International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 4, Issue 2, Apr 2014, 109-114

TRANS STELLAR
• Journal Publications • Research Consultancy

WEB-SCALE INFORMATION EXTRACTION FROM UNSTRUCTURED AND UNGRAMMATICAL DATA SOURCES

MADHAVI K. SARJARE[1] & S. L. VAIKOLE[2]

[1]Department of Computers, Yadavrao Tasgoankar College of Engineering and Management, Karjat, Mumbai, Maharashtra, India

[2]Department of Computers, Datta Meghe College of Engineering, Airoli, Navi Mumbai, Maharashtra, India

ABSTRACT

Information Extraction (IE) is the task of automatically extracting knowledge from text. The massive body of text now available on the World Wide Web presents an unprecedented opportunity for information extraction. However, information extraction on the Web is challenging due to the enormous variety of distinct concepts and structured expressed. The explosive growth and popularity of the worldwide web has resulted in a huge amount of information sources on the Internet. However, due to the heterogeneity and the lack of structure of Web information sources, access to this huge collection of information has been limited to browsing and searching.

Information extraction from unstructured and ungrammatical text on the Web, such as classified Ads, Auction listings, and web postings forums. Since the data is unstructured and ungrammatical, this information extraction precludes the use of rule-based methods that rely on consistent structures within the text or natural language processing techniques that rely on grammar. Posts are full of useful information, as defined by the attributes that compose the entity within the post.

Currently accessing the data within posts does not go much beyond keyword search. This is precisely because the ungrammatical and unstructured nature of posts makes extraction difficult, so the attributes remain embedded within the posts. These data sources are ungrammatical, since they do not conform to the proper rules of written language. Therefore, Natural Language Processing (NLP) based information extraction techniques are not appropriate.

As more and more information comes online, the ability to process and understand this information becomes more and more crucial. Data integration attacks this problem by letting users query heterogeneous data sources within a unified query framework, combining the results to ease understanding. However, while data integration can integrate data from structured sources such as databases, semi-structured sources such as that extracted from Web pages, and even Web Services, this leaves out a large class of useful information: unstructured and ungrammatical data sources.

We proposed a system based Machine Learning technique to obtain the structured data records from different unstructured and non-template based websites. The proposed approach will be implemented by collection of known entities and their attributes, which refer as "*reference set*," A reference set can be constructed from structured sources, such as databases, or scraped from semi-structured sources such as collections of Web pages. A reference set can even be constructed automatically from the unstructured, ungrammatical text itself. This project implements methods to exploit reference sets for extraction using machine learning techniques. The machine learning approach provides higher accuracy extractions and deals with ambiguous extractions, although at the cost of requiring human effort to label training data.

KEYWORDS: Natural Language Processing, Reference set, Nested String List, Hypertrees