Abstract

The rapid development of the internet and web publishing techniques create numerous information sources published as HTML document on World Wide Web. WWW is now a popular medium by which people all around the world can spread and gather the information of all kinds. But web document of various sites that are generated. Contain undesired information also. This information is called noisy or irrelevant content. The need for innovative and effective technologies to help find and use the useful information and knowledge from a large variety of data sources is continually increasing. Web information has become increasingly diverse. In order to utilize the Web information better, people pursue the latest technology, which can effectively organize and use online information. Classification is one of the vital and important data mining techniques that grouped various items in a collection to predefined classes or groups. The main goal of classification is to exactly predict the target class for each case in the data. Web Document Classification is technique of data mining to discover classification of Web Documents. The information providers on the web will be interested in techniques that could improve the effectiveness of the web search engine. In this paper, the relationships among the
Number 26

techniques used in data mining are studied. A study of web usage is also done on optimization of this web classification.

References

10. Daniele Riboni, D.S.I., University’ degli Studi di Milano, Italy,“Feature Selection for Web Document Classification”.

Index Terms

Computer Science

Information Sciences

Keywords
Structured data, Unstructured data, Rough Set Theory, Web Document Classification.