Andreas Blumauer

Von Taxonomien über Text Mining zu Linked Data

Taxonomien werden in vielen Fällen dazu verwendet, um Inhalte konsistent zu klassifizieren. In Kombination mit Text Mining und Linked Data Technologien gelingt der Schritt hin zur Graph-basierten Repräsentation ausdrucksstarker Wissensmodelle, um Zusammenhänge in großen Informationsspeichern besser erschließen zu können. Wie dies gelingt und welchen Nutzen dies stiften kann, erläutert dieser Beitrag.

➜ Taxonomien als Ausgangspunkt

Taxonomien entsprechen einfachen Wissensmodellen (Ontologien) und werden üblicherweise zur Annotation und Klassifikation von Dokumenten verwendet. Der SKOS-Standard (Simple Knowledge Organization System) zur Beschreibung von kontrollierten Vokabularen erweitert die Möglichkeiten einfacher Taxonomien: Hier werden Entitäten unterschiedlichster Kategorien (Personen, Organisationen, Produkte, Orte, ...) nicht nur hierarchisch, sondern als vernetzter Wissensgraph organisiert. Jeder Entität können pro Sprache auch mehrere Bezeichnungen (z. B. Synonyme) zugeordnet werden.

Verfahren des automatischen Text Minings helfen u.a., Inhalte besser zu erschließen. Seit vielen Jahren wird auch von der vollautomatischen Erstellung semantischer Netze gesprochen, die mittels computer-linguistischer und statistischer Verfahren aus Texten abgeleitet werden sollen, letztendlich aber nur relativ flache Wortnetze darstellen, die Korrelationen, aber keine semantischen Zusammenhänge repräsentieren. Höher entwickelte Verfahren des Text Minings setzen diese Verfahren schon kombiniert mit Wort- bzw. Synonymlisten (z. B. Gazetteers) ein.

➜ Wissensgraphen dynamisch generieren

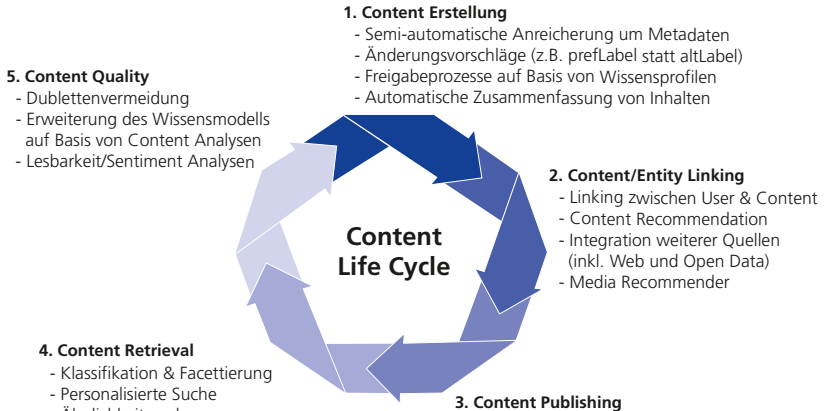
Mit dem Aufkeimen von Linked Data wurden in den letzten Jahren umfassende Wissensgraphen in unterschiedlichsten Fachbereichen verfügbar gemacht (z. B. Geonames, MeSH, Eurovoc, DBpedia, etc.). Einerseits werden diese nun zunehmend als Grundstock für unternehmensspezifische Wissensgraphen herangezogen, andererseits auch als Basis zur automatischen Extraktion von Entitäten aus großen Dokumentbeständen.

Damit gelingt es, nicht einfach nur Terme und ihre Korrelationen aus Texten automatisch zu extrahieren, sondern Wissensgraphen dynamisch generieren und laufend erweitern zu können. Dies bildet die Grundlage hochwertiger semantischer Services entlang eines typischen Content Life Cycles.

➜ Konsistente, vernetzte Metadaten

Im Rahmen der Content-Verwertung dienen graph-basierte Linked Data-Standards der Anreicherung von Informationsbeständen um wertvolle, da konsistente und vernetzte Metadaten. Diese ermöglichen es erst, ähnliche oder verwandte Objekte zueinander mit hoher Präzision in Beziehung zu setzen. Im Zentrum der aktuellen Entwicklung stehen zwar immer noch Such- und Empfehlungsdienste, die das Dokument im Zentrum ihrer Informationsarchitektur haben, jedoch findet allmählich auch innerhalb von Unternehmensgrenzen eine Transformation hin zur graph-basierten Verarbeitung von strukturierten und unstrukturierten Informationen und ihren Metadaten statt.

➜ Einsatzmöglichkeiten von Linked Data entlang des Content Life Cycles



➜ Der Autor

Andreas Blumauer ist Wirtschaftsinformatiker und Geschäftsführer der Semantic Web Company GmbH. Die Semantic Web Company ist anerkannter Pionier im Semantic Web und bietet seit 2009 ein Produkt am globalen Markt semantischer Technologien an: Mit der PoolParty Semantic Suite gelingt es, unternehmensinterne und -externe Informationsbestände sinnvoll zu verknüpfen und komfortabel durchsuchbar zu machen. Wissensarbeiter profitieren von intelligenteren Software-Anwendungen.

✉ *blumauer@wissensmanagement.net*