ZUSAMMENFASSUNGEN COLLOQUIUM, REPETITORIUM

Pascal Guldener

10.7.2017

1 Colloquium

1.1 WebSuche: PageRank & HITS

In ihrem Vortrag verglich Anastasia Bespala zwei Indizierungsalgorithmen für die Suche im Web miteinander. Zunächst stellte Sie die Grundannahmen über den zugrundeliegenden Suchraum und die Funktion von Crawlern vor. Crwaler tokenisieren Dokumente im Netz und erstellen über den Resultaten Zuordnungen zwischen den Token und Mengen von Dokumenten in welche diesen vorkamen. Diese werden in sog Term-Dokument-Dateien gespeichert. Bei einer Suchanfrage kann es dann zu einer grