Christoph Jan Richter

Über die Vermeidung redundanter Betrachtungen beim Approximate String Matching

Kurzfassung

Im heutigen Informationszeitalter liegen fast beliebig viele Daten und Informationen digital gespeichert vor und es werden von Tag zu Tag mehr. Zur Verarbeitung und Nutzung dieser Daten haben sich Suchverfahren als essenzielle Werkzeuge etabliert. Allgemein werden Daten in Form von Zeichenketten, auch Strings genannt, gespeichert und so basieren die meisten Suchverfahren auf Zeichenketten. Durch lange Forschung stellt das Problem der exakten Suche in der Praxis kein Problem mehr dar. Doch noch viel größer in der Bedeutung wurde in den letzten Jahren die fehlertolerierende Suche, üblicherweise Approximate String Matching genannt. Das Approximate String Matching ist durch die Beachtung möglicher Fehler deutlich aufwendiger als die exakte Suche, weshalb insbesondere bei großen Datenmengen schnelle Algorithmen gefordert sind.

Das Ziel dieser Arbeit ist es, durch die Ausnutzung von Redundanzen beschleunigte Algorithmen für das Approximate String Matching zu erreichen. Die Betrachtungen fokussieren sich dabei auf die Gruppe der Filter-Algorithmen, die sich in der Praxis als die schnellsten erwiesen haben. Dabei wird für diese Filter-Algorithmen ein Klassifikationsschema vorgestellt, welches nach den verschiedenen, prinzipiellen Ansatzmöglichkeiten unterscheidet.

Filter-Algorithmen arbeiten zweiphasig, wobei in der ersten Phase Bereiche mit möglichen Lösungsstellen identifiziert werden, die in der zweiten, aufwendigeren Phase dann überprüft werden. In dieser zweiten Phase kann es zur mehrfachen - und damit redundanten - Betrachtung mancher Bereiche im zu durchsuchenden Text kommen. Das in dieser Arbeit vorgestellte Verfahren der Patchwork Verifikation stellt einen allgemeinen Ansatz dar, diese redundanten Betrachtungen zu vermeiden und dadurch Berechnungszeit zu sparen. Die sich für dieses Verfahren ergebenden konkreten Nutzungsgrenzen werden berechnet und anhand einer Implementierung demonstriert.

Weiterhin wird in dieser Arbeit der Algorithmus GraI vorgestellt, der auf der Idee basiert, die Betrachtung von redundanten Bereichen zu vermeiden, die im zu durchsuchenden Text identifiziert werden können. Zur Erkennung redundanter Textbereiche wird eine Grammatik verwendet, bei der sich aus jeweils identischen Textstellen eine Regel ableitet. Damit ist es möglich, mehrfach vorhandene Textbereiche in beiden Phasen des Filter-Ansatzes jeweils immer nur einmal zu betrachten. Auch für diesen Algorithmus wurden die Nutzungsgrenzen berechnet und demonstriert. Zudem wurde die Anwendbarkeit anhand unterschiedlichster Textdaten aufgezeigt.