
Domäne-spezifische hybride automatische Indexierung von bibliographischen Metadaten

1. März 2019

Dimitri Busch

Fraunhofer-Informationszentrum
Raum und Bau (IRB), Stuttgart

Digitale Bibliothek 2019,
Digitale Horizonte,
Graz, 28. Februar- 1. März 2019

Einführung

- Im Fraunhofer IRB wird Fachliteratur im Bereich Planen und Bauen bibliographisch erschlossen.
- Die Dokumente (bibliographische Metadaten) werden in bibliographischen Datenbanken IRB verwendet.
- Zu den Dokumenten werden Deskriptoren von einer Nomenklatur (Schlagwortliste IRB) zugeordnet.
- Momentan wird die Indexierung intellektuell durchgeführt. Die Intellektuelle Indexierung ist zeitaufwendig und teuer.
- In der Präsentation geht es um ein automatisches Indexierungssystem, das entwickelt wurde, um o.g. Probleme zu lösen.

Beispiel-Dokument

Originaltitel: Jagdburg. Begriffe erkunden

Autor: Laß, Heiko

Abstract: Der Begriff "Jagdburg" ist von der modernen Wissenschaft in Analogie zum Begriff "Jagdschloss" gebildet und bezeichnet einen Funktionstypus eines meist adeligen Profanbaus des Mittelalters. Es handelt sich um Bauten, von denen aus vornehmlich gejagt wurde. Ihre Funktion ist in den zeitgenössischen Schriftquellen meist nicht explizit benannt und die Anlagen wurden oft multifunktional genutzt. Spätestens im 13. Jahrhundert finden sich Jagdburgen im römisch-deutschen Reich nicht mehr nur beim Kaiser, sondern auch bei den Fürsten.

Schlagwörter: Kulturgeschichte; Burg; Schloss; Terminologie; Architektur; Jagdschloss; Forschung; Begriff; Funktion; Verbreitung; Besonderheit; Beispiel

