Unplagged Developers Manual – Part 2

Building the Plagiarism Detection Cockpit

Term paper for the master project I

Mentoring Teacher: Prof. Dr. Debora Weber-Wulff

Department Economics II

HTW Berlin – University of Applied Sciences

Elsa Mahari (s0534556) <Elsa.Mahari@gmx.de>

Dominik Horb (s0534217) <horb@htw-berlin.de>

Tien Nguyen (s0512510) <s0512510@htw-berlin.de>

Benjamin Oertel (s0522720) <contact@benjaminoertel.com>

Heiko Stammel (s0534218) <heiko.stammel@googlemail.com>

Contents

List of Figures

1. Introduction and Overview

@dominik

2. Features

2.1. Notifications and Comments

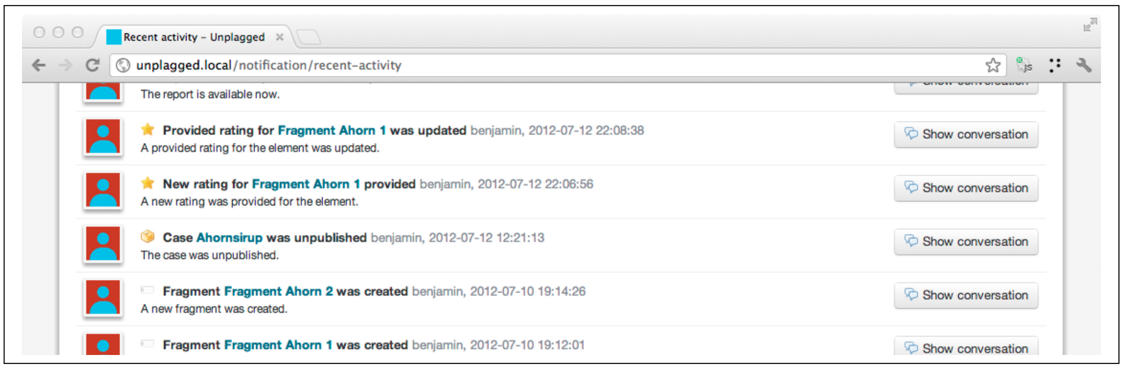
The following section describes an important part for registered users, the notification system and the opportunity for adding comments basically on any resource in the system.

2.1.1. Recent activity stream

It displays the most recent events in the portal in the order they happened. Each activity (figure 2.1), one line in the stream, consists of 3 parts: meta information about the activity itself, information about the initiator and comments. The content of each of these parts is being determined automatically, when a new notification is being persisted to the database.

Besides the initiator information and comments, each activity has an own title, description and icon. These meta information texts can be edited in the portal in the 'Administration' > 'Acions' section. Although it is not possible to remove them, because the action references are being used hard-coded within the system. That means a set of notifications is provided and can be ouput anywhere in the workflow.

The following code snippet shows, how to create a new notification when a new automatic plagiarism detection report was created. The static method takes in 3 parameters: a unique name for the notification type, the content object related to the notification and a user object, as the third parameter. A list of all available notification types can either


Recent activity – Unplagged

unplagged.local/notification/recent-activity

The report is available now.

Provided rating for Fragment Ahorn 1 was updated benjamin, 2012-07-12 22:08:38
A provided rating for the element was updated.
Show conversation

New rating for Fragment Ahorn 1 provided benjamin, 2012-07-12 22:06:56
A new rating was provided for the element.
Show conversation

Case Ahornsirup was unpublished benjamin, 2012-07-12 12:21:13
The case was unpublished.
Show conversation

Fragment Fragment Ahorn 2 was created benjamin, 2012-07-10 19:14:26
A new fragment was created.
Show conversation

Fragment Fragment Ahorn 1 was created benjamin, 2012-07-10 19:12:01
Show conversation

Figure 2.1.: Single activity in the activity stream

be found in the previousely mentioned actions section or in the scripts/build/initdb.php file, where all notifcation types are being declared.

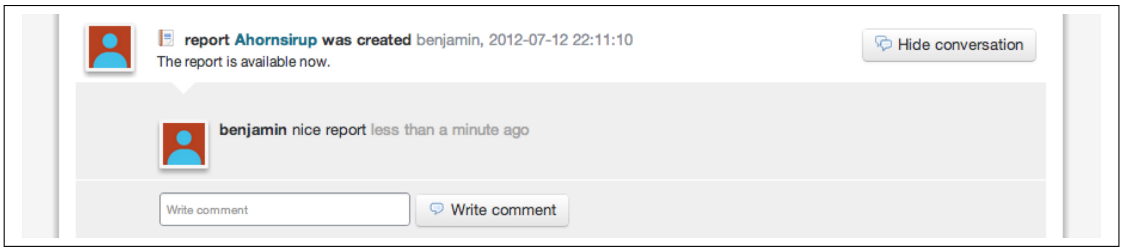
Listing 2.1: Creating a notification for a created report

```
Unplagged_Helper::notify("detection_report_created", $report,
    $report->getUser());
```

Unplagged does have an extensive role and permission management. Therefore the grant on each resource is being verified, before it is being displayed in the activity stream. Usually the resource related to the notification is the resource where the permission check is being performed on. Although in some cases, e.g. when rating a fragment, the resource will be the rating itself, but the permission check is done on the fragment. In this case the notify-method is being called with a fourth, optional parameter, another resource, in this case the fragment. When the user has access on the fragment, all ratings can be accessed as well, automatically.

