Aufbau einer Infrastruktur für Information Retrieval-Evaluationen

Von Nils-Peter Schirrmeister und Stefan Keil

Hochschule Darmstadt, Fachbereich Media

Abstract

Das Projekt „Aufbau einer Infrastruktur für Information Retrieval-Evaluationen“ (AIIRE) bietet eine Softwareinfrastruktur zur Unterstützung von Information Retrieval-Evaluationen (IR-Evaluationen). Die Infrastruktur basiert auf einem Tool-Kit, das bei GESIS im Rahmen des DFG-Projekts IRM entwickelt wurde. Ziel ist es, ein System zu bieten, das zur Forschung und Lehre am Fachbereich Media für IR-Evaluationen genutzt werden kann.

Einleitung

Das Projekt AIIRE[1], gefördert durch das Zentrum für Forschung und Entwicklung der Hochschule Darmstadt[2], bietet eine Softwareinfrastruktur zur Unterstützung von IR-Evaluationen.

Die Retrieval-Infrastruktur, die im Projekt eingesetzt wird, ist ein fertig entwickeltes Software-Tool-Kit, das im DFG-Projekt IRM[3] bei GESIS entwickelt wurde, und Open Source zur Verfügung steht. Details zum Tool-Kit und damit durchgeführten Untersuchungen unter anderem in: (Mutschke et al. (2011); Mayr et al. (2011)).

Die Infrastruktur besteht aus einer Suchmaschine (Apache Solr[4]) und mehreren Retrieval-Modulen. Besonders hervorzuheben ist ein Modul, ein sogenanntes Information Retrieval Service Assessment-Tool (IRSA), das themenbezogene (term-basierte) Retrieval-Studien mit einer Bewertungsfunktion unterstützt.

Apache Solr ist eine Suchmaschine, die sich unter anderem durch eine schnelle Volltextsuche auszeichnet, da sie nicht mit einer relationalen Datenbank arbeitet, sondern einen Index generiert. Eine Suchanfrage kann somit über alle (zur Suche zugelassenen) Elemente erfolgen oder durch Anwendung von Suchschlüsseln auf bestimmte Metadatenelemente beschränkt werden. Auf eine Suchanfrage liefert Solr eine XML-Ausgabe, die durch den Solr-PHP-Client in ein für Menschen besser lesbares Format gebracht wird.

---

[1] http://aiire.media.h-da.de/SolrPHP/
[2] http://zfe.h-da.de/
[3] http://www.gesis.org/irm/
[4] http://lucene.apache.org/solr/

Die thematische Breite des Projekts lässt sich in folgenden Arbeitsfeldern, hier in chronologischer Abfolge aufgeführt, zusammenfassen:

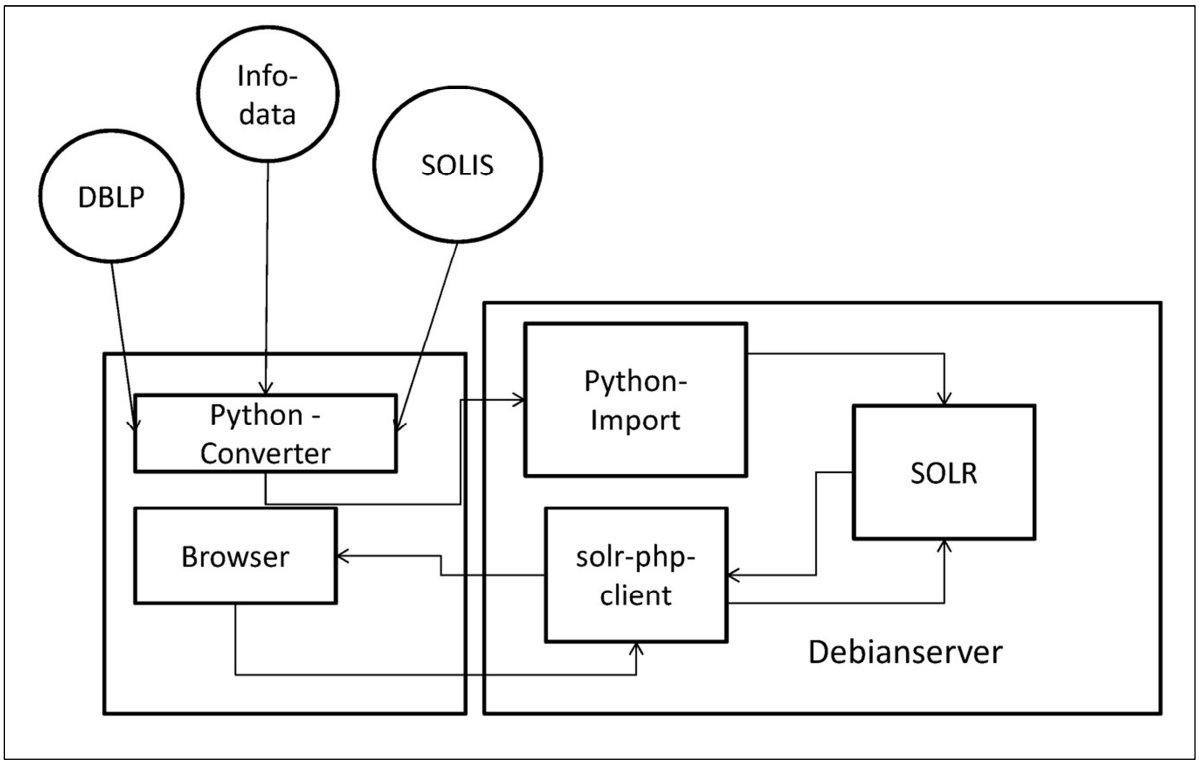
Installation und Bereitstellung der Infrastruktur-Komponenten und Datenbanken
Anpassungen und Customizing für den Fachbereich
Konvertierung und Import der Testkorpora in die Infrastruktur
Erprobung von Verfahren zur Generierung spezifischer Korpora

Aktueller Stand

Zum Zeitpunkt dieser Einreichung befindet sich die Infrastruktur bereits in einem funktionsfähigen Zustand und die Korpora sind erfolgreich importiert. Das System wird noch durch weitere Anpassungen optimiert, während parallel die spezifischen Korpora generiert werden.

Die spezifischen Korpora bestehen aus (Teil-)Mengen der Sammlungen Infodata[5][6], SOLIS[7], und der DBLP[8], die durch Extraktion der Datensätze, die inhaltlich für die Informationswissenschaft relevant sind, erstellt werden.

Datenimport und Systemfunktionalität



[5] http://www.infodata-edepot.de/
[6] Es handelt sich um eine Teilmenge von ca. 17.000 Datensätzen aus den Erfassungsjahren 2000-2005 unter CC-Lizenz
[7] http://www.gesis.org/unser-angebot/recherchieren/solis/
[8] http://www.informatik.uni-trier.de/~ley/db/

Abb. 1 - Schematische Darstellung des Datenimports

Abb. 1 zeigt den Datenimport in schematischer Form. Die bereits genannten Sammlungen (in XML vorliegend) werden mittels eines XSLT-Stylesheets sowie einem Pythonscript auf das AIIRE-Schema angepasst. Daraufhin werden die Daten serverseitig in SOLR eingespielt. Mittels des Solr-PHP-Client Framework werden die Daten für die Nutzer als Websuche zugänglich gemacht.

Abb. 2 stellt ein typisches IR-Evaluierungsszenario schematisch dar. Bezieht man dieses Schema auf das AIIRE-System, so ist anzumerken, dass das „document set“ von SOLR bereitgestellt wird sowie der ganze Prozess des Rankens und der Erstellung des „document pools“ vom IRSA-Modul ausgeführt wird. Die Relevanzbewertungen der „human assesors“ geben daraufhin Aufschluss auf die Qualität der angewandten Ranking-Algorithmen.

