Automatische Indexierung von deutschsprachigen bibliographischen Metadaten im Bereich Bauwesen

14. Juni 2018

Dimitri Busch

Fraunhofer-Informationszentrum
Raum und Bau IRB, Stuttgart

107. Bibliothekartag,

Berlin, 12.-15. Juni 2018

Fraunhofer
IRB

Einführung

- Im Fraunhofer IRB wird Fachliteratur im Bereich Planen und Bauen bibliografisch erschlossen.
- Die Dokumente (bibliografische Metadaten) werden u.a. in der Datenbank RSWB®plus verwendet.
- Zu den Dokumenten werden Deskriptoren von einer Nomenklatur (Schlagwortliste IRB) und Notationen von einem Klassifikationssystem (Fachgliederung IRB) zugeordnet.
- Momentan wird die Indexierung intellektuell durchgeführt. Die Intellektuelle Indexierung ist zeitaufwendig und teuer.
- In der Präsentation geht es um ein (halb-)automatisches Indexierungssystem, das entwickelt wurde, um o.g. Probleme zu lösen.

Beispiel-Dokument

Originaltitel	Praxisbeispiel "Umbau Hauptbahnhof Hannover" - Anwendung der BIM-Methodik
Autor	Horstmann, Wolfgang
Referat	Im Hauptbahnhof Hannover sind sechs Bahnsteige mit den Bahnsteigdächern, Beleuchtungsanlagen, Aufzügen, Fahrtreppen, Fahrgastinformationsanlagen und den Bahnsteigausstattungen zu erneuern, fünf Eisenbahnbrücken sind zu ersetzen. Ein Gewölbe sowie die Entwässerungsleitungen des Bahnhofes sollen erneuert bzw. teilerneuert werden. Im Projekt wird die BIM-Methodik angewendet…
Schlagwörter	Hauptbahnhof; Personenverkehr; Verkehrsanlage; Bahnsteig; Eisenbahnbrücke; Umbau; Planungsmethode; Erneuerung; Building Information Modeling…
Fachgebiet	31.160 Ingenieurhochbau: Verkehrsanlage; 36.010 Bahnbau: Allgemein
Publikationstyp	Buchkapitel; Konferenzbeitrag
Quelle	Vorträge zum Deutschen Bautechnik-Tag am 27. und 28. April 2017 in Stuttgart, Seiten 109-110

Kontrollierte Vokabularien

- **Schlagwortliste IRB;** Typ: Nomenklatur. Autor: Fraunhofer IRB; Themen: Bauwesen, Raumordnung, Städtebau, Wohnungswesen; ca. 42400 Terme (Stand 06.2018). Beziehungen: nicht unterstützt; zweisprachig: Deutsch und Englisch.
- **Fachgliederung IRB;** Typ: Klassifikationssystem; Autor: Fraunhofer IRB; Themen: Bauwesen, Raumordnung, Städtebau, Wohnungswesen ; ca. 1000 Notationen (Stand 06.2018); Beziehungen: Hierarchie, 2 Ebenen; zweisprachig: Deutsch und Englisch.
- **FINDEX BAU/Raum**; Typ: Facettierte Thesauri. Legacy. Autor: Fraunhofer IRB; Thema: Bauwesen, Raumordnung; ca. 7500 Terme (auch in Schlagwortliste IRB enthalten); Beziehungen: u.a. Äquivalenz, Hierarchie

Vorarbeiten im Fraunhofer IRB

- Halbautomatische Indexierung von englischsprachigen Dokumenten mit Deskriptoren von der Schlagwortliste IRB
- Das Indexierungssystem verwendet JEX, eine freie Indexierungssoftware, die eine Java-Programmierschnittstelle hat [Steinberger/Ebrahim/Turchi 2012].