2.1.2. Comments plugin

Comments are simply a small text related to a specific user and a resource. They can be used to share ideas on an object collaboratively. The most prominent part at Unplagged, where comments are being used, is the activity stream. A comment can be added by any user having access to the notification.


report Ahornsirup was created benjamin, 2012-07-12 22:11:10
The report is available now.

benjamin nice report less than a minute ago

Figure 2.2.: Creating a comment on a resource

For providing a better workflow to the user, the comments can be refreshed and added in place. That means, the position where the user scrolled to in the browser does not get affected. The in-place refreshing is being realized through AJAX. The comments container is being loaded empty and displaying a small loading image only. Not before the user clicks the 'show conversation' button, the comments are being fetched through a post request to the server. Whenever the result is being fetched completely, the spinner graphic is being hidden and the comments are being appended. The parsing of the comments markup is being done in Javascript as well. So the server requests are kept small and the server can return JSON only without any HTML.

Listing 2.2: Refreshing the comments of a resource

```
target.show();
    conversation.hide();
    loading.slideDown(800, function() {
      // get the whole conversation
      $.post('/notification/conversation', {
        'source': sourceId
      }, function(data) {
        if(!data.errorcode) {
          conversation.html("");
          $.each(data, function(index, value) {
            conversation.append(renderConversation(value));
          });
          loading.slideUp(800, function() {
            conversation.slideDown(300);
          });
        } else {
          conversation.html('<div class="comment">' + data.
              message + '</div>');
          loading.slideUp(800, function() {
            conversation.slideDown(300);
          });
        }
      }, "json");
```

Listing 2.3: Creating the markup of a single comment

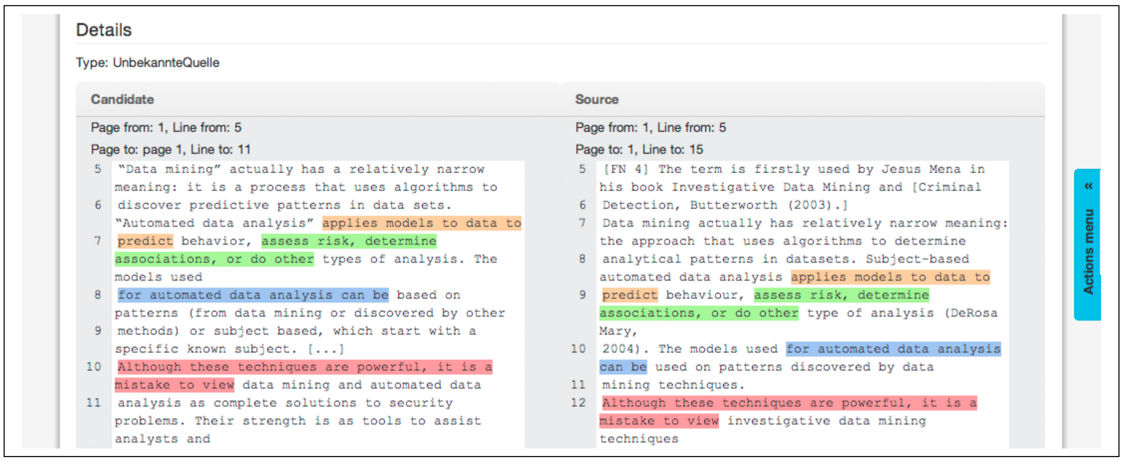
```
function renderConversation(data, target) {
    var tpl;

    switch(data.type) {
      case 'comment':
        tpl = '<div class="comment">' +
        '<div class="image"><img class="avatar-small" src="' +
            data.author.avatar + '" /></div>' +
        '<div class="details">' +
        '<div class="title"><b>' + data.author.username + '</b
            > ' + data.text +
        ' <span class="date">' + data.created.humanTiming +
            '</span>' +
        '</div>' +
        '</div>' +
        '</div>';
        break;
    }
    if(!target) {
      return tpl;
    } else {
      target.append(tpl);
    }
  }
```

2.2. Fragments

Fragments are the part of the application where found text passages that are plagiairism or potential plagiarism are being documented.

A single fragment contains a candidate and a source document. Each of the two documents is being saved with a starting position (page number / line number combination) and an ending position. These two positions can be used to determine exactly the text being involved in a fragment. To visualize this, the figure 2.3) below shows a sample fragment.



Details

Type: UnbekannteQuelle

Candidate

Page from: 1, Line from: 5

Page to: page 1, Line to: 11

```
5  "Data mining" actually has a relatively narrow
   meaning: it is a process that uses algorithms to
6  discover predictive patterns in data sets.
   "Automated data analysis" applies models to data to
7  predict behavior, assess risk, determine
   associations, or do other types of analysis. The
   models used
8  for automated data analysis can be based on
   patterns (from data mining or discovered by other
9  methods) or subject based, which start with a
   specific known subject. [...]
10 Although these techniques are powerful, it is a
   mistake to view data mining and automated data
11 analysis as complete solutions to security
   problems. Their strength is as tools to assist
   analysts and
```

Source

Page from: 1, Line from: 5

Page to: 1, Line to: 15

```
5  [FN 4] The term is firstly used by Jesus Mena in
   his book Investigative Data Mining and [Criminal
6  Detection, Butterworth (2003).]
7  Data mining actually has relatively narrow meaning:
   the approach that uses algorithms to determine
8  analytical patterns in datasets. Subject-based
   automated data analysis applies models to data to
9  predict behaviour, assess risk, determine
   associations, or do other type of analysis (DeRosa
   Mary,
10 2004). The models used for automated data analysis
   can be used on patterns discovered by data
11 mining techniques.
12 Although these techniques are powerful, it is a
   mistake to view investigative data mining
   techniques
```

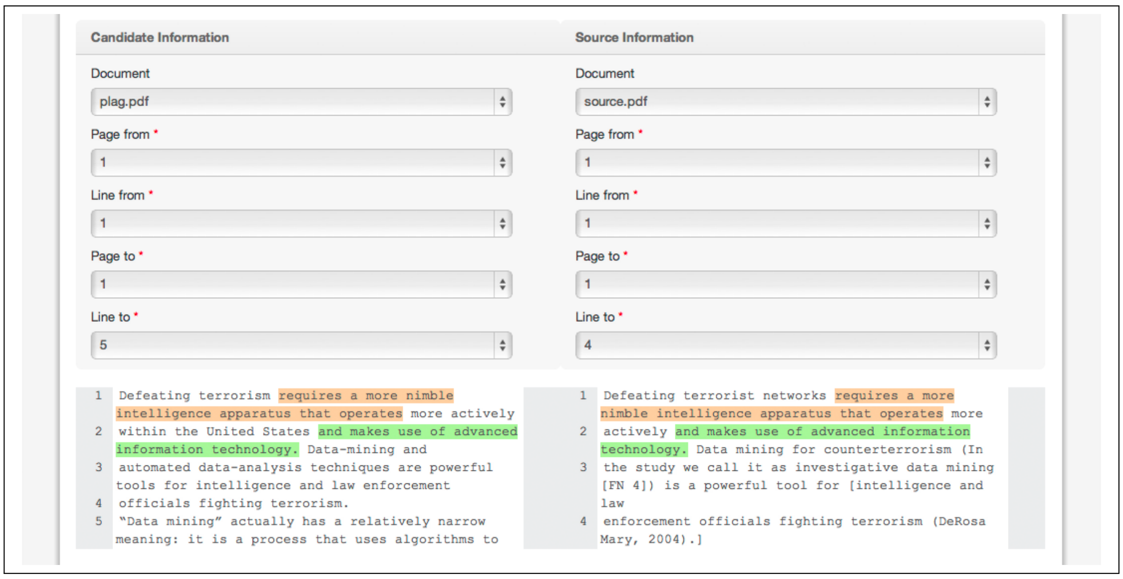
Figure 2.3.: Single fragment with highlighted similarities

2.2.1. Creating a fragment

Such a fragment can be created in two ways, one for people that like using the mouse and another one that can be used with the keyboard only.

The old-fashioned way

The basic way, which can be accessed through the keyboard only, offers a two-column form to the user where the source and potential plagiarism information can be selected by hand. Once a page or line number is being changed, the text shown below is being updated instantly through AJAX and the similarities are highlighted automatically. Although the big disadvantage of this method is, that the whole page is never displayed and the user actually has to guess where the starting and ending point of the fragment in the text really is. Therefore the values of the line from and line to fields have to be increased or decreased by hand, until they are adjusted properly.


Candidate Information

Document: plag.pdf
Page from *: 1
Line from *: 1
Page to *: 1
Line to *: 5

Source Information

Document: source.pdf
Page from *: 1
Line from *: 1
Page to *: 1
Line to *: 4

```
1  Defeating terrorism requires a more nimble
   intelligence apparatus that operates more actively
2  within the United States and makes use of advanced
   information technology. Data-mining and
3  automated data-analysis techniques are powerful
   tools for intelligence and law enforcement
4  officials fighting terrorism.
5  "Data mining" actually has a relatively narrow
   meaning: it is a process that uses algorithms to
```

```
1  Defeating terrorist networks requires a more
   nimble intelligence apparatus that operates more
2  actively and makes use of advanced information
   technology. Data mining for counterterrorism (In
3  the study we call it as investigative data mining
   [FN 4]) is a powerful tool for [intelligence and
   law
4  enforcement officials fighting terrorism (DeRosa
   Mary, 2004).]
```

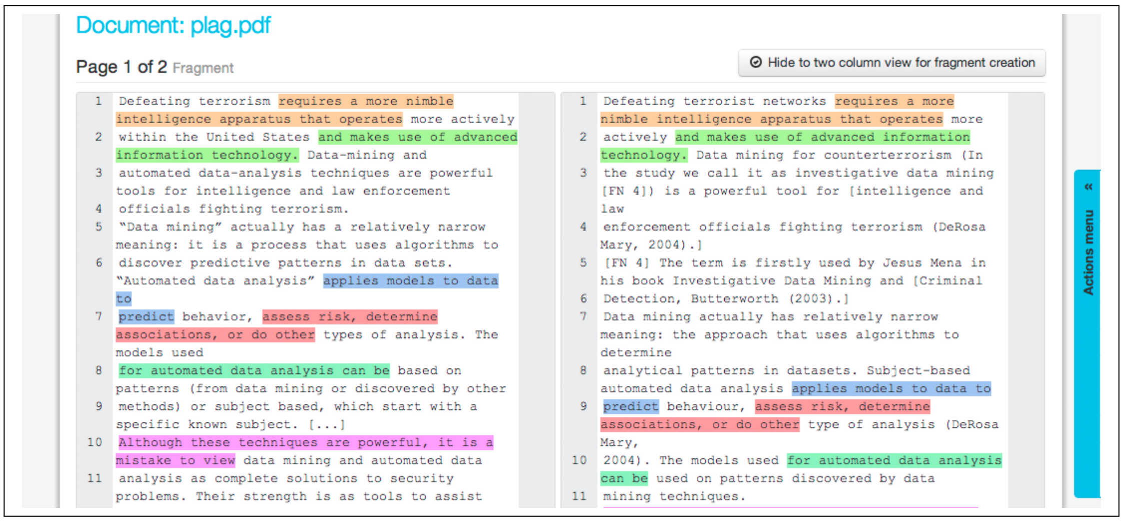
Figure 2.4.: Form for creating a fragment by hand

A more comfortable workflow

Wouldn't it be cool to select text by just marking it with the mouse and having this previousely described form being filled out automatically? We though it would, so we implemented it.

The user has to go to the document being inspected in the current case, select a page to start with and then hit the button 'Switch to two-column view for fragment creation'. At this point a second document can be selected on the right side and the similarities in both texts are once again being highlighted. (figure 2.5) In this two-column view it is also possible to iterate through the pages of the left-side or right-side document to compare page 1 from the left with page 2 from the right and page 1 on the left with page 3 on the right just by one click.

When there are sufficient similarities in an area of the page, a fragment can be created by marking the text, then making a click with the right mouse key to open the context


Document: plag.pdf

Page 1 of 2 Fragment

Hide to two column view for fragment creation

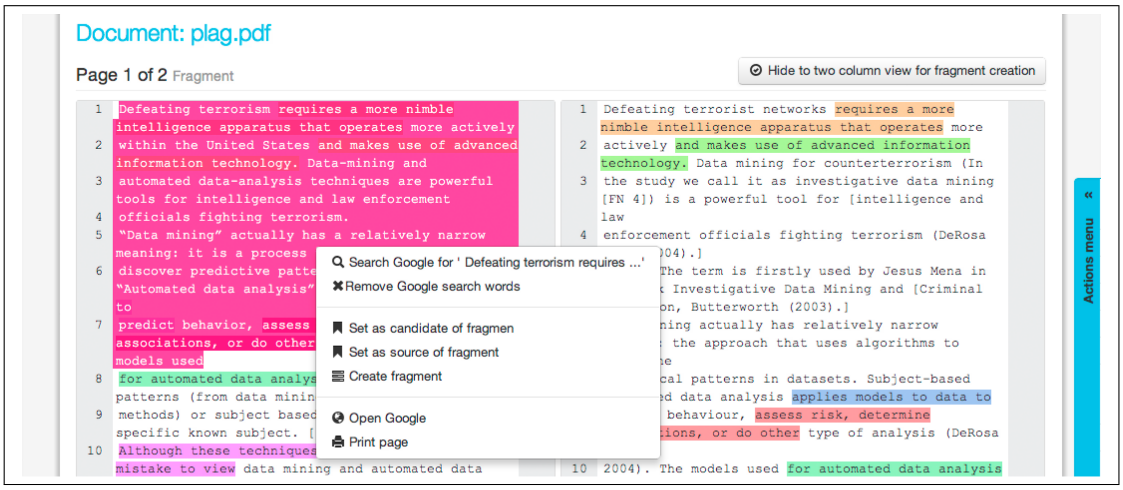
```
1  Defeating terrorism requires a more nimble
   intelligence apparatus that operates more actively
2  within the United States and makes use of advanced
   information technology. Data-mining and
3  automated data-analysis techniques are powerful
   tools for intelligence and law enforcement
4  officials fighting terrorism.
5  "Data mining" actually has a relatively narrow
   meaning: it is a process that uses algorithms to
6  discover predictive patterns in data sets.
   "Automated data analysis" applies models to data
   to
7  predict behavior, assess risk, determine
   associations, or do other types of analysis. The
   models used
8  for automated data analysis can be based on
   patterns (from data mining or discovered by other
9  methods) or subject based, which start with a
   specific known subject. [...]
10 Although these techniques are powerful, it is a
   mistake to view data mining and automated data
11 analysis as complete solutions to security
   problems. Their strength is as tools to assist
```

```
1  Defeating terrorist networks requires a more
   nimble intelligence apparatus that operates more
2  actively and makes use of advanced information
   technology. Data mining for counterterrorism (In
3  the study we call it as investigative data mining
   [FN 4]) is a powerful tool for [intelligence and
   law
4  enforcement officials fighting terrorism (DeRosa
   Mary, 2004).]
5  [FN 4] The term is firstly used by Jesus Mena in
   his book Investigative Data Mining and [Criminal
6  Detection, Butterworth (2003).]
7  Data mining actually has relatively narrow
   meaning: the approach that uses algorithms to
   determine
8  analytical patterns in datasets. Subject-based
   automated data analysis applies models to data to
9  predict behaviour, assess risk, determine
   associations, or do other type of analysis (DeRosa
   Mary,
10 2004). The models used for automated data analysis
   can be used on patterns discovered by data
11 mining techniques.
```

Figure 2.5.: Creating a fragment the modern way - Step 1

menu and 'set as candidate/source of fragment'. This stores the marked text temporarily until the 'create fragment' button in the context menu is being pressed (figure 2.6). The selection of the 'create fragment' button opens the same form as described in the section before and pre-fills it with the start and end values for page and line for both sides automatically.

This techchnique makes it much easier to create a fragment, since the text can be seen in the context of the whole page, before it is added to the fragment form.


Document: plag.pdf

Page 1 of 2 Fragment

Hide to two column view for fragment creation

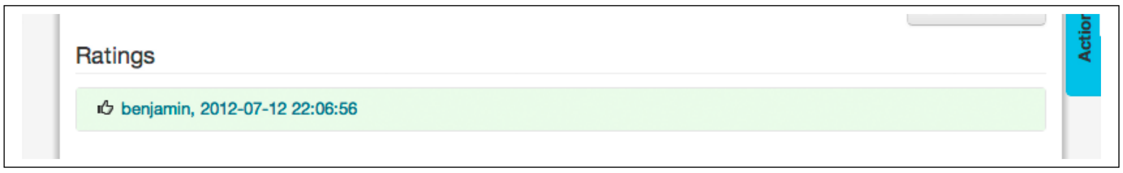
Search Google for ' Defeating terrorism requires ...'
Remove Google search words
Set as candidate of fragmen
Set as source of fragment
Create fragment
Open Google
Print page
Actions menu

Figure 2.6.: Creating a fragment the modern way - Step 2

2.2.2. Rating a fragment

A created fragment has to be verified by other collaborators of the case in order to be approved for containing plagiarism. This process is described in the current section.

A rating contains information about the user who made the rating, a flag whether it approves or declines the fragment and an optional property that can contain a description, why the user gave the rating. Each fragment can be approved by a user only once. However the reason and rating type can be changed by the initiator at any time, until the fragment is approved by a certain amount of people. The amount of ratings to lock a fragment and its ratings for further edits is defined in the case administration form. Whenever this amount is reached, the fragment gets locked automatically and can be unlocked by administrators only.


Ratings

benjamin, 2012-07-12 22:06:56

Figure 2.7.: List of fragment ratings

2.3. User avatar

@elsa

2.3.1. Avatar cropping

While the previous section described how the user avatar can be set, this one explains how the cropping of images that are not in the correct aspect ratio is being done. All avatars have to be in a square aspect ratio. However, not all users have to skills to provide an image that meets these requirements. So we decided to crop uploaded images into the appropriate format as they are being uploaded.

