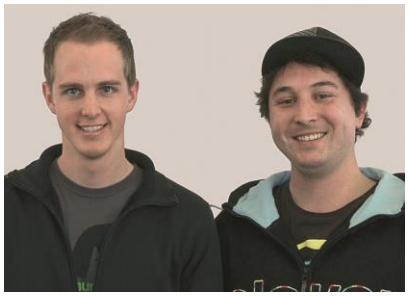
Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

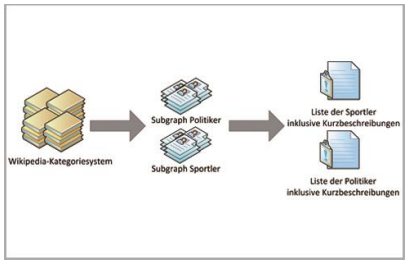
War da was? Mining mittels Wikipedia



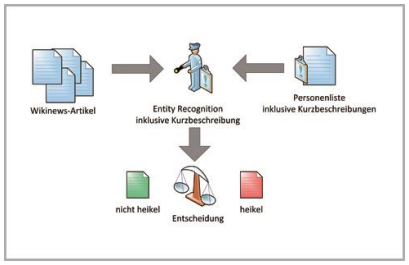
Diplomierende
Matthias Hug
Severin Wirth

Dozent
Martin Braschler

In dieser Bachelorarbeit wird ein bestehendes System erweitert, das Personen in News-Artikeln automatisch erkennt und die Dokumente zusätzlich auf charakteristische Schlagwörter überprüft. Anhand der Schlagwörter kann entschieden werden, ob ein Dokument für ein bestimmtes Thema heikle Aspekte beinhaltet oder nicht. Bei der Personenerkennung besteht die Schwierigkeit, dass in den News-Artikeln unterschiedliche Personen mit identischem Namen nicht differenziert werden können. Um dieses Problem zu beheben, extrahiert man aus Wikipedia-Artikeln zusätzliche Informationen zu den Tätigkeiten der Personen und speichert diese ergänzend zu den Namen in XML-Listen.
In einer vorgängigen Masterarbeit wurde ein entsprechendes System entwickelt, das aus zwei Komponenten besteht. Die erste Komponente dient zur Generierung von Personenlisten, die aus Wikipedia exportiert werden. Dazu wurde ein GUI erstellt, das anhand des Wikipedia-Kategoriesystems gerichtete Graphen darstellt und daraus die Listen exportiert. Die auf "Lucene" basierende zweite Komponente analysiert News-Artikel, wobei die Nachrichten von einer freien Quelle bezogen werden.
Die entwickelte Applikation ist eine vereinfachte Variante eines sogenannten "Named Entity Recognition"-Systems (NER-System), das in einem unstrukturierten und in natürlicher Sprache verfassten Textdokument Personennamen (Named Entities, NE) erkennt. Thomas Müller ist einerseits ein Schweizer Politiker und andererseits ein deutscher Fussballspieler. In diesem Zusammenhang sollte das System die im Textdokument enthaltene Zeichenkette "Thomas Müller" als NE erkennen und der Kategorie "Person" zuweisen. Zusätzlich sollen auch die Wörter "Politiker" oder "Fussballspieler" erkannt und in der Kategorie "Tätigkeit" eingeteilt werden. Anhand dieser Bestimmung kann das System feststellen, um welchen Thomas Müller es sich handelt.
Die aktuelle Lösung findet heikle Inhalte in News-Artikeln, ohne dass danach gesucht werden muss. Das kann insbesondere für ein Risk Assessment sehr nützlich sein. Der Umgang mit Political Exposed Persons (PEP) verlangt eine gewisse Sorgfaltspflicht, wenn es um die Frage geht, ob eine Person zum Kundenstamm gehören soll oder nicht. Hier soll der Prototyp als Warnsystem dienen, das Auskunft über kritische Geschehnisse von PEP gibt.



Das erste Schema zeigt den Ablauf der Generierung von Personenlisten, die aus Wikipedia exportiert werden.



Das zweite Schema zeigt den Ablauf der Analyse von frei bezogenen News-Artikeln.