Pornography web pages classification with principal component analysis and independent component analysis

Abstract

The impressive growth of internet has made a new evolution of human life. Internet is an information superhighway but also the most unsecured place. Web users always need to take the risk for theft of information, spamming, virus threat and mental pollution of harmful resource. The illicit web content such as pornography, violence, gambling, etc. have greatly polluted the mind of immature web users. Pornography perhaps is one of the biggest threats related to current children's and teenagers' healthy mental life. There are thousands of pornography sites on the internet can be easily found and detected. This will certainly become a detrimental factor to letting children and teenagers access internet without proper guidance. This fact makes web filtering systems are highly required in family and education environment. Web filter normally provide two major services which are protection against inappropriate content and preventing misuse of network [1]. Current filtering approaches such as URL blocking, keyword matching and rating system like PICS (Platform for Internet Content Selection) rating are widely implemented in today commercialize web filtering systems. The URL blocking technique will restrict or allow the web users to access web sites by checking required URL with sets of URL list stored in database. The problem of this technique is with current limited technology, it is hard to obtain the complete up–to-date URL list since there are an estimated 1 billion web pages being added daily [2].

This technique is costly in maintaining and insufficient against unknown web content. On the other hand, the trust issue is always an argument for PICS rating technique since the web publishers have the right to label whatever content to the metadata. Hence PICS is only suggested as a supplementary filtering technique due to its weakness against ever-changing web content. The keyword matching technique is designed to overcome the dynamic content issues; however it is not efficient during different subjects but having similar terminologies web pages. For instances this technique will block both pornography and gynecology web pages since intentionally we only need to block pornography web pages. Under-block and over-block are always the issues for this technique. Ordinary illicit web pages are constructed by mixing textual hyperlinked content with visual content. We could tackle the dynamic web content issues by

using content based analysis approaches. Since most of the web pages contain textual information, so, we mainly focus on textual content based analysis. In fact, the approaches of current content based analysis mostly rely on machine learning process. Yu et al. [3] classify web pages by implementing their proposed framework, Positive Example Based Learning (PEBL) which uses support vector machine (SVM) as a classifier. Lee et al. [4] classify the documents with fuzzy learning technique and Selamat et al. [5] categorize the Japanese sport news with artificial neural network (ANN).