Yogita Rahulsing Chavan, J Data Mining In Genomics & Proteomics 2015, 6:4
http://dx.doi.org/10.4172/2153-0602.C1.003

OMICS International
conferenceseries.com

2nd International Conference on
Big Data Analysis and Data Mining
November 30-December 01, 2015 San Antonio, USA

Effective web data extraction

Yogita Rahulsing Chavan
New Horizon Institute of Technology and Management, India

Web data extraction is one of the very popular research activities that aim at extracting useful information from web pages. Such extracted information is then stored into the database that can be used for faster access to the data in the applications like comparison shopping, information integration, etc. Several efforts have already been carried out and used in the past. Some of the techniques are record level while the others are page level. An efficient algorithm has been proposed by W. Su et al., for extracting useful information from web pages using the concepts of tags and values. The algorithm constructs a DOM tree from the source code associated with the page (that is HTML code). Data regions are formed by inspecting similar nodes in the tag tree. One or more data regions formed during this step are then merged if similarity is found. However the method discards non matching first node that represents non auxiliary information in the data region and thus results in loss of information. The research work deals with implementation of the algorithm mentioned above. It also extends the algorithm to overcome the problem of loss of information.

Biography

Yogita Rahulsing Chavan has completed her Master's in Computer Engineering from Pune University, Maharashtra. She is currently working as an Assistant Professor in New Horizon Institute of Technology and Management, Thane (W), Maharashtra, India. She has published around 7 papers in several conferences and journals. She has been working in teaching field in Engineering Institute for more than 10 years.

yogita84@gmail.com

Notes: