Chapter 32
Research on Digital Forensics Based on Uyghur Web Text Classification

Yasen Aizezi
Xinjiang Police college, Urumqi, China

Anwar Jamal
Xinjiang Police College, Urumqi, China

Ruxianguli Abudurexiti
Xinjiang Police College, Urumqi, China

Mutalipu Muming
Xinjiang Police College, Urumqi, China

ABSTRACT

This paper mainly discusses the use of mutual information (MI) and Support Vector Machines (SVMs) for Uyghur Web text classification and digital forensics process of web text categorization: automatic classification and identification, conversion and pretreatment of plain text based on encoding features of various existing Uyghur Web documents etc., introduces the pre-paratory work for Uyghur Web text encoding. Focusing on the non-Uyghur characters and stop words in the web texts filtering, we put forward a Multi-feature Space Normalized Mutual Information (M-FNMI) algorithm and replace MI between single feature and category with mutual information (MI) between input feature combination and category so as to extract more accurate feature words; finally, we classify features with support vector machine (SVM) algorithm. The experimental result shows that this scheme has a high precision of classification and can provide criterion for digital forensics with specific purpose.

DOI: 10.4018/978-1-7998-3025-2.ch032

INTRODUCTION

Due to the rapid development of information and storage technologies, especially wide appli-cation of cloud computing technology, a lot of Uyghur Web information has been stored in various major information systems with the Internet as carrier. To prevent, strike and control harmful information better, it is required to use digital forensics technology to conduct deep analysis on data stored and discover the law and relationship of various case analyses (Chun-Hui & Qin-Ming, 2009). How to classify Uyghur Web documents rapidly, analyze specific type accurately and extract useful information in the process of digital forensics when facing a lot of Uyghur Web documents is a major problem to be solved by forensics. Text classification technology in data mining is an effective method for solving such problems (Lu et al., 2013).

With great investment of the state into Xinjiang region, the infrastructure construction has developed rapidly. A lot of textual information of minority language such as Uyghur starts to be presented in digital form. There are many Uyghur websites on the Internet, which provide many different types of services. Though most information is useful, there are some objectionable contents on the webpage of some Uyghur websites. Classifying a lot of Web textural data on Uyghur websites so as to conduct digital forensics and hit relevant unlawful acts effectively can maintain the stability in Xinjiang region and has great significance (Parhat et al., 2014). Currently, text classification technologies for majority language such as English and Chinese have been studied greatly and tend to be mature. However, relevant researches on the classi-fication of Uyghur digital texts are still in its early stage. Uyghur is an adhesive language which has complicated change of tenses and rich morphological structure (Tohti et al., 2014). Therefore, literature (Tohti et al., 2014) puts forward a classification method of Uyghur texts based on feature extraction of semantic words and uses a composite statistic (DME) to measure the (Aysa et al., 2012) degree of correlation between adjacent words in text so as to extract feature words. Literature uses statistic $\chi 2$ to extract stem and establishes Uyghur text classifier with support vector machine (SVM) algorithm. Literature (Aysa et al., 2015) puts forward a new statistic (CHIMI), combines statistic $\chi 2$ and mutual information (MI) to constitute CHIMI, extracts Bigram as textual feature and uses SVM algorithm to classify Uyghur text.

This paper puts forward a Uyghur Web document classification scheme based on text classify-cation for digital forensics of Web text based on the improvement of traditional MI feature extraction. It extracts features of Uyghur web text with improved normalized mutual information algorithm and uses SVM for text classification so as to extract illegal or objectionable text infor-mation effectively.

SCHEME IN THIS PAPER

This paper puts forward a Uyghur digital forensics scheme based on text classification, mainly including 3 parts: (1) Uyghur text pre-processing; (2) feature extraction; (3) text classification. In the stage of feature extraction, this paper improves traditional MI feature extraction which only considers MI of single feature and category and fails to consider the relevance between contextual features, and replaces MI of single feature and category with MI between combination and category of input features.