Web Mining: A Brief Survey on Data Extraction Techniques

Abstract

In last few years, we faced problem regarding extraction of data from web pages. In this paper we proposed to address problem of web data extraction techniques related to areas such as natural language processing, language and grammar, machine learning, information retrieval and Ontologies. As consequence they represent very distinct feature and capabilities which make direct comparison difficult to be done.

References

- Bray. T. , Pauli. J. And Mcqueen M. S Extensible Markup Language (Xml) 1. 0
- Crescenzi V. And Mecca G. Grammars have Exceptions, Information Systems 23,8.
- Embley D. W. , Jiang Y. S. And Ng. Y. K. Record Boundary Discovery in web documents. In proceedings ACM SIGMOD International Conference on Management of Data (Pholadephia, Penssylvania, USA, 1999).
- Golgher P. B. , Di Silva A. S. , Laender A. H. F. And Ribiero Neta. B. A Bootstrapping for example based Data extraction. In proceedings of tenth ACM International Conference on Information and Knowledge management (Atlanta, Georgia, 2001)
- Ion Muslea, Rise:Repositionary of Online Information Sources used in Information
Web Mining: A Brief Survey on Data Extraction Techniques

- Liu, L., Pu., C And Han, W. Xwrap. An Xml Enabled Wrapper Construction System for web Information Sources. In proceeding of 16th IEEE International Conference on Data Engineering (SanDiego, California, 2000)
- Papakonstantinou Y., Garcia-Molina, H., And Widom J. Object Exchange Across Heterogeneous Information Sources. In proceeding of IEEE 11th International Conference on Data Engineering. (Taipei, Taiwan, 1995)
- Teixeira, J. S. A Comparative study of Approaches for semostructured Data Extraction. Master’s Thesis. Department of Computer Science, Federal University of Minas Gerais, Brazil, 2001
- Satyajeet Nimgaonkar and SuryaprakhDuppala, &quot;A survey on web content mining and extraction of Structured and Semi structured data&quot;

Index Terms

Computer Science
Emerging Trends in Technology
Keywords
Ontology  Web Data Extraction  Web Crawler  Road Runner