Karlsruher Institut für Technologie
Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB)
Andreas Thalhammer
thalhammer@kit.edu

16.02. 2015

Bachelorarbeit

Intelligentes Extrahieren von Seiten-Links aus Wikipedia

Um was geht es?

In Wikipedia sind viele Seiten miteinander verknüpft. Aus dem dadurch entstehenden Link-Graphen kann mit Verfahren der Graphanalyse neues Wissen über die Wichtigkeit von Seiten gewonnen werden. Diese Analyseverfahren liefern jedoch nur dann gute Ergebnisse, wenn die gesetzten Links tatsächlich einen direkten Bezug zur aktuellen Seite aufweisen. Ein Beispiel: Im Englischen Wikipedia sind auf vielen Seiten Links vorhanden, welche auf die Nomenklatur von Carl von Linné verweisen[1]. Diese Links sind zwar notwendig wenn es darum geht, zoologische Namensgebungen aufzuklären, tragen aber nicht die gleiche Bedeutung wie Links zu Seiten, welche einen klareren Bezug zum Artikel haben. Thema der Arbeit ist es, ein verbessertes Verfahren zur Linkextraktion zu implementieren und zu testen. Hier kann auf bestehende Verfahren und Code aus der DBpedia-Gemeinschaft[2] zurückgegriffen werden.

Was sollten Sie mitbringen?

- Interesse an Datengraphen und modernen Ranking-Verfahren
- Interesse an der Verarbeitung von großen Datenmengen
- Gute Programmierkenntnisse

Weitere Materialien

- http://people.aifb.kit.edu/ath/#DBpedia_PageRank
- https://github.com/dbpedia/extraction-framework

Kontaktperson:
Andreas Thalhammer
thalhammer@kit.edu
Tel.: 0721/608 47363

[1] http://en.wikipedia.org/wiki/Special:WhatLinksHere/Carl_Linnaeus
[2] http://dbpedia.org