Automatic Folder Allocation System for Electronic Text Document Repositories Using Enhanced Bayesian Classification Approach

Wou Onn Choo, Faulty of Information Technology and Sciences, INTI International University, Nilai, Malaysia
Lam Hong Lee, School of Computing, Faculty of Science and Technology, Quest International University Perak, Ipoh, Malaysia
Yen Pei Tay, School of Computing, Faculty of Science and Technology, Quest International University Perak, Ipoh, Malaysia
Khang Wen Goh, School of Computing, Faculty of Science and Technology, Quest International University Perak, Ipoh, Malaysia
Dino Isa, Department of Electrical and Electronic Engineering, Faculty of Engineering, The University of Nottingham, Semenyih, Malaysia
Suliman Mohamed Fati, INTI International University, Nilai, Malaysia

ABSTRACT

This article proposes a system equipped with the enhanced Bayesian classification techniques to automatically assign folders to store electronic text documents. Despite computer technology advancements in the information age where electronic text files are so pervasive in information exchange, almost every single document created or downloaded from the Internet requires manual classification by the users before being deposited into a folder in a computer. Not only does such a tedious task cause inconvenience to users, the time taken to repeatedly classify and allocate a folder for each text document impedes productivity, especially when dealing with a huge number of files and deep layers of folders. In order to overcome this, a prototype system is built to evaluate the performance of the enhanced Bayesian text classifier for automatic folder allocation, by categorizing text documents based on the existing types of text documents and folders present in user’s hard drive. In this article, the authors deploy a High Relevance Keyword Extraction (HRKE) technique and an Automatic Computed Document Dependent (ACDD) Weighting Factor technique to a Bayesian classifier in order to obtain better classification accuracy, while maintaining the low training cost and simple classifying processes using the conventional Bayesian approach.

KEYWORDS

Automatic Folder Allocation, Bayes Formula, Document Classification, High-Relevance Keyword Extraction Technique

1. INTRODUCTION

With meteoric growth of personal computers and rapid advancements of cloud computing technologies, the use of electronic data most notably electronic text documents are so pervasive in Internet information exchange. Living in the Big Data era today, the high-volume and high-velocity nature of
electronic data exchange motivates creation of a utility to automatically categorize and assign newly created or incoming electronic text documents into the most appropriate folders on a computer. In the absence of such feature, an average computer user shall perform at least three manual steps in (1) reviewing the type of the file or filename to be stored, (2) determining the appropriate folder to store the file, and (3) moving the actual file to the desired folder on the computer. Not only does such tedious task causes inconvenience to many modern computer users, the time taken to repeatedly review, classify and allocate folder for each text document impedes productivity, especially dealing with huge number of files and deep layer of folders. In order to overcome this, an enhanced Bayesian text classifier system is built to perform automatic folder allocation by categorizing new incoming text documents based on the existing types of text documents and folders present in user’s computer hard drive.

With the automatic folder allocation technique, computers are able to recognize and classify incoming text documents and determine the most appropriate folder for storing, without requiring extensive manual interventions from the user. This can greatly reduce the time taken for human–computer interaction in allocating the appropriate folder path to store similar documents while also improving the overall information retrieval process.

Text document classification denotes the task of assigning text documents to one or more predefined categories. This is a direct concept from machine learning, which implies the declaration of a set of labelled categories as a way to represent the documents, and a statistical classifier trained with a labelled training set. Classification is the process in which objects are initially recognized, differentiated and understood, and implies that objects are grouped into categories, usually for some specific purposes. Ideally, a category represents a relationship between the subject and object of knowledge. Classification is fundamental in prediction, inference, and decision-making. However, there are varieties of ways to approach the classification task. An increasing number of supervised approaches have been developed for document classification, such as decision tree induction (Greiner and Schaffert, 2011), rule induction (Apte et al., 1994), k-nearest neighbour classification (Han et al., 1999), maximum entropy (Nigam et al., 1999) artificial neural network (Chen et al., 2005; Diligenti et al., 2003), support vector machines (Isa et al., 2008; Joachims, 1998; Joachims, 1999) and Bayesian classification (Androustopoulos et al., 2000; Chen et al., 2009; Domingos and Pazzani, 1997; Eyheramendy et al., 2003; Kim et al., 2002; Lee et al., 2010; Lee et al., 2012-a; Lee and Isa, 2010; McCallum and Nigam, 2003; Rish, 2001; Sahami et al., 1998). Besides the supervised classification approaches, unsupervised clustering approaches such as k-means and self-organizing map, have also been introduced for text document segmentation (Adami et al., 2005; Isa et al., 2009; Lee and Yang, 2003; Takamura, 2003).

Among the approaches mentioned above, Bayesian classification approach has gained its popularity and has been widely implemented in multitude of text document classification and folder allocation applications in recent years (Al-Mubaid and Shenify, 2016; Choo et al., 2014; Lee et al., 2012-b, Tang et al., 2016, Diab and El-Hindi, 2017; Viegas et al., 2018), due to its simplicity in both training and classifying stage. Although this generative method has been reported to be less accurate than the discriminative methods such as support vector machines (Isa et al., 2008; Joachims, 1998; Joachims, 1999; Kim et al., 2002; Brücher et al., 2002; Chakrabarty et al., 2003; Lin, 1999; Yang and Pedersen, 1997; Yang and Liu, 1999), many empirical evidence have proven that Bayesian classification approach provides intuitive text generation models and therefore performs well in many application domains, under specific conditions (Domingos and Pazzani, 1997; Kim et al., 2002; Lee et al., 2010; Lee et al., 2012-a; Lee and Isa, 2010; McCallum and Nigam, 2003; Rish, 2001). Bayesian classifier is a simple probabilistic classifier based on the implementation of Bayes’ Theorem. Depending on the precise nature of the probability model, the Bayesian classifier can be efficiently trained with small training data set to estimate the parameters required for classification. Due to the independent variables assumed, only the variances of variables for each class need to be determined and not the entire covariance matrix. Due to such apparently over-simplified assumptions, Bayesian classifier
Performance Analysis of Structured, Un-Structured, and Cloud Storage Systems
www.igi-global.com/article/performance-analysis-of-structured-un-structured-and-cloud-storage-systems/216467?camid=4v1a

DEAL: A Distributed Authorization Language for Ambient Intelligence
www.igi-global.com/article/deal-distributed-authorization-language-ambient/61137?camid=4v1a