Karlsruher Institut für Technologie

Referenzsuche und Referenzgrapherstellung

Seminar Anwendungen von Semantic MediaWiki

Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB)

Agenda

1. Ausgangslage
    - 1.1 Aufgabenstellung
    - 1.2 Spezifikation der Erweiterung
2. Konzept
    - 2.1 Konkreter Ablauf
    - 2.2 Metadatenextraktion
3. Umsetzung
    - 3.1 Vorstellung von „ReferenceHelper“
    - 3.2 Aufbau der Erweiterung
4. Fazit

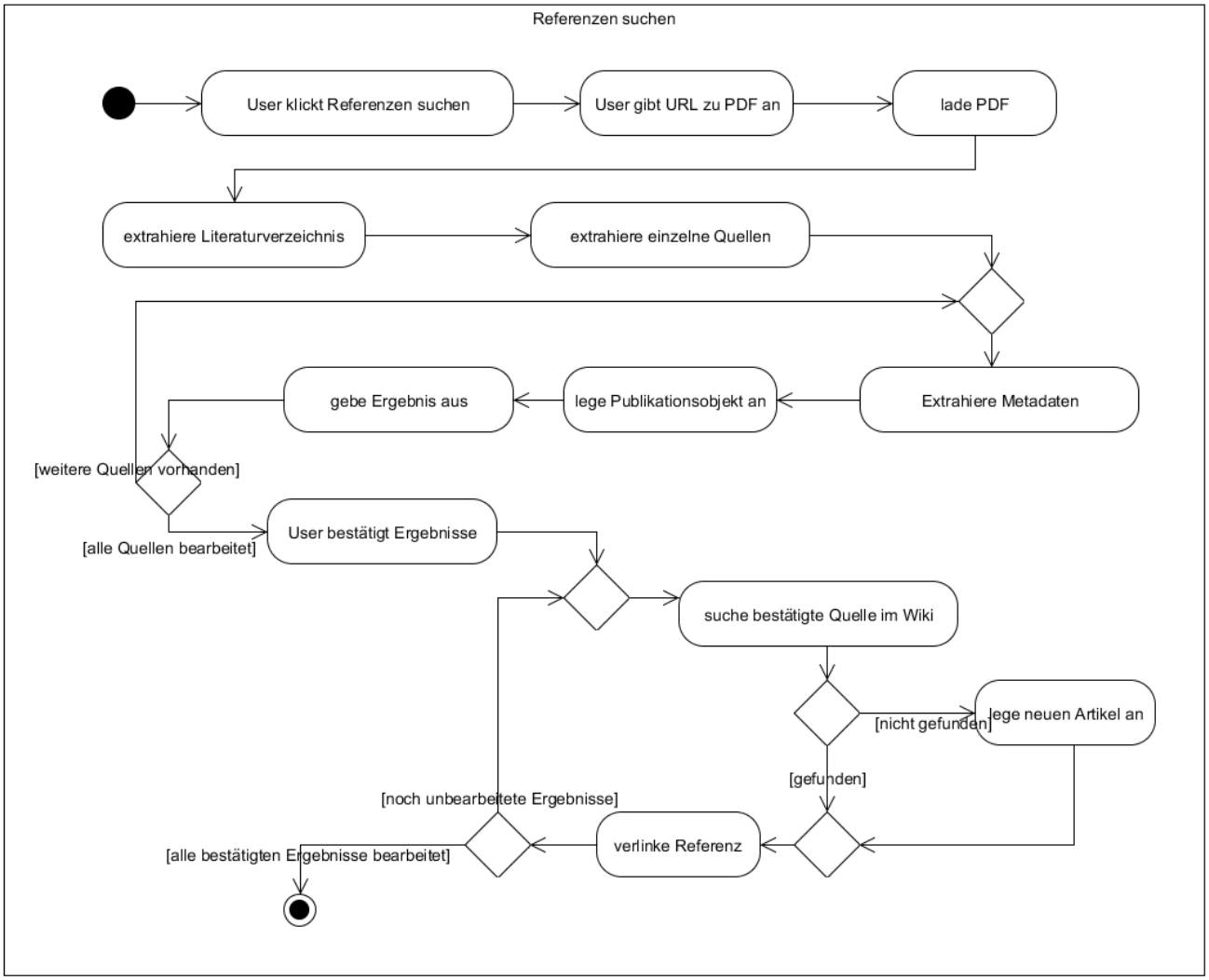
1.1 Aufgabenstellung

- In einem SMW sind Publikationen als Wiki-Seiten (mittels einer Vorlage) gespeichert.
- Die Referenzen aus dem damit verlinkten Publikations-PDF sollen extrahiert werden.
- Gefundene Referenzen sollen entweder auf schon im Wiki vorhandene Publikations-Wiki-Seite verlinken oder es sollen entsprechende Publikations-Wiki-Seiten angelegt werden.
- Darauf aufbauend kann ein Referenzgraph erstellt werden, der die Referenzstruktur widerspiegelt .

1.2 Spezifikation der Erweiterung

- Funktionale Anforderungen an die Erweiterung:
  - /F10/: Einlesen von PDF-Dokumenten.
  - /F20/: Extraktion einzelner Referenzen aus einer PDF.
  - /F30/: Verbinden einer Publikation mit deren Referenzen.
  - /F40/: Anlegen neuer Publikationen aus gefundenen Referenzen.
  - /F50/: Visualisierung der sich aus Referenzen ergebenden Beziehungen zwischen Publikationen.

2.1 Konkreter Ablauf



2.2 Metadatenextraktion

- Extraktion von Informationen (bspw. Autor, Titel, Jahr, Journal) aus einer Literaturangabe
- Informationen sind durch Signalwörter oder Zeichen voneinander getrennt
- Vielfalt verschiedener Zitierstandards und unbeabsichtigte Fehler der Autoren, erschweren den Prozess
- ein sogenanntes „Sequence Labeling Problem“ liegt vor

Zitierweise nach APA-Richtliniene:

Alon, Uri. “How to choose a good scientific problem.” *Molecular cell* 35.6 (2009) : 726-728. Print.

Zitierweise nach MLA:

Alon, U. (2009). How to choose a good scientific problem. *Molecular cell*, *35*(6), 726-728.

2.2 Metadatenextraktion

- „FreeCite“ stellt Lösung dieses Problems bereit
- bietet Service über REST-Schnittstelle an
- XML als Ausgabeformat:

```
<citations>
  <citation valid=true>
  <authors>
    <author>Alon, Uri</author>
  </authors>
  <title>How to choose a good scientific problem</title>
  <journal>Molecular cell</ journal>
  <pages>726-728</pages>
  <year>2009</year>
  …
  </citation>
</citations>
```

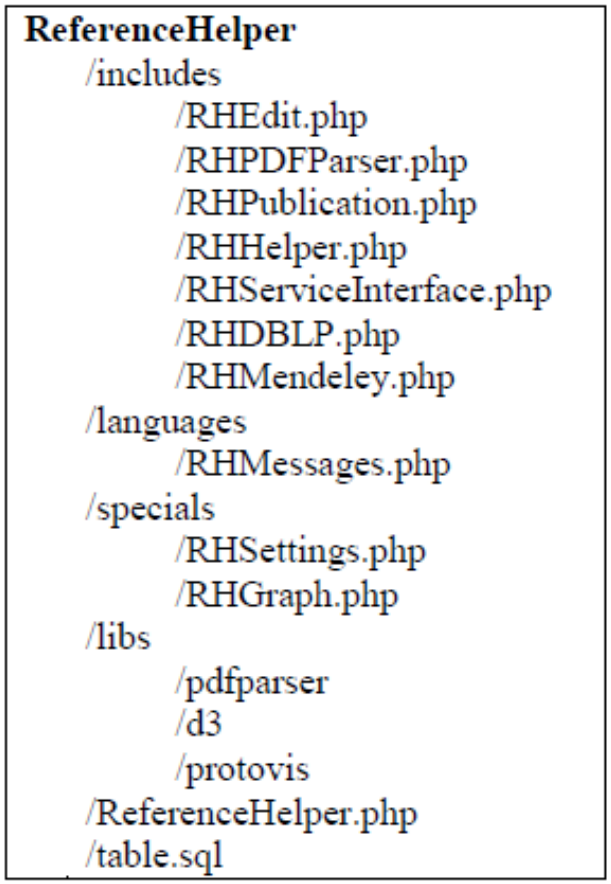
3.1 Vorstellung von „ReferenceHelper“

3.2 Aufbau der Erweiterung

```
ReferenceHelper
    /includes
        /RHEdit.php
        /RHPDFParser.php
        /RHPublication.php
        /RHHelper.php
        /RHServiceInterface.php
        /RHDBLP.php
        /RHMendeley.php
    /languages
        /RHMessages.php
    /specials
        /RHSettings.php
        /RHGraph.php
    /libs
        /pdfparser
        /d3
        /protovis
    /ReferenceHelper.php
    /table.sql
```

- RHEdit steuert den gesamten Prozess des Referenzen Suchens und Speicherns
- RHPDFParser stellt Werkzeuge zum Auslesen des Textes und Finden von Referenzen in PDFs bereit
- repräsentiert Publikationen und stellt Werkzeuge zur Bearbeitung bereit
- RHHelper stellt verschiedene statische Methoden bereit
- RHServiceInterface definiert Methoden zur Anbindung verschiedener Services zur Informationsgewinnung

3.2 Aufbau der Erweiterung



- /languages beinhaltet RHMessages mit allen benötigten Textbaustein-Variablen
- /specials enthält alle Klassen zur Bereitstellung der Wiki-Spezialseiten
- /libs enthält benötigte Bibliotheken

4. Fazit

- funktionale Anforderungen konnten umgesetzt werden
- nicht alle PDF-Formatierungen konnten abgedeckt werden
- Autorennamen nicht einheitlich hinterlegt, kann zu Redundanzen führen
- Erweiterung der verfügbaren Schnittstellen

Danke für eure Aufmerksamkeit.

