Detecting and Removing Noisy Data on Web Document using Text Density Approach

Abstract

The web documents content are useful resources for many applications. However, this content could be classified into relevant content and irrelevant content with respect to the involved application. The irrelevant content, like advertisements banner, copyright information, and navigation menus assumed as noisy data. Noisy data that found among the content of the web document affects negatively the performance of most of applications that deals with the content of web pages. The process of detecting and removing noisy data is an important pre-processing step in many applications such as web page classifications, clustering of web pages and information retrieval tasks. We developed a unified algorithm able to detect automatically the noisy data and eliminate them out of the web page and produce a clear web document that could be used effectively in later steps. The suggested approach examined using a dataset composed of different classes. The results of the conducted experiments showed a significant enhancement in the problem of detecting and removing noisy.

References

- David Gibson, Kunal Punera, and Andrew Tomkins. 2005. The Volume and Evolution of
Detecting and Removing Noisy Data on Web Document using Text Density Approach


Index Terms

Computer Science
Web Technology
Keywords
Web Page Segmentation  Noise Removal  Information Extraction