Publikationstyp: Zeitschriftenartikel

Quelle: Burgen und Schlösser (2018), Jg.59, Nr.3, S.188-189

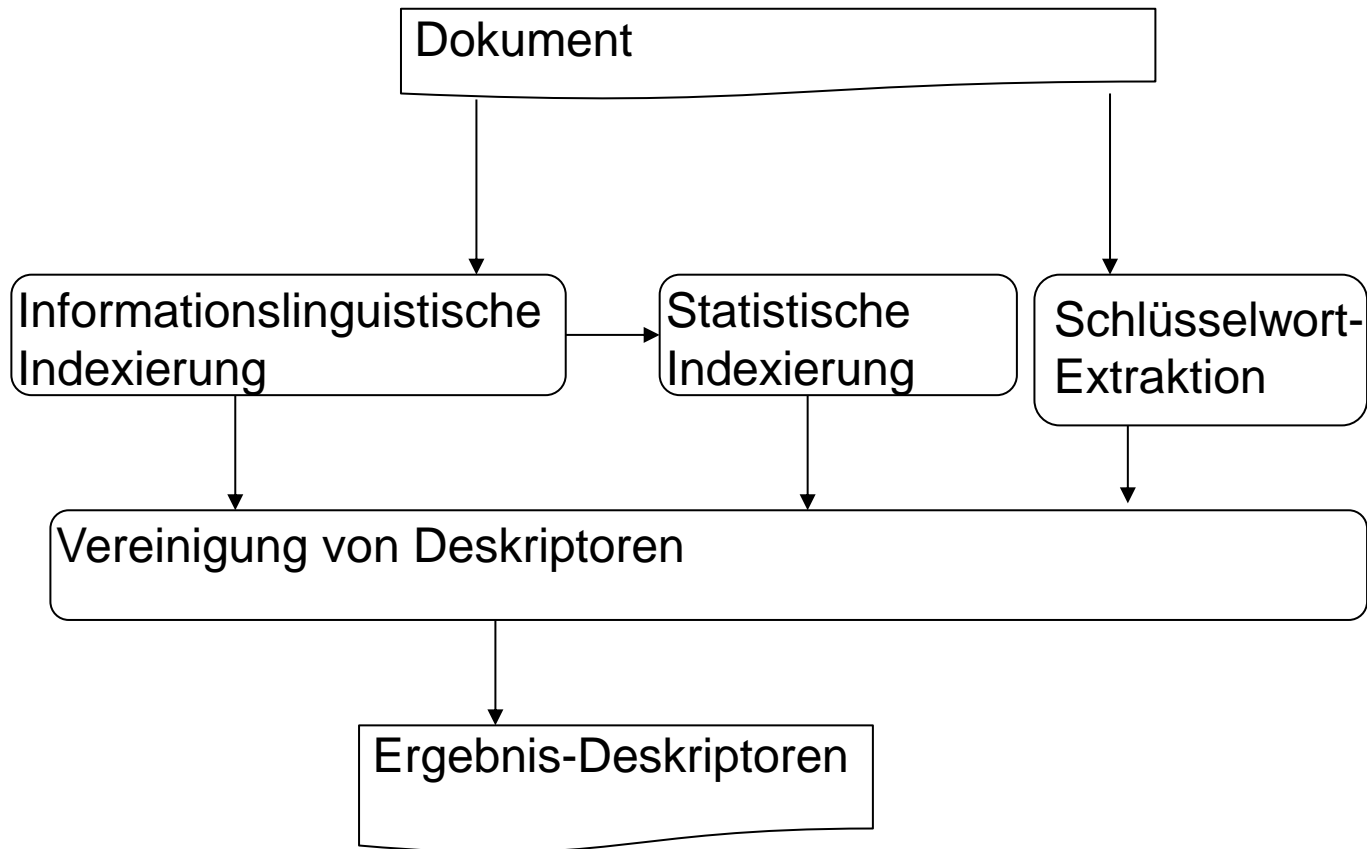
Schlagwortliste IRB

- Typ: Nomenklatur.
- Autor: Fraunhofer IRB.
- Themen: Bauwesen, Raumordnung, Städtebau, Wohnungswesen.
- Anzahl der Terme: ca. 38600 Terme (Stand 02.2019).
- Sprachen: Deutsch und Englisch.
- Besonderheiten:
 - Keine Beziehungen zwischen Termen sind unterstützt.
 - Viele Deskriptoren werden selten oder nie verwendet.

Probleme mit automatischer Indexierung

- Für automatische Indexierung werden u.a. informationslinguistische und statistische Verfahren verwendet [vgl. Gödert/Lepsky/Nagelschmidt 2012].
- Probleme bei der Anwendung von informationslinguistischen Verfahren auf die Schlagwortliste IRB:
 - Viele wichtige Deskriptoren können nicht erkannt werden, da in der Schlagwortliste keine Beziehungen zwischen Termen unterstützt sind.
- Probleme bei der Anwendung von statistischen Verfahren auf die Schlagwortliste IRB:
 - Für viele Deskriptoren gibt es nicht genügend Trainingsdokumente.
 - Inkonsistente Indexierung von Trainingsdokumenten.
- Lösung: Ensemble aus mehreren Indexierungsprozeduren, die nach verschiedenen Verfahren funktionieren.

Indexierung mit Deskriptoren: Ablauf



Software

- Informationslinguistische Indexierung: freie Software **Lingo**.
[Gödert,/Lepsky/Nagelschmidt 2012]
- Schlüsselwort-Extraktion: freie Software **KEA** [Medelyan/Witten 2006]
- Statistische Indexierung:
 - Java-Implementierung des „Parameterized Rocchio Classifier“ (**PRC**)
[Basili/Moschitti 2005]
 - freie Software **JEX** [Steinberger/Ebrahim/Turchi 2012].

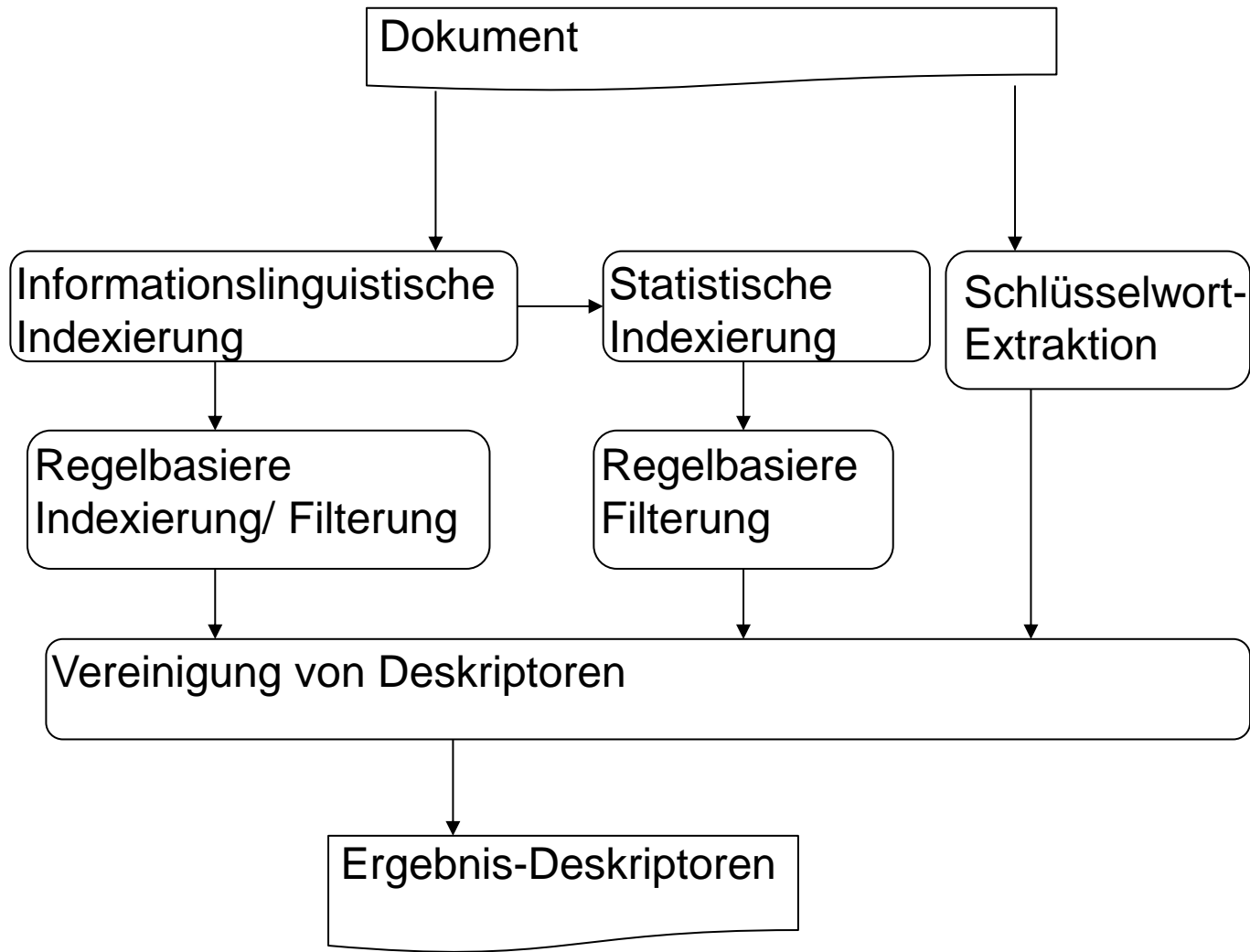
Indexierung eines Dokuments: Beispiel

Titel: Jagdбург. Begriffe Erkunden

Abstract: Der Begriff "Jagdбург" ist von der modernen Wissenschaft in Analogie zum Begriff "Jagdschloss" gebildet und bezeichnet einen Funktionstypus eines meist adeligen Profanbaus des Mittelalters. Es handelt sich um Bauten, von denen aus vornehmlich gejagt wurde...

