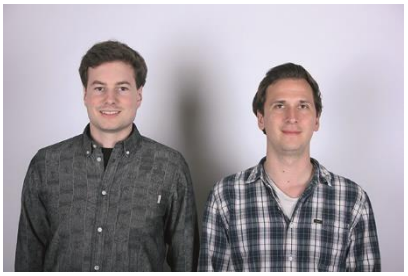
Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

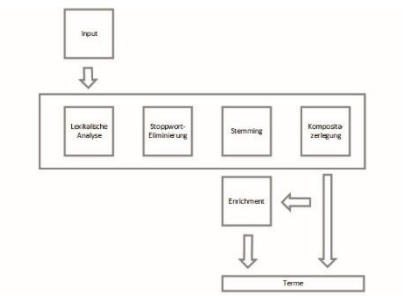
Sprechen Sie Deutsch? - Retrieval-Suchtechnologie für natürliche Sprache anpassen



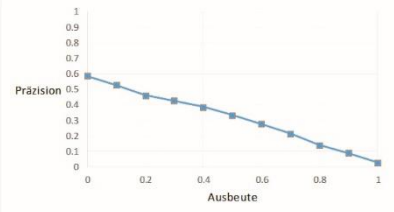
Diplomierende
Benjamin Sutter
Timo Visscher

Dozent
Martin Braschler

Information Retrieval behandelt Methoden, um Informationen aus grossen Mengen an unstrukturierten Daten durch gezielte Verarbeitung von Dokumenten und Suchanfragen zu gewinnen. Das verarbeitende System muss, unabhängig von der verwendeten Sprache, ein treffendes Ergebnis liefern. Auf der einen Seite gibt es hierfür sprachneutrale und auf der anderen fortgeschrittene, computerlinguistische und somit sprachabhängige Ansätze. Im Vergleich zum Englischen gehört Deutsch zu den stark flektierenden Sprachen, weil die deutsche Grammatik eine grosse Menge an Wortformen hergibt. So wird aus „Baum“ im Genitiv „Baumes“ oder das Verb „gehen“ konjugiert in der Vergangenheitsform zu „gegangen“. Hinzu kommt, dass in Deutsch aufgrund der Kompositabildung fast beliebige Wörter zusammengesetzt werden können. Es muss also speziell beachtet werden, wie mit Komposita wie „Fussballweltmeisterschaftseröffnungsspiel“ oder „Donaudampfschiff“ umgegangen wird.
Die im Rahmen der Bachelorarbeit durchgeführten Experimente veranschaulichen, dass ein Suchsystem grundsätzlich von einem leichten Stemmer (Stammformreduktion von Wörtern), einer Stoppwort-Eliminierung, einer Kompositazerlegung und von einer Normalisierung der Dokumente beim Indexierungsprozess profitiert. Die Suchanfrage muss weiter unterschieden werden, je nachdem, ob nach einem oder mehreren Wörtern gesucht wird. Bei der Suchanfrage mit nur einem Stichwort soll zwingend ein Blind Relevance Feedback durchgeführt werden. Auch eine gezielte Kompositazerlegung wirkt sich positiv auf die Retrievaleffektivität aus. Teile von Suchanfragen, welche einen Eigennamen (Personennamen, Ortschaftsnamen oder Organisationsnamen) beinhalten, dürfen nicht gestemmt oder normalisiert werden. Dasselbe gilt für Phrasen wie zum Beispiel „der schiefe Turm von Pisa“. Akronyme oder Abkürzungen profitieren vom Blind Relevance Feedback, so können weitere relevante Begriffe gefunden werden.
Diese Erkenntnisse wurden aufgrund von Experimenten mit den Testkollektionen des als Verarbeitungsempfehlungen dokumentiert und in einem Prototyp mittels in implementiert. Anschliessend wurden die Verarbeitungsempfehlungen auf zwei weiteren Dokumentensammlungen ohne Relevanzbewertung evaluiert.



Für das Information-Retrieval-System wurde eine Verarbeitungspipeline definiert.



Nach jedem Experiment wurde die Effektivität des angepassten Information-Retrieval-Systems ausgewertet.