Abstract

Nowadays the users of the WWW are not only the human. There are other users or visitors like web crawlers and robots which are generated by the search engines or information retrievers. The direct visitors of your website are very less than those who reach to your website by using search engines or through other links. To collect information from your website search engines use crawlers or robots to access your website. There must be an access mechanism or protocol for such robots which restrict them to access unwanted content of the website. robots. txt is a partial mechanism for such facilities but not fully functional. This paper gives an enhancements to fully make use of the functionality of robots. txt file.

References

- Standard for the Format of ARPA Internet Text Messages: https://www.ietf.org/rfc/rfc0822.txt
- W3C http://www.w3.org/
Restructuring robots. txt for better Information Retrieval

- Timestamp http://en.wikipedia.org/wiki/Timestamp

Index Terms

Computer Science

Web Services

Keywords

Crawling agents robots spammer harvesters User Agent tag Directive

Overriding
Web Crawling
Web Tree
Web Spamming
Crawling
Querying.