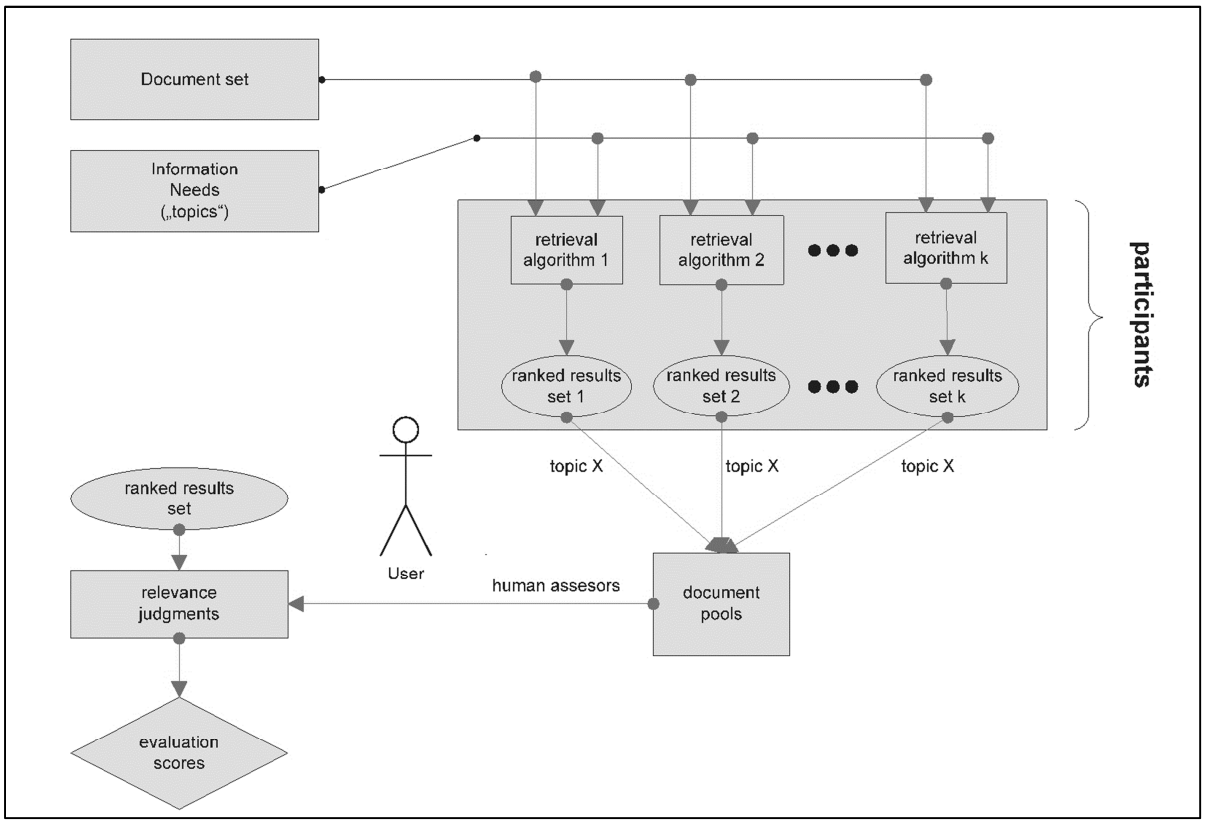


Abb. 2 - Typisches IR-Evaluationsszenario nach dem Verfahren TREC aus Mayr (2010, S. 101)

Ausblick

Der Mehrwert der entwickelten Infrastruktur liegt in der Möglichkeit gezielt verschiedene Aspekte des IR zu evaluieren. Neben der Möglichkeit das System in der Lehre anzuwenden und damit spezifische Sachverhalte des IR zu verdeutlichen, können auch Such- und Rankingalgorithmen evaluiert werden. Neben dieser Funktionalität ist es denkbar auch noch weitere Evaluationen anhand des Systems vorzunehmen.

Desweiteren können am Informationssystem Usability-Studien an Weboberflächen und grafischen Zugängen durchgeführt werden. Hierfür würden sich unter anderem Logfileanalysen und Eyetracking als Methoden anbieten, wobei die Implementierung der Logfileerstellung noch aussteht.

Es wurden auch Schritte eingeleitet um innovative Dienste am System zu testen. Besonders hervorzuheben ist die Einbindung von QR-Codes zur schnellen Weitergabe der Metadaten eines Dokuments. Hierbei ist nicht nur die technische Implementierung eine Herausforderung. Die größere Problematik besteht aus der Limitierung von 300 Zeichen, die ein QR-Code darstellen kann. Durch die Technik sollen Anwendungsmöglichkeiten für QR-Codes in einem Informationssystem aufgezeigt werden.

Referenzen:

Mutschke, P.; Mayr, P.; Schaer, P.; Sure, Y. (2011). Science Models as Value-Added Services for Scholarly Information Systems. Scientometrics, 89 (1), 349-364

Mayr, Philipp (2010): Information Retrieval. Mehrwertdienste für Digitale Bibliotheken. Crosskonkordanzen und Bradfordizing. Bonn: GESIS

Mayr, P.; Mutschke, P.; Petras, V.; Schaer, P.; Sure, Y. (2011). Applying Science Models for Search. In J. Griesbaum, T. Mandl & C. Womser-Hacker (Hrsg.), Information und Wissen: global, sozial und frei? (S. 184-196). 12. Internationales Symposium für Informationswissenschaft, Hildesheim, 9.—11. März 2011. Hildesheim. Boizenburg: Verlag Werner Hülsbusch