World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:10, 2015

A Methodology for Automatic Diversification of Document Categories

Authors : Dasom Kim, Chen Liu, Myungsu Lim, Su-Hyeon Jeon, ByeoungKug Jeon, Kee-Young Kwahk, Namgyu Kim

Abstract : Recently, numerous documents including unstructured data and text have been created due to the rapid increase in the usage of social media and the Internet. Each document is usually provided with a specific category for the convenience of the users. In the past, the categorization was performed manually. However, in the case of manual categorization, not only can the accuracy of the categorization be not guaranteed but the categorization also requires a large amount of time and huge costs. Many studies have been conducted towards the automatic creation of categories to solve the limitations of manual categorization. Unfortunately, most of these methods cannot be applied to categorizing complex documents with multiple topics because the methods work by assuming that one document can be categorized into one category only. In order to overcome this limitation, some studies have attempted to categorize each document into multiple categories. However, they are also limited in that their learning process involves training using a multi-categorized document set. These methods therefore cannot be applied to multi-categorization of most documents unless multi-categorized training sets are provided. To overcome the limitation of the requirement of a multi-categorized training set by traditional multi-categorization algorithms, we previously proposed a new methodology that can extend a category of a single-categorized document to multiple categorizes by analyzing relationships among categories, topics, and documents. In this paper, we design a survey-based verification scenario for estimating the accuracy of our automatic categorization methodology.

Keywords : big data analysis, document classification, multi-category, text mining, topic analysis
Conference Title : ICDE 2015 : International Conference on Data Engineering
Conference Location : Bali, Indonesia
Conference Dates : October 11-12, 2015