Non IT configurable 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 24, 2014

Contents

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 challenging and, relevant problem.

1.2 HyperLeap

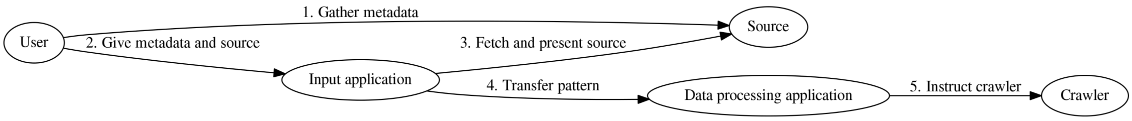
Hyperleap is a small company that is specialized in infotainment (information + entertainment) and administrates several websites which bundle information about entertainment in an ordered way and as complete as possible. Right now, most of the input data is added to the database by by hand which is very labor intensive. Therefore Hyperleap is looking for a smart solution to automate a part of the data injection in the database, the crux however is that the system must not be too complicated from the outside and be useable for a non IT professional(NIP).

1.3 Research question and practical goals

This brings up the main research question: *How can we make an adaptive, autonomous and programmable data mining program that can be set up by a NIP which is able to transform raw data into structured data.*

In practice the goal and aim of the project is to create an application that can, with NIP input, give computer parseable patterns which a separate crawler can periodically crawl. The NIP has to be able to enter the information about the data source in a user friendly interface which sends the information together with the data source to the data processing application. The dataprocessing application then in turn processes the data into a extraction pattern which is sent to the crawler. The crawler can visit sources specified by the NIP accompanied by the extraction pattern created by the dataprocessing application. This workflow is described in graph 1.1.

Figure 1.1: Workflow within the applications



In this way the NIP can train the crawler to periodically crawl different data sources without too much technical knowledge. The main goal of this project is to extract the underlying structure rather then to extract the substructures. The project 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 neven taken out of the experimental phase. The techniques described by Roelofs et al. are more focussed on extracting data from substructures so it can be an addition to the current project.

As a very important sidenote, the crawler needs to notify the administrators if a source has become problematic to crawl, in this way the NIP can very easily retrain the application to fit the latest structural patterns.

1.4 Scientific relevance

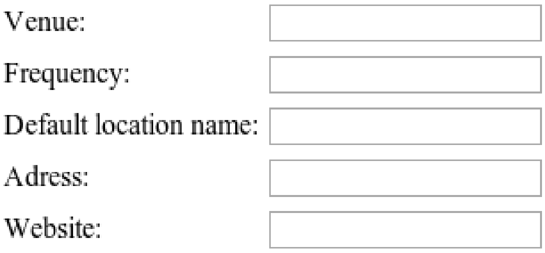
Currently the techniques for conversion from non structured data to structured data are static and mainly only usable by IT specialists. There is a great need of data mining in non structured data because the data within companies and on the internet is piling up and are usually left to catch dust.

2 Methods

2.1 Input application

The user input all goes through the familiar interface of the user's preferred web browser. By visiting the crawler's train website the user can specify the metadata of the source it wants to be periodically crawled through simple web forms as seen in figure 2.1

Figure 2.1: Webforms for source metadata



Venue:

Frequency:

Default location name:

Adress:

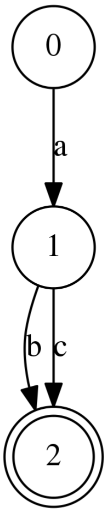
Website:

2.2 Data processing application

2.2.1 Directed acyclic graphs and finiti automata

Directed acyclic graphs(DAG) and finite state automaton(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 2.2 for example shows a FSA that matches on the words *ab* and *ac*.

Figure 2.2: Example DAG/FSA

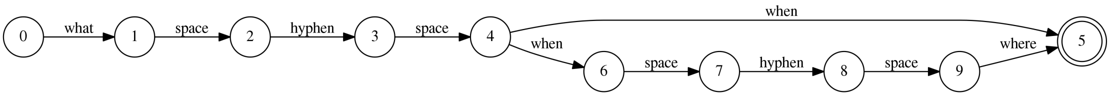


With this FSA we can test if a word fits to the constraints it the FSA describes. And with a little adaptation we can extract dynamic information from semi-structured data.

