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

Bachelor's Thesis in Artificial Intelligence
Radboud University Nijmegen

Franc Grootjen
RU

Alessandro Paula
Hyperleap

Mart Lubbers
s4109053

July 23, 2014

Contents

Abstract

Abstract

1 Introduction

1.1 Introduction

Within the entertainment business there is no consistent style of informing people about the events. Different venues display their, often incomplete, information in entirely different ways. Because of this, converting raw information from venues to structured consistent data is a relevant problem.

1.2 HyperLeap

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.

1.3 Research question

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.*

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.

1.4 Scientific relevance

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

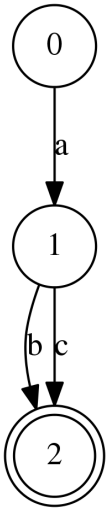
non structured data because the data within companies and on the internet is piling up and are usually left to catch dust.

2 Methods

2.1 Directed acyclic graphs and finitie automata

Directed acyclic graphs(DAG) and finite state automatas(FSA) have a lot in common concerning pattern recognition and information extraction. By feeding words into an algorithm a DAG can be generated so that it matches certain patters present in the given words. Figure 2.1 for example shows a FSA that matches on the words *ab* and *ac*.

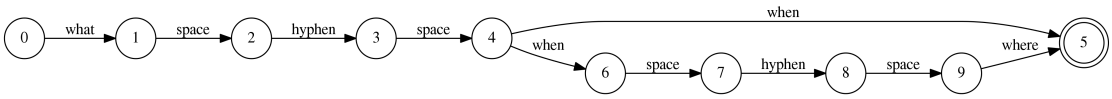
Figure 2.1: Example DAG/FSA



With this FSA we can test if a word fits to the constraints it the FSA describes. And with a little adaptation we can extract dynamic information from semi-structured data.

Nodes in this datastructure can be single letters but also bigger constructions. The example in Figure 2.2 describes different separator pattern for event data with its three component: what, when, where.

Figure 2.2: Example event data



2.2 Algorithm

Hello Wordl

3 Results

4 Discussion

5 Appendices