System für automatische Indexierung deutschspracher Dokumente

- Weiterentwicklung des bestehenden Indexierungssystems, um die Besonderheiten Deutscher Sprache zu berücksichtigen und Indexierungsqualität zu verbessern.
- Indexierung mit Deskriptoren
  - profilbasiertes Verfahren
  - Informationslinguistische Verfahren
- Indexierung mit Notationen nach k-NN-Verfahren

Indexierung mit Deskriptoren: Profilbasiertes Verfahren

Beispielprofil für den Deskriptor „Baustil“

Term	Gewicht
Gotik	0.263
Barock	0.255
Klassizismus	0.234
Historismus	0.187
Jugendstil	0.185
...	...

Erzeugung von Profilen aus Trainingsdokumenten nach dem parametrischen Rocchio- Algorithmus [Basili/Moschitti 2005].

Trainingsdokumente sind Dokumente, die bereits indexiert sind.

Berechnung von Ähnlichkeiten zwischen Dokumenten und Profilen.

Informationslinguistische Indexierungsverfahren

Informationslinguistische Verfahren ermitteln Indexterme auf Basis sprachlicher Gesetzmäßigkeiten.

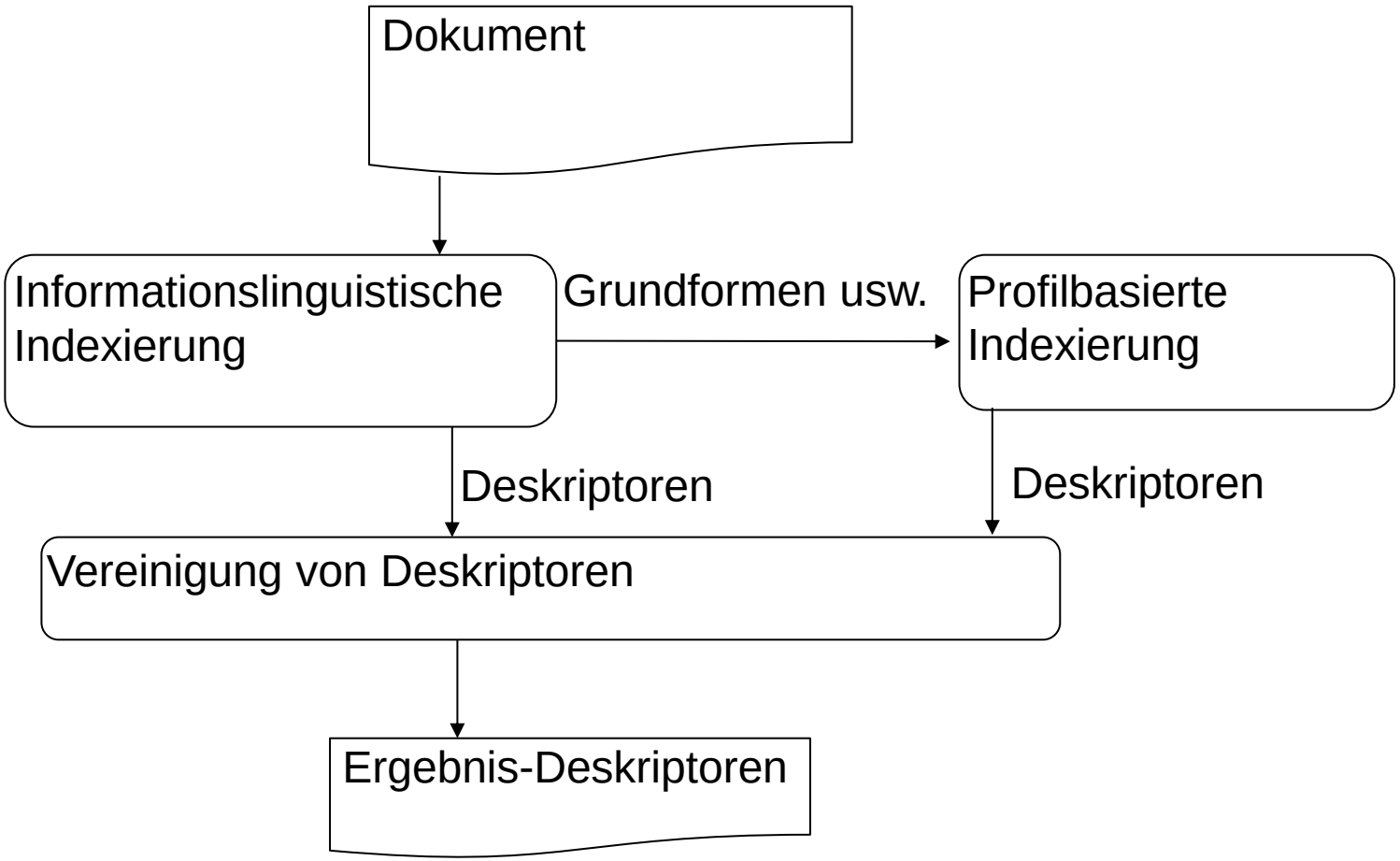
Zu Informationslinguistischen Verfahren gehören u.a.:

- Lemmatisierung (Grundformreduktion)
- Erkennung und Zerlegung von Komposita
- Mehrwortgruppenerkennung
- Erkennung von Synonymen

Beispiele:

- Schlösser->Schloss
- Residenzschloss->Residenz+Schloss
- Wärmeisolierung->Wärmedämmung

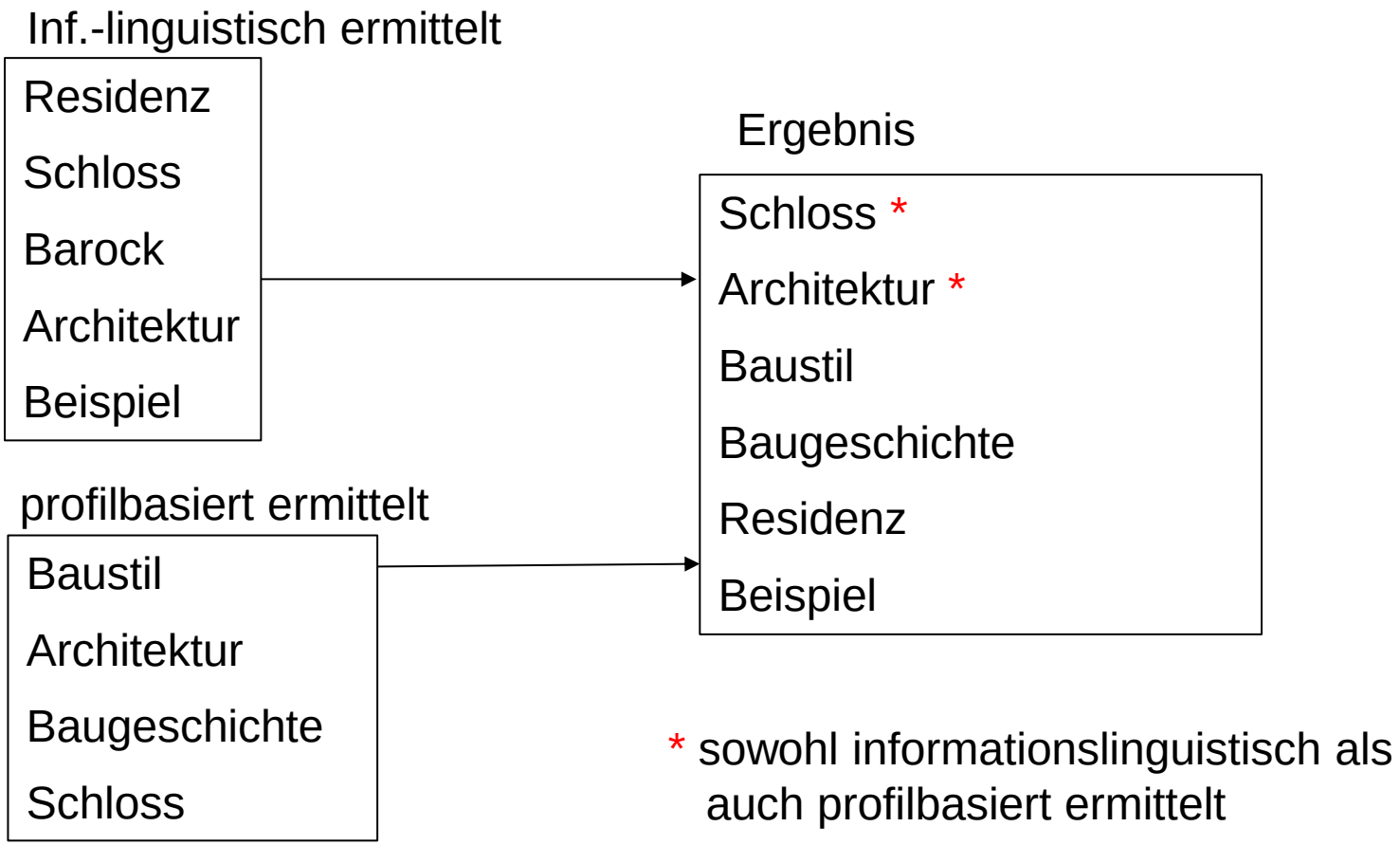
Indexierung mit Deskriptoren: Ablauf



Vereinigung von Deskriptoren: Beispiel

Titel: Ludwigsburger Residenzschloss

Abstract: Der Artikel befasst sich mit barocker Architektur am Beispiel …



Halbautomatische Zuordnung von Deskriptoren: Beispiel

Dokument:

Titel:

Ludwigsburger Residenzschloss

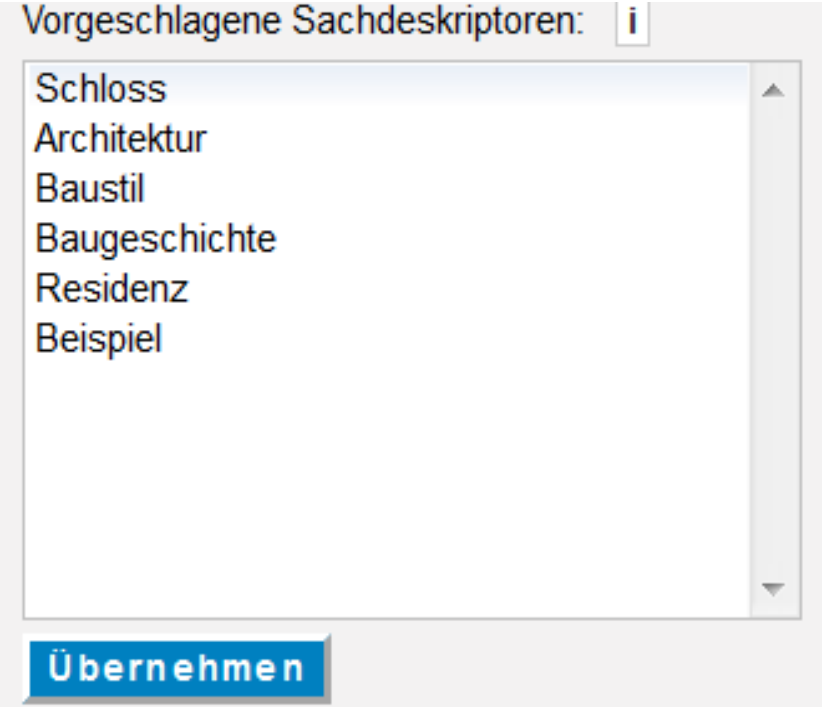
Abstract:

Der Artikel befasst sich mit barocker Architektur am Beispiel Ludwigsburger Residenzschlosses

Zugeordnete Deskriptoren:

Schloss, Architektur, Baustil, Baugeschichte, Residenz

Vorgeschlagene Deskriptoren:



Indexierung mit Notationen: Instanzbasiertes Verfahren

Das Verfahren orientiert sich an die Notationen, die jene Trainingsdokumente aufweisen, die dem zu indexierenden Dokument am ähnlichsten sind.

k Nearest Neighbors (k-NN):

- Finden einer bestimmten Anzahl (k) der Trainingsdokumente, die dem zu indexierenden Dokument am ähnlichsten sind.
- Zuordnung von Notationen, die in den gefundenen Trainingsdokumenten am häufigsten vorkommen.