2.2.2 Back to DAG's and FSA's

Nodes in this data structure can be single letters but also bigger constructions. The example in Figure 2.3 describes different separator pattern for event data with its three component: what, when, where. In this example the nodes with the labels *what, when, where* can also be complete subgrahps. In this way data on a larger level can be using the NIP markings and data within the categories can be processed autonomously.

Figure 2.3: Example event data



2.2.3 Algorithm

2.3 Crawler application

3 Results

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

4 Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

5 Appendices

5.1 Input application

Python front/back-end

```
#!/bin/env python
# −*− coding: utf−8 −*−

from mod_python import apache, util
import feedparser
import re
import urllib
import subprocess
import os

def req_pre_pos(req):
    req. log_error (’handler’)
    req.content_type = ’text/html’
    req.send_http_header()
    args = util.FieldStorage(req)
    req.write(”””\
<html>
␣␣␣␣<head>
␣␣␣␣␣␣␣␣<title>VER:␣0.01␣−␣HyperFrontend␣RSS␣feed␣POSTREQUEST</title>
␣␣␣␣</head>
␣␣␣␣<body>
␣␣␣␣␣␣␣␣Thanks␣submitting:␣<br␣/>
␣␣␣␣␣␣␣␣<a␣href=”index.html”>Enter␣new␣rss␣feed</a>
␣␣␣␣␣␣␣␣<pre>
{}
␣␣␣␣␣␣␣␣</pre>
␣␣␣␣</body>
</html>
”””.format(args))
    os.chdir(’/var/www/py/files’)
    with open(’raw_out.txt’, ’w’) as f:
        f.write(str(args))


def req_pre(req):
    req. log_error (’handler’)
    req.content_type = ’text/html’
    req.send_http_header()
    req.write(”””\
<html>
␣␣␣␣<head>
␣␣␣␣␣␣␣␣<title>HyperFrontend␣RSS␣feed␣input</title>
␣␣␣␣␣␣␣␣<script␣src=”contextmenu_o.js”></script>
␣␣␣␣</head>
␣␣␣␣<body>

␣␣␣␣<table>
```

```
        <tr><td>Venue: </td><td>
            <input type="text" name="venue" class="indat"></td></tr>
        <tr><td>Frequency: </td><td>
            <input type="text" name="freq" class="indat"></td></tr>
        <tr><td>Default location name: </td><td>
            <input type="text" name="dloc" class="indat"></td></tr>
        <tr><td>Adress: </td><td>
            <input type="text" name="adress" class="indat"></td></tr>
        <tr><td>Website: </td><td>
            <input type="text" name="website" class="indat"></td></tr>
    </table>

    Selecteer iets en klik de link<br />
    <button style="color:blue" onclick="javascript:f_waar()">Waar</button>
    <button style="color:green" onclick="javascript:f_wat()">Wat</button>
    <button style="color:red" onclick="javascript:f_wanneer()">Wanneer</button>
    <br />

<div style="position:absolute; left:12px;width:500px;"></div>
<script language="javascript" type="text/javascript">
    var content='<b>Categorize</b><br />';
    content+=' <a href="#" onclick="javascript:f_waar()">Waar</a><br />';
    content+=' <a href="#" onclick="javascript:f_wat()">Wat</a>';
    content+=' <a href="#" onclick="javascript:f_wanneer()>Wanneer</a><br />';
    content+=' Test 123';
    init(content,120);
</script>
""")


def req_post(req):
    req.write("""\
            <button onclick="javascript:post_all()" method="post" target="_bla\
nk">Submit</button>
    </body>
</html>
""")


def feed2html(req, url, name):
    url = urllib.unquote(url)
    url = url if re.match('https?://', url) else 'http://{}'.format(url)
    req.write(
        '\tLoading "{}" as <p id="rssname">"{}"</p><br />\n'.format(url, name))
    feed = feedparser.parse(url)
    channel = feed.feed
    req.write('\t<table id="content-table" border="1" id="htab">\n')
    req.write('\t\t<tr><th>Title</th><th>Summary</th></tr>\n')
    for i in feed.entries:
        req.write('\t\t<tr><td>{}</td><td>{}</td></tr>\n'.
                  format(i['title'].encode('ascii', 'xmlcharrefreplace'),
                         i['summary'].encode('ascii', 'xmlcharrefreplace')))
    req.write('\t</table>\n<br />')


def handler(req):
    if req.method == "POST":
        req_pre_pos(req)
    else:
        req_pre(req)
        args = util.FieldStorage(req)
        if 'url' not in args and 'name' not in args:
            req.write('Something_went_wrong,_empty_fields?<br_/>')
            req.write('<a_href="index.html">back</a>')
        else:
            feed2html(req, args['url'], args['name'])
        req_post(req)
    return apache.OK
```