Term	Lingo	KEA	PRC	JEX	Zuordnung
Jagd	*		*		Ja
Burg	*		*		Ja
Begriff	*		*		Ja
Wissenschaft	*				Nein
Jagdschloss	*		*	*	Ja
Profanbau	*	*			Ja
Mittelalter	*	*	*		Ja
Baugeschichte			*	*	.Ja
...					

Regelbasierte Tuning und Filterung



Regeltypen

- Ausreichende Regeln

Beispiel:

Barock -> Architekturstil

(Wenn das Dokument das Wort „Barock“ enthält,
wird der Deskriptor „Architekturstil“ zugeordnet)

- Notwendige Regeln

Beispiel:

Abfall OR Müll <- Abfallbeseitigung

(Der Deskriptor „Abfallbeseitigung“ darf nur dann zugeordnet werden,
wenn das Dokument das Wort „Abfall“ oder das Wort „Müll“ enthält)

- Negative Regeln

Beispiel:

in der Regel – Regel

(Der Deskriptor „Regel“ darf nicht zugeordnet werden,
wenn das Dokument die feste Sequenz „in der Regel“ enthält)

Anpassungen an thematische Domänen

- Die Indexierungsprozeduren können an bestimmte thematische Domänen angepasst werden.
- Trainieren von statistischen Indexierungsprozeduren und der Schlüsselwort-Extraktion mit bereits indexierten Dokumenten, die Artikel aus domäne-spezifischen Fachzeitschriften beschreiben.
- Beispiel:
 - Domäne: Architekturgeschichte
 - Zeitschriften: „Architectura“, „Burger und Schlösser“
- Aufbau von domäne-spezifischen Wörterbüchern auf Basis von Deskriptoren, die zu Dokumenten zugeordnet wurden, die Artikel aus domäne-spezifischen Fachzeitschriften beschreiben.

Indexierung eines Dokuments: Testumgebung

Titel: Jagdбург. Begriffe Erkunden

Abstract: Der Begriff "Jagdбург" ist von der modernen Wissenschaft in Analogie zum Begriff "Jagdschloss" gebildet und bezeichnet einen Funktionstypus eines meist adeligen Profanbaus des Mittelalters. Es handelt sich um Bauten, von denen aus vornehmlich gejagt wurde...

Sachdeskriptoren:

Jagd; Burg; Begriff;
Jagdschloss; **Profanbau;
Mittelalter; Baugeschichte

Vorgeschlagene Sachdeskriptoren: i

Jagd
Burg
Begriff
Jagdschloss
Profanbau
Mittelalter
Baugeschichte

Passt gut

Kommt in Frage (**)

Irrelevant (***)

Falsch (****)

Zusammenfassung

- Das vorgestellte System erlaubt die vollautomatische Indexierung ohne die Beteiligung eines menschlichen Indexierers.
- Das System verwendet mehrere Indexierungsverfahren, die erlauben Probleme zu lösen, die wegen mangelnder Trainingsdaten und fehlender Angaben über Beziehungen zwischen Termen entstehen.
- Indexierungsprozeduren können an thematische Domänen angepasst werden.
- Das System befindet sich derzeit in dem Alphaversions-Stadium.
- Künftig können evtl. neue fortgeschrittene Verfahren in das System integriert werden.
- Obwohl der Ansatz hauptsächlich für den Bereich Bauwesen bestimmt ist, ist er auch auf andere Bereiche prinzipiell übertragbar.

Literatur

- Basili, R. ; Moschitti, A. (2005). Automatic Text Categorization. Rom: Aracne
- Gödert, W.; Lepsky, K.; Nagelschmidt, M. (2012): Informationserschließung und Automatisches Indexieren. Heidelberg: Springer
- Medelyan, O., Witten I. H. (2006) Thesaurus Based Automatic Keyphrase Indexing. In Proc. of the Joint Conference on Digital Libraries 2006, Chapel Hill, NC, USA, pp. 296-297.
- 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), Istanbul, 21-27. Mai 2012, S.798-805

Vielen Dank für Ihre Aufmerksamkeit