Halbautomatische Zuordnung von Notationen: Beispiel

Dokument:

Titel:

Ludwigsburger Residenzschloss

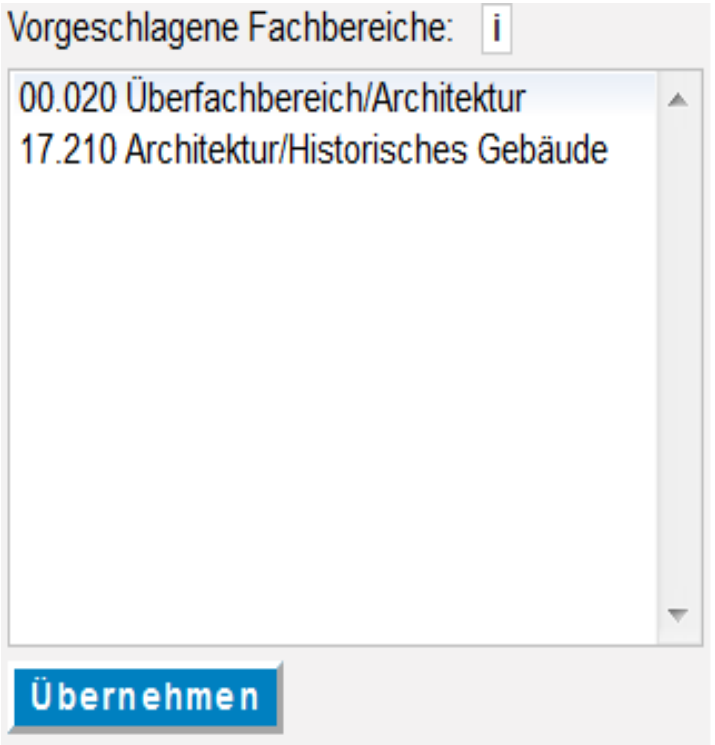
Abstract:

Der Artikel befasst sich mit barocker Architektur am Beispiel Ludwigsburger Residenzschlosses

Zugeordnete Fachbereiche:

00.020; 17.210

Vorgeschlagene Fachgebiete:



Anpassungen an thematische Domänen

- Die Indexierungsprozeduren können an bestimmte thematische Domänen angepasst werden, z.B. Betonbau, Gebäudetechnik
- Trainieren von profilbasierten Klassifikatoren mit Dokumenten aus einschlägigen Zeitschriften
- Aufbau von domäne-spezifischen Wörterbüchern (z.B. aus entsprechenden Facetten von FINDEX-Thesauri)

Implementierung

- Die meisten Programme sind in Java implementiert.
- Informationslinguistische Indexierung mit freier Software LINGO . [Gödert,/Lepsky/Nagelschmidt 2012]
- LINGO benötigt eine Ruby-Ungebung
- K-NN Indexierung mit LUCENE
- Automatische Ermittlung von Indextermen im Batch-Modus
- Computerunterstützte Zuordnung der Indexterme über eine webbasierte Benutzeroberfläche, die mit JSF implementiert wurde

Zusammenfassung

- Der vorgestellte Einsatz eignet sich insbesondere für halbautomatische Indexierung, bei der ein menschlicher Indexierer eine endgültige Entscheidung über die Zuordnung von Deskriptoren und Notationen trifft.
- Nach entsprechenden Anpassungen wird es voraussichtlich möglich, auch eine vollautomatische Indexierung für bestimmte thematische Domänen durchzuführen.

Literatur

- Basili, R. ; Moschitti, A. (2005). Automatic Text Categorization. Rom: Aracne
- Gödert, W.; Lepsky, K.; Nagelschmidt, M. (2012): Informationserschließung und Automatisches Indexieren. Heidelberg: Springer
- Steinberger, R.; Ebrahim, M.; Turchi, M. (2012). JRC EuroVoc Indexer JEX – A freely available multi-label categorisation tool. In Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012), Istambul, 21-27. Mai 2012, S.798-805

Vielen Dank für Ihre Aufmerksamkeit