Semi-guided Self Learning Data Structuring System

About

TicketCatcher is a young and vibrant company in an inspiring environment. We are currently building a system in which semi-structured event data in PDFs should be scraped and structured. Because the PDFs with the event information are constantly changing we are looking for a system which by way of partial guidance from its users can act in a self-learning way so that new PDF-structuring by the ticket issuers can be continuously adapted to.

Introduction

Extraction and structuring of relevant information from an unstructured and dynamic source has been a challenge for the field of data mining. Although advances have been made through the usage of NoSQL techniques and synonym-cloud building, the usage of loosely based rules in the form of fuzzy logic in combination with decentralized and incremental user guidance and evolution strategies remains a field of research.

Proposed Method

Seeing that the data at hand is partly structured, the proposed system could benefit from a rather structured start. Rules are laid out for a small subset of data files which can extract and structure the relevant information. From this point on, the system should rely on feedback from users which comes in two forms: A) Boolean type true/false, and B) suggested corrections. Seeing that the feedback will not be fully correct, nor will it judge or correct all outcomes, the usage of fuzzy logic will be used to determine the expected correctness of information based on several to be determined (and evolved) indicators. Based on the feedback and the outcomes of the fuzzy logic, the rules for data extraction can be evolved using evolution strategies. This means that the confirmation of the correctness of the rules can contract the mutation process to focus on incremental improvements. Similarly, the rejection and correcting of the results can be lead to an increase of mutation strength, thereby exploring more experimental extraction techniques. Finally, the system should not only converge using incremental evolutionary methods, but should also cater for exceptions and outliers. If exceptional high levels of negative feedback (i.e.: “falses”) is registered. The system should recognize these high levels and pair these cases with similar cases of high disapproval rates, finding common

denominators by eliminating all variant information in the data files. If common denominators are found, the suggested corrections could then be employed for similar data files by comparing hashes of cleaned (variant eliminated) data files.
Combining these methods of fixed starting rules, incremental but incomplete user guidance, fuzzy logic, evolution strategies, and outlier handling is expected to demonstrate superior self learning data extraction capabilities.

Research method

The research will be employed with a real life dynamic and evolving data set, from which (music/cultural) event data should be extracted. Systems can be tested with guidance given both by controlled and informed users, as well as uninformed real life users. Research results are based on manual counting of correctly extracted and recognized data.

We are looking for

An enthusiastic person who can turn the box inside out. A person which likes to tinker about new solutions to existing problems. Making it work is key, making it perfect is next. Of course we work with the latest technologies, but foremost we work with technologies we like; Linux, Python, Bash, and whatever you like yourself. Interested? Come and talk to us!

Contact

Please contact us for further information:

TicketCatcher BV
Hielke Kramer (Lead Developer)
+31618449998
kramerh@gmail.com
Johan Huizingalaan 763a
1066VH
Amsterdam