engineering, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia.
[2] Centre for Intelligent Systems Research, Deakin University, Australia
E-mail: *kmtay@feng.unimas.my

***Abstract*— An improved evolving model, i.e., Evolving Tree (ETree) with Fuzzy c-Means (FCM), is proposed for undertaking text document visualization problems in this study. ETree forms a hierarchical tree structure in which nodes (i.e., trunks) are allowed to grow and split into child nodes (i.e., leaves), and each node represents a cluster of documents. However, ETree adopts a relatively simple approach to split its nodes. Thus, FCM is adopted as an alternative to perform node splitting in ETree. An experimental study using articles from a flagship conference of Universiti Malaysia Sarawak (UNIMAS), i.e., Engineering Conference (ENCON), is conducted. The experimental results are analyzed and discussed, and the outcome shows that the proposed ETree-FCM model is effective for undertaking text document clustering and visualization problems.**

Keywords*— Evolving tree; text document clustering; visualization; online learning; fuzzy c-means

I. INTRODUCTION

Clustering is a task of assigning data objects into a number of groups (or clusters) so that the objects in the same cluster share the same similarities, as compared with those in other clusters [1]. It converts a set of non-linear data into a human and/or machine understandable format, which can be very useful for unsupervised learning systems. Examples of some popular clustering methods are the Self-Organizing Map (SOM) [2, 3], k-mean clustering [4], and fuzzy c-mean clustering (FCM) [5, 6]. With respect to SOM, it is an artificial neural network that maps a set of high-dimensional data onto a predefined low-dimensional grid of nodes [2, 7], and retains the topological relationship of the data. From the literature, various applications of SOM, e.g., speech recognition [8, 9], feature extraction [10], robotic arm [11], noise reduction in telecommunication [12], and textual documents clustering [13], have been reported. Indeed, various extensions of SOM, e.g., hierarchical search [14, 15], growing SOM [16, 17], growing hierarchical SOM [18], and evolving tree (ETree) [19], have been proposed over the years. Examples of applications of ETree to a variety of application domains can be found in [20-22]. In general, these approaches increase the flexibility of SOM and improve the learning time for processing large data samples.

Text document clustering (also known as text categorization) is a procedure to assemble similar text documents into groups based on their similarity [23]. Many text document clustering methods are available. Examples include the naive Bayes-based document clustering model [24], WEBSOM [25], and support vector machines-based model for imbalanced text document classification [26]. These approaches allow a collection of documents to be clustered (and visualized). Regardless of the popularity of these approaches, it is not sure how these approaches can be extended to evolving or online learning. Thus, it is important to develop a text document clustering model with evolving capabilities for the following reasons: (1) new documents are generated or created everyday; and (2) it is not practical to perform re-training of a model whenever a new document appears. Thus, an evolving model is useful for tackling text document clustering problems.

The focus of this paper is on an improved evolving model, i.e., ETree combined with FCM (denoted as ETree-FCM), as an alternative to SOM as well as other offline learning methods, for document clustering. Instead of SOM-based models (e.g. WEBSOM [25]), ETree-FCM allows the clustering method to have evolving features. Besides that, it serves as a solution to a few shortcomings of WEBSOM, i.e., the learning time [19], and the difficulty in determining the map size before learning [19]. Even through ETree has been proposed as an alternative to SOM, its application is still limited. To the best of our knowledge, this study is a new attempt to use ETree-FCM in document clustering [27].

It is worth mentioning that ETree adopts a relatively simple approach to allow a trunk node to be split into child nodes. We have previously developed an ETree-based text document clustering and visualization procedure [27]. In this paper, we further extend our previous work by adopting FCM to allow a trunk node to be split into child nodes. FCM is chosen because it is a reliable clustering method [28]. In our proposed ETree-FCM model, some salient features of WEBSOM, e.g., text pre-processing, are retained. A case study with information/data from a flagship conference of Universiti Malaysia Sarawak (UNIMAS), i.e., the Engineering Conference (ENCON), is conducted to evaluate the effectiveness of the proposed approach.

This paper is organized as follows. The background of ETree and FCM are provided in Section II. In Section III, the use of ETree-FCM for text document clustering and visualization is described. An experimental study to evaluate the usefulness of ETree-FCM in text document clustering is presented in Section IV, with the results analyzed and discussed. A summary of conclusion is included in Section V.

II. BACKGROUND

A. *Structure of ETree*

Fig. 1 depicts an example of the ETree structure. The tree structure consists of n_{node} nodes. Each node is denoted as $N_{l,j}$, where $l = 1,2,3, \dots n_{node}$ is the identity of the node and