HTML landing page

```
<html>
    <head>
    </head>
    <body>
        <form method="get" action="./hyper.py">
            <table>
                <tr><td><p>RSS URL: </td><td><input type="text" name="url"
                        value="localhost/py/paradiso.rss.xml"></td></tr>
                <tr><td>RSS Name: </td><td><input type="text" name="name"></td></tr>
                <tr><td><input type="submit" value="Submit"</p>
            </table>
        </form>
    </body>
</html>
```

Javascript frontend

```
var selection;
var mouse_x = 0;
var mouse_y = 0;
var mouse_left = false;
var mouse_right = false;
if (document.addEventListener != undefined) document.addEventListener('mousemove', mouseMove, true);
else if (document.layers) document.captureEvents(Event.MOUSEDOWN | Event.MOUSEMOVE | Event.MOUSEUP);
document.onmousemove = mouseMove;
document.oncontextmenu = RightMouseDown;
document.onmousedown = mouseDown;
document.onmouseup = mouseUp;

function mouseMove(a) {
    mouse_x = document.all ? event.clientX + document.body.scrollLeft : document.layers ? a.x + window.pageXOffset : a.clientX + window.pageXOffset;
    mouse_y = document.all ? event.clientY + document.body.scrollTop : document.layers ? a.y + window.pageYOffset : a.clientY + window.pageYOffset
}

function RightMouseDown() {
    mouse_right = true;
    return false
}

function init(a, w) {
  console.log(a)
    var b = document.createElement("DIV");
    b.id = "contextmenu";
    if (!w) var w = 120;
    b.style.width = w + "px";
    var c = '<div_style="position:relative;left:5px;top:-4px;">';
    c += a;
    c += '</div>';
    b.innerHTML = c;
    b.style.position = "absolute";
    b.style.left = "0px";
    b.style.top = "0px";
    b.style.visibility = "hidden";
    b.style.overflow = "hidden";
    b.style.padding = "4px";
    b.style.backgroundColor = "#ffffff";
    b.style.border = "1px solid #6a6868";
    document.body.appendChild(b);
    delete b
}

function mouseUp(e) {
    var curselection = window.getSelection().getRangeAt(0);
    if (curselection.endOffset - curselection.startOffset > 0)
  selection = curselection;
    console.log(selection)
    if (e.which == 1) document.getElementById("contextmenu").style.visibility = "hidden";
    else if (e.which == 3) mouse_right = false
}

function mouseDown(e) {
    if (e.which == 3) {
        mouse_right = true;
        document.getElementById("contextmenu").style.left = mouse_x + "px";
        document.getElementById("contextmenu").style.top = mouse_y + "px";
        document.getElementById("contextmenu").style.visibility = "visible"
    }
}


function stylizeHighlightedString(range, col)
{
    var selectionContents = range.extractContents();
    var span = document.createElement("span");
    span.appendChild(selectionContents);
    span.setAttribute("class","uiWebviewHighlight");
    span.style.backgroundColor = col;
    span.style.color = "white";
    range.insertNode(span);
}

function f_wanneer() {
    stylizeHighlightedString(selection, "red")
}

function f_wat() {
    stylizeHighlightedString(selection, "green")
}

function f_waar() {
    stylizeHighlightedString(selection, "blue")
}

function post_all() {
    var xmlhttp = new XMLHttpRequest();
    xmlhttp.onreadystatechange=function()
    {
        if (xmlhttp.readyState==4)
        {
            document.write(xmlhttp.responseText);
        }
    }
    var params = ”content=”+encodeURIComponent(document.getElementById(”content−table”).innerHTML);
    params += ”&name=”+encodeURIComponent(document.getElementById(”rssname”).innerHTML);
    var indatarray = document.getElementsByClassName(’indat’)
    for (var i = 0; i<indatarray.length; i++) {
        params += ”&” + indatarray[i].name + ”=” + indatarray[i].value;
    }
    xmlhttp.open(”POST”, ”hyper.py”, true);
    xmlhttp.setRequestHeader(”Content−type”, ”application/x−www−form−urlencoded”);
    xmlhttp.setRequestHeader(”Content−length”, params.length);
    xmlhttp.send(params);
}
```