Gogineni K Chaitanya, J Comput Sci Syst Biol 2015, 8:3
http://dx.doi.org/10.4172/0974-7230.S1.012

OMICS International
conferenceseries.com

International Conference on
Big Data Analysis and Data Mining

May 04-05, 2015 Kentucky, USA

Hidden web data extraction using word net ontologies

Gogineni K Chaitanya
NRI Institute of Technology, India

In response to the search engine crawler's queries, the application servers generate the information and deliver it directly to the user. The generated information forms the hidden web (deep web or invisible web) because the information is usually enwrapped in Hyper Text Markup Language (HTML) pages as data records. Due to the dynamic nature of the generated data records from the hidden web, current search engines (either general or commercial) are unable to index the HTML page accordingly. Propose to develop an Ontological Wrapper (OW) for the extraction and alignment of data records using lightweight ontological technique driven by word net repositories. Main component of the wrapper involves checking the similarity of data records and not just visual cues by stripping the html aspects. There are three main components in our wrapper design, namely, parsing process performed with TEXT MDL Algorithm, extraction initiated with irrelevant HTML stripping, and alignment of data records for classification. After the three step process, we are left with pure text data records stripped of the html content which can be searched over by humans or search engine crawlers. Our approach is almost adaptable to most websites of distinguished visual cues and yields better data extraction results at better speeds than prior systems and a practical implementation validates our claim.

Biography

Gogineni K Chaitanya completed his Masters from JNTU Kakinada University. He is Microsoft certified programmer and is working as Assistant Professor at NRI Institute of Technology, in the Department of Computer Science Engineering. He has presented papers at various international, national conferences and has published papers in reputed journals.

mail2gogineni@aol.com

Notes: