Abstract

The World Wide Web is the most increasingly growing and accessible source of information. Web contents of different fields which can offer important information to users are available in the Web like multimedia data, structured, semi-structured and unstructured data. But only a part of the information is useful for a particular application and the remaining information are considered as noises. Data on web pages contain formatting code, advertisement, navigation links, etc. This collection of unwanted noise with the real content in a web page complicates the task of automatic information extraction and processing. This requires the extraction of useful noise-free information. Otherwise, it can ruin the effectiveness of Web mining techniques. This paper proposes a novel method to filter web pages and retrieve the actual content of a web page. This research work proposed an approach for removing the noises from a given web page which will improve the performance of web content mining. At first, the web page information is divided into various blocks which then tokenized to separate the informative content from noises. This paper presents algorithm for removing noises from the web page and automatically extract important web content. This paper also presents the algorithm for global noise removal.

References
Noise Reduction and Content Retrieval from Web Pages


L. Yi, B. Liu, X. Li. 2003. Eliminating Noisy Information in Web Pages for Data Mining, SIGKDD. 03, August 24-27, IEEE.


World Wide Web: Internet and Web Information Systems, 8, 179–224, Springer.

Index Terms

Computer Science

Information Sciences

Keywords

Web content mining  content retrieval  noises  outlier  redundancy  precision  recall  accuracy