What we do is taking the shortest of the two rectangle sides and cut a partial from the center of the longer one of the two sides, which has the same length as the shortest side. To make it more visual, what that means, figure 2.8 shows an example. The red area is the part that will be cropped.

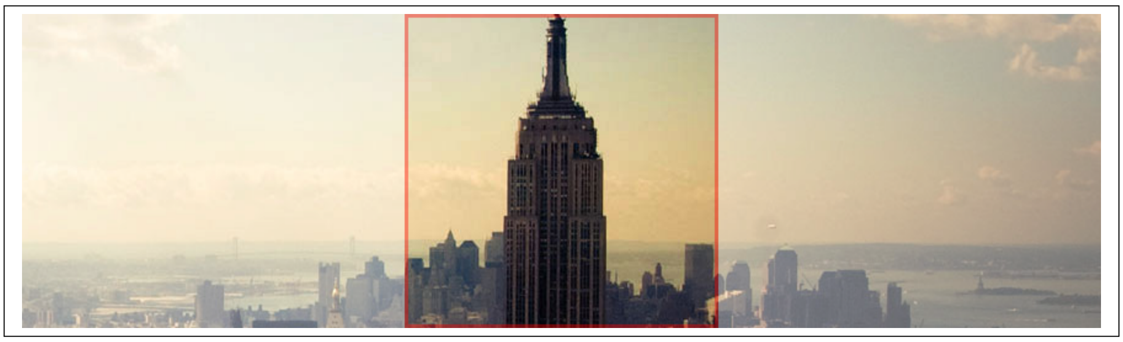


Figure 2.8.: Cropping avatar

The cropped image then is being scaled to the needed width and height, currently 50 x 50 pixels. The algorithm for cropping that has been developed, is shown in the following code snippet.

Listing 2.4: Cropping an image to a square aspect ratio

```php
public function crop($thumbWidth, $thumbHeight){
    //getting the image dimensions
    list($width, $height) = getimagesize($this->file->
        getFullPath());

    //saving the image into memory (for manipulation with GD
        Library)
    $myImage = imagecreatefromjpeg($this->file->getFullPath())
        ;

    // setting the crop size
    if($width < $height) {
      $twidth = $width;
      $theight = $width;
      $x = 0;
      $y = $height / 2. - $width / 2.;
    } else {
      $twidth = $height;
      $theight = $height;
      $x = $width / 2. - $height / 2.;
      $y = 0;
    }

    // creating the thumbnail
    $thumb = imagecreatetruecolor($thumbWidth, $thumbHeight);
    imagecopyresampled($thumb, $myImage, 0, 0, $x, $y,
        $thumbWidth, $thumbHeight, $twidth, $theight);

    imagejpeg($thumb, $this->file->getFullPath());

    imagedestroy($thumb);
    imagedestroy($myImage);

    return $this->file;
  }
```

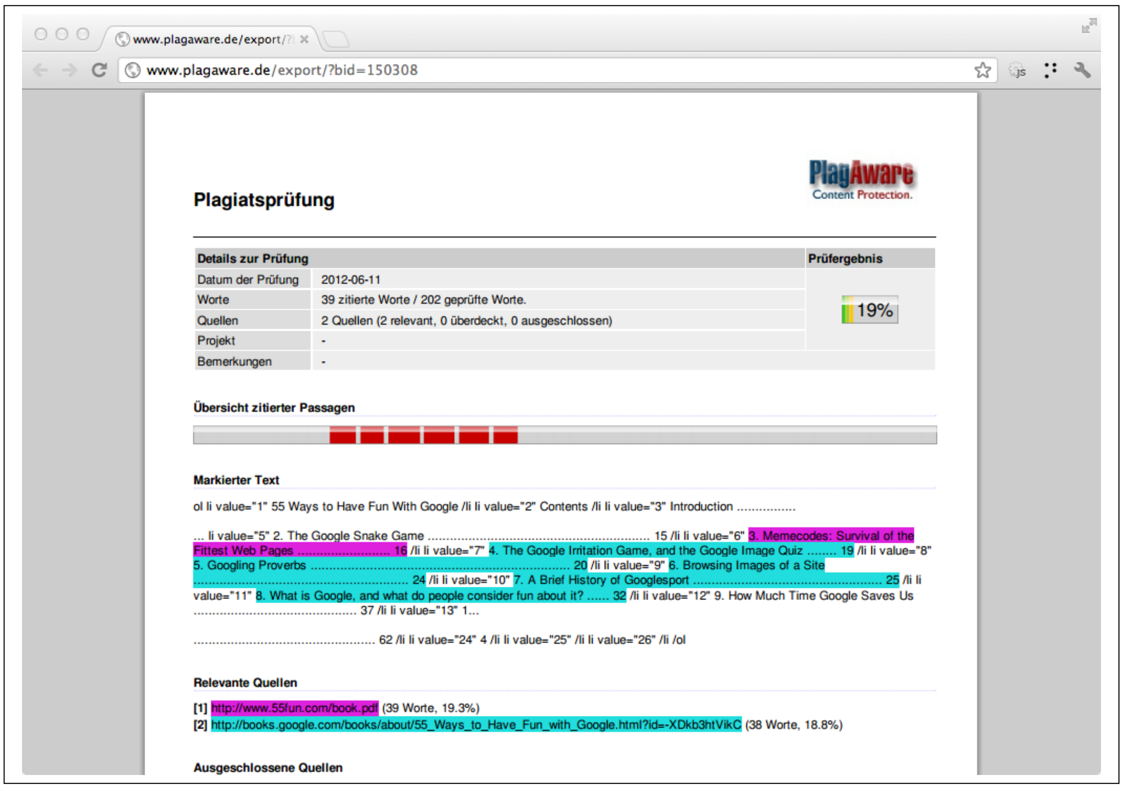
2.4. Automatic Plagiarism Detection Webservices

Unplagged itself is a workbench, where the plagiairism detection is done by hand. Although it provides useful automatic tools that help the user to make the process of plagiairism detection easier. One of these tools are external webservices that check a specific text for plagiairism and try to find sources which can be used for further inspections.

We were talking to 3 companies offering such a webservice: Docoloc, PlagScan and PlagAware. As a proof of concept the PlagAware webservice has been implemented and added to our application, it will be described in the following section.

2.4.1. PlagAware

PlagAware, a website for automatic plagairism detection is a commercial website which costs money depending on the amount of text to analyze. It figures out possible sources of the text handed in on their website and creates a PDF report with a list of all the found sources (figure 2.9).


www.plagaware.de/export/?bid=150308

Plagiatsprüfung

PlagAware
Content Protection.

Details zur Prüfung		Prüfergebnis
Datum der Prüfung	2012-06-11	19%
Worte	39 zitierte Worte / 202 geprüfte Worte.	
Quellen	2 Quellen (2 relevant, 0 überdeckt, 0 ausgeschlossen)	
Projekt	-	
Bemerkungen	-	

Übersicht zitierter Passagen

Markierter Text

ol li value="1" 55 Ways to Have Fun With Google /li li value="2" Contents /li li value="3" Introduction ....

... li value="5" 2. The Google Snake Game .... 15 /li li value="6" 3. Memecodes: Survival of the Fittest Web Pages .... 16 /li li value="7" 4. The Google Irritation Game, and the Google Image Quiz .... 19 /li li value="8" 5. Googling Proverbs .... 20 /li li value="9" 6. Browsing Images of a Site .... 24 /li li value="10" 7. A Brief History of Googlesport .... 25 /li li value="11" 8. What is Google, and what do people consider fun about it? ...... 32 /li li value="12" 9. How Much Time Google Saves Us .... 37 /li li value="13" 1...

.... 62 /li li value="24" 4 /li li value="25" /li li value="26" /li /ol

Relevante Quellen

[1] http://www.55fun.com/book.pdf (39 Worte, 19.3%)
[2] http://books.google.com/books/about/55_Ways_to_Have_Fun_with_Google.html?id=-XDkb3htVikC (38 Worte, 18.8%)

Ausgeschlossene Quellen

Figure 2.9.: PlagAware result document

The website also offers a webservice which takes text as input and notifies the user when the analyzing is finished. Although it does not provide any information about the sources through the webservice response call, there is only a status code and the percentage of plagairism responded. Whenever the response was sent, the user has to go to the PlagAware website and check out the detailed results there.

Since the PlagAware webservice is a simple HTTP-Webservice, the connection is done through an HTTP request with curl.

Listing 2.5: Sending a request through curl to the PlagAware webservice

```
public function detect(
    Application_Model_Document_Page_DetectionReport &$report){
     $url = "http://www.plagaware.de/service/submittext";
     $fields = array(
       'UserCode'=>urlencode($this->paUserCode),
       'ResultUrl'=>urlencode($this->paResultUrl . $report->
           getId()),
       'TestText'=>urlencode($report->getPage()->getContent()),
       'DryRun'=>urlencode($this->paDryRun)
     );

     // url-ify the data for the POST
     $fields_string = "";
     foreach($fields as $key=>$value){
       $fields_string .= $key . '=' . $value . '&';
     }
     rtrim($fields_string, '&');

     $ch = curl_init();
     curl_setopt($ch, CURLOPT_URL, $url);
     curl_setopt($ch, CURLOPT_RETURNTRANSFER, 1);
     curl_setopt($ch, CURLOPT_CONNECTTIMEOUT, 10);
     curl_setopt($ch, CURLOPT_POST, count($fields));
     curl_setopt($ch, CURLOPT_POSTFIELDS, $fields_string);

     $output = curl_exec($ch);
     $info = curl_getinfo($ch);
     curl_close($ch);
}
```

The response is sent through a GET-Request to a pre-defined request URL, in our case '/document/response-plagiarism/report/<report-id>'. The call of this action stores the result in our database and creates a notifcation for the user, indicating that a new report is available.

2.5. Permission and role management

@dominik

- types of permission, types of roles - collaborators - new role created from global case role

2.6. Barcode

The barcode is a visual representation of the amount of plagairism in each page of the target document. In our application it is used in two representations, which use the same data source but have another visual layout and details shown within.

The barcode has 4 different colors, which represent another amount of plagairism within the page:

- light blue: the page is disabled for the barcode
- white: page not available or no plagiarism
- black: more than 0 percent plagiarized
- dark red: more than percent plagairized
- light red: more than percent plagairized

Representation with labels on the home page

The first representation can be found at the home page, here are the barcodes of all published cases being displayed. The barcode is being displayed with page numbers below and the width is dynamic, so the barcode increases when the page gets wider and the barcode gets smaller, when the page width decreases.

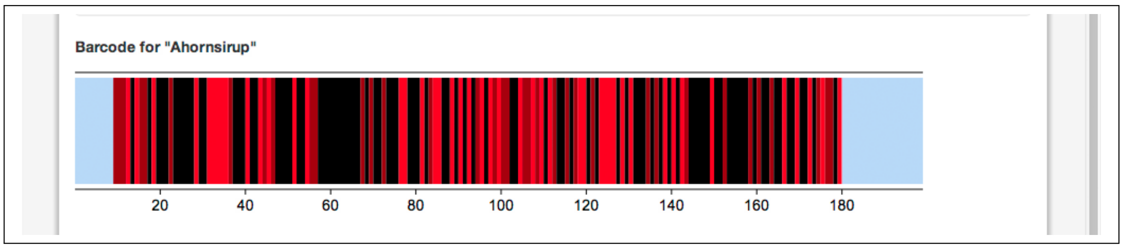


Figure 2.10.: Barcode on home page

Since this mechanism is very dynamic, an algorithm needed to be developed to calculate the x axis values taking into consideration the page width and the number of pages in the document being visualized. It assumes that a label needs at least 45px of width and the whole barcode has to look well to a width of 500px.

So what we do is, we start with a setpsize of 10 for the page numbers: 10 20 30 40 50, now we check if the amount of pages can be displayed with such a fine scale to stay in the 500px range. For example a page count of 200 leeds to 20 values on the x axis and if we assume 45px are necessary to display a value on the axis properly, we would need 900px of width, but we have only 500px available. Otherwise we would have crossovers when we get below 900px. So the stepsize is increased by 10, until we stay in the 500px maximum width. As it turns out, a step size of 20 pages is sufficent to get a width of 450px in the end. The code for this calculation is being shown below.

Listing 2.6: Generating the barcode x axis

```
private function generateAxis() {
        $labelStepsize = 10;

        while (true) {
            $count = sizeof($this->pages);
            $labelCount = floor($count / $labelStepsize);

            // we assume a label needs 45px and 500px is the
               width that needs to be displayable without
               crossovers
            if (45 * $labelCount > 500) {
                $labelStepsize += 10;
                continue;
            }
```

```
                break;
            }

            $label = $labelStepsize;
            $x = 0;
            while ($x < ($this->width - ($this->initWidth *
                $labelStepsize))) {
                $x += ($this->initWidth * $labelStepsize);
                $this->result .= '<text x="' . $x . $this->
                    widthUnit . '" y="' . $this->y . '" font-family
                    ="Arial" font-size="14" text-anchor="middle">'
                    . $label . '</text>';
                $this->result .= '<line x1="' . $x . $this->
                    widthUnit . '" y1="' . ($this->y - 20) . '" x2=
                    "' . $x . $this->widthUnit . '" y2="' . ($this
                    ->y - 15) . '" stroke-width="1" stroke="#000000
                    "></line>';

                $label += $labelStepsize;
            }
        }
```

Representation without labels in the report

The second area where barcodes are used, are the final reports. They will be explained in the next section. The space available in the PDF is much less than on the website and the library we are using for generating the PDF report does not support text in scable vectors graphics, so we decided to display the barcodes without labels for now. However the data visualized is exactly the same.

3. Showtime

4. Summary and Outlook

summary and outlook

A. Meetings

The following tables show the minutes of most of the team meetings.

Bibliography

[Google 2012] GOOGLE: *News Search Interest: plagiat.* http://www.google.com/insights/search/#q=Plagiat&geo=DE&date=1%2F2011%2013m&gprop=news&cmpt=q, 2012. – [Online; accessed 07-February-2012]