Gedanken zu einem Whitepaper „KI in Bibliotheken“

Wie kommen wir zu einer modernen Inhaltserschließung? ungehobene Schätze in neuen Technologien

Dr. Anna Kasprzik
InnoCamp 2020
05.11.2020

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Status quo 1: Wie wird Information kodiert?

Wie werden Metadaten aktuell strukturiert und abgelegt?

- Ablage in gefelderten Datensätzen
- Inhalte der Felder nicht standardmäßig strukturiert / typisiert
  (simples Bsp.: Language Tagging!)

→ vieles noch sehr von Menschenlesbarkeit her gedacht
(Angaben in Textform, implizite Annahmen zu Reihenfolge, Bezügen etc.)

keine echte Maschinenlesbarkeit und damit
fehlende Ansatzpunkte für neue semantische Technologien / KI !

Chance 1: Wie wird Information kodiert?

- stärker strukturierte Ablage und weitergehende Datentypisierung nach modernen Standards und nach dem Vorbild des Semantic Web – Richtung:
    - Ablage anhand des RDF-Modells
    - Linked-Data-Prinzipien
    - FAIR-Prinzipien
- mittelfristig: nicht mehr dokumentenbasiert, sondern entitätenbasiert denken – keine künstliche Trennung zwischen Formal- und Inhaltserschließung, Sachschlagworten und Normierung von Personen und Körperschaften, … (s.a. Open Research Knowledge Graph der TIB als Schritt in diese Richtung)

→ *neuro-symbolische Integration …*

Status quo 2: Beziehung Inhaltserschließung – Retrieval

Klassische Beziehung Inhaltserschließung – Retrieval an Bibliotheken:

- Fachreferat erschließt mit Schlagwörtern aus (poly-)hierarchischem Thesaurus
- Schlagwörter werden in einem Datenfeld im Metadatensatz abgelegt
- textuelle Felder im Metadatensatz werden fürs Retrieval aufbereitet *(tf-idf*, Indexierung)
- hierbei wird höchstens der Inhalt des Schlagwortfeldes als Ganzes (i.e., als Text) etwas gepusht bei der Berechnung des Rankings im Discovery System

Chance 2: Beziehung Inhaltserschließung – Retrieval

Inventur: Welche Informationen aus der Inhaltserschließung (IE)
könnte das Retrieval noch aufgreifen und auswerten?

- semantische Strukturinformationen, z.B. hierarchische Relationen im Thesaurus (perspektivisch: Relationen und logische Restriktionen aus einer Ontologie)

- mit zunehmender Automatisierung der IE:
  - NLP-Preprocessing vorschalten und dann für beide Vorgänge nachnutzen?
  - mitgelieferte Konfidenzwerte aus der maschinellen IE nutzen fürs Ranking? weitere Provenienzinformationen nutzen? (z.B. verwendete Qualitätsfilter, Menge der Trainingsdaten, … ?)

siehe auch Beiträge von Christian Wartena und Jan Frederik Maas
hier: https://wiki.dnb.de/pages/viewpage.action?pageId=149292829

Status quo 3: Formen der Recherche

Recherche mit Hilfe von

- Eingabe in einen Textschlitz (+ Boolesche Operatoren)
- erweiterte Suche: Eingaben in mehrere Textschlitze (~ Datenfelder)
- Drilldown über angebotene Facetten

Chance 3: Formen der Recherche

- vielfältige Arten der Eingabe, z.B. auch
  - Eingabe längerer Textpassagen für Suche nach Ressourcen mit ähnlichem semantischen Profil
  - natürlichsprachliche Interaktion
  - … (*insert your own idea here*)
- fließender Übergang zwischen gezielter und explorativer Suche (*serendipity*)
  - (ggf. von Ressource/Entität aus) Hangeln durch ein Wissensnetz anhand verschiedener Entitätentypen, z.B. Konzepten, Autor*n, …
  - … (*insert your own idea here bzw. siehe Impulsvortrag Ralph Ewerth*)