3.3 Funktionsweise ausgewählter Programmteile

Extrahieren einzelner Publikationen aus Literaturverzeichnis:

- wird von der Klasse RHPDFParser ausgeführt
- zuerst Literaturverzeichnis abtrennen - *getReferenceSection()*
- Literaturverzeichnis nach verwendetem Muster durchsuchen - *getCitationStyle()*

REFERENCES

Aleven, V., & Koedinger, K. R. (2002). An effective meta-cognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26(2), 147-179. hier trennen

Aleven, V., & Koedinger, K. R. (2000). Limitations of Student Control: Do Students Know when they need help? In G. Gauthier, C. Frasson & K. VanLehn (Eds.) *Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000* (pp. 292-303). Berlin: Springer Verlag. hier trennen

Aleven, V., McLaren, B. M., & Koedinger, K. R. (2006). Towards Computer-Based Tutoring of Help-Seeking Skills. In S. Karabenick & R. Newman (Eds.) *Help Seeking in Academic Settings: Goals, Groups, and Contexts* (pp. 259-296). Mahwah, NJ: Erlbaum.

3.3 Funktionsweise ausgewählter Programmteile

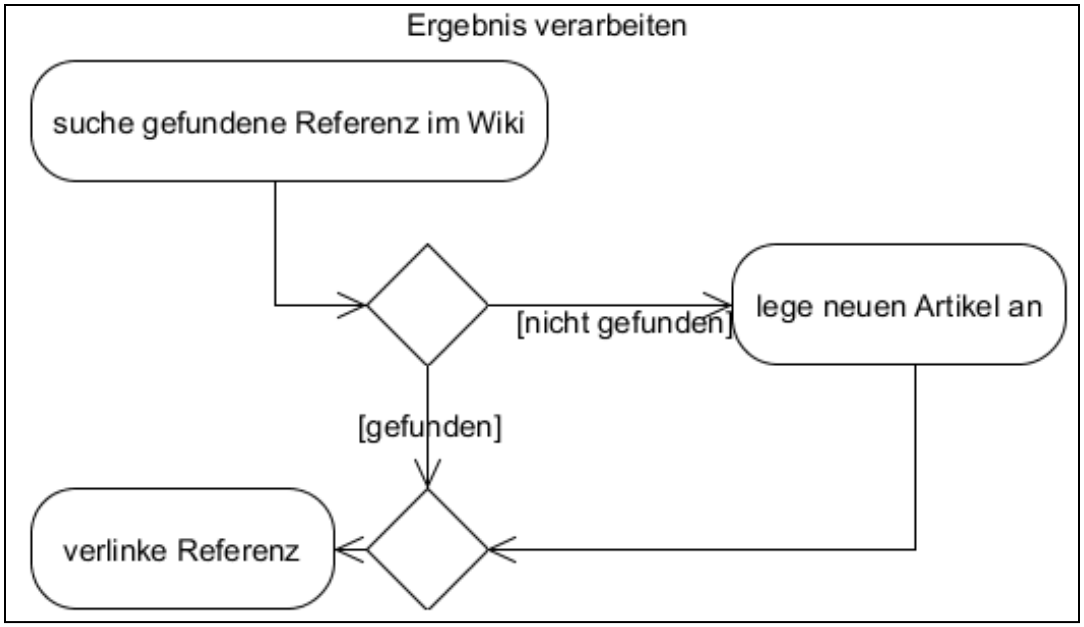
Extrahieren einzelner Publikationen aus Literaturverzeichnis:

```
private function getCitationStyle($references) {
      $styles = array('@\[[0-9]{1,3}\]@' => 0,
            '@\.\n@' => 0,
            '@\. \n@' => 0);
      foreach ($styles as $regex => $quantity) {
            $styles[$regex] = preg_match_all($regex, $references);
      }
      arsort($styles);
      $keys = array_keys($styles);
      return array_shift($keys);
}
```

- Literaturverzeichnis wird auf drei Muster untersucht (Reguläre Ausdrücke)
- Rückgabewert ist der am häufigsten aufgetretene reguläre Ausdruck
- mit *preg_split (*Rückgabewert*)* wird Literaturverzeichnis dann in Array aufgeteilt

3.3 Funktionsweise ausgewählter Programmteile

Verarbeiten der Ergebnisse:



- Prozess wird von der Methode *saveResults()* der Klasse RHEdit gesteuert
- Referenz liegt dabei als RHPublikationsobjekt vor

3.3 Funktionsweise ausgewählter Programmteile

Verarbeiten der Ergebnisse - Suchprozess:

- die in MediaWiki integrierte Suche und deren Funktionen werden verwendet
- gibt Übereinstimmungen mit Seitentiteln zurück
- berücksichtigt dabei Unterschiede in Groß- und Kleinschreibung, sowie der Zeichensetzung
- dazugehörige Methode heißt *getTitle()*
- wird in der MediaWiki-Klasse SearchMySQL bereitgestellt