Non IT configurable adaptive data mining solution used in transforming raw data to structured data

A proposal

Mart Lubbers
0651371972
s4109503
mart@martlubbers.net

May 1, 2014

Contents

1 Supervisors

Franc Grootjen
Radboud University Nijmegen
Nijmegen, The Netherlands
f.grootjen@psych.ru.nl

Signature

Alessandro Paula
Hyperleap
Nijmegen, The Netherlands
aldo@hyperleap.nl

Signature

2 Abstract73 words

Raw data from information providers is usually hard to interpret for a software solution and the conversion of raw data to structured data is usually done by hand. This project aims towards an adaptable, configurable data transformation program optionally in combination with a webcrawler that can perform the conversion from raw data to structured data. This is all done in under supervision of Franc Grootjen and Alessandro Paula and under commissioned by Hyperleap.

3 Project Description 484 words

3.1 Research Question and Motivation

The main research question is: *How can we make an adaptive, autonomous and programmable data mining program that can be set up by a non IT professional which is able to transform raw data into structured data.*
Hyperleap is a small company that is specialized in infotainment (information+entertainment) and administrates several websites which bundle information about entertainment in a ordered and complete way. Right now, most of the data input is done by hand and takes a lot of time to type in.

3.2 Aim

The practical goal and aim of the project is to make a crawler(web or other document types) that can autonomously gather information after it has been setup by a, not necessarily IT trained, employer via an intuitive interface. Optionally the crawler shouldn't be susceptible by small structure changes in the website, be able to handle advanced website display techniques such as javascript and should be able to notify the administrator when the site has become uncrawlable and the crawler needs to be reprogrammed for that particular site. But the main purpose is the translation from raw data to structured data. The projects is in principle a continuation of a past project done by Wouter Roelofs[**?**] which was also supervised by Franc Grootjen and Alessandro Paula, however it was never taken out of the experimental phase and therefore is in need continuation.

3.3 Research Plan and Schedule

The schedule or plan for the project can be divided into 4 stages namely the initial, developmental, testing and writing stage. These stages are not mutually exclusive and therefore can and will overlap.

- Initiating stage: In this stage we will look at the past project and present literature on the subject and create a explicit plan for the eventual software. There probably is a lot of literature written on how to parse certain information fields such as dates, places and artist information. The date parsing and recognizing was a main part in the past project.

- Developmental stage: The developmental stage is the stage where most of the programming is done and the where the algorithms for crawling and transformation are implemented. For web-frontend the framework choice has fallen upon firefox extensions which are mainly written in javascript and cfx. The data transformer will probably be written in python due to the robust natural language tools and portability.

- Testing stage: This stage will overlap greatly with the developmental stage because this will save a lot of time.
- Writing stage: The last stage will be the stage in which the thesis is written and the project presented. During all other stages certain parts of the thesis can already be written down.

3.4 Weekly planning

Because of some mandatory courses in the first semester of the next year the schedule can be seen as provisional meaning that there is room to extend the schedule.(in practice at maximum up to december 2014).

#	Wk	Date	Task	Deliverables
1	15	2014-04-07	proposal and references	proposal signed by both parties
2	16	2014-04-14	references and test environment setup	test environment
3	17	2014-04-21	planning for writing the tool	software design
4	18	2014-04-28	writing thesis and programming	introduction
5	19	2014-05-05	writing thesis and programming	
6	20	2014-05-12	idem	methods
7	21	2014-05-19	idem	first prototype software
8	22	2014-05-26	testing, programming and thesis	
9	23	2014-06-02	testing, implementation bigger picture	
10	24	2014-06-09	testing	working tool and results and abstract
11	25	2014-06-16	presentation and thesis	discussion and presentation
12	26	2014-06-23	presentation	presentiation
13	27	2014-06-29	presentation	

There will also be bi-weekly meetings with both supervisors to make sure we are on schedule. If necessary the frequency of meetings with the external supervisor can be increased.

4 Scientific relevance52 words

Currently the techniques for conversion from non structured data to structured data are static and mainly only usable by IT specialists. There is a great need of data mining in non structured data because the data within companies and on the internet is piling up and are usually left to catch dust.

References

[1] W. Roelofs, A. T. Paula, and F. Grootjen, “Programming by Clicking,” in *Proceedings of the Dutch Information Retrieval Conference*, pp. 2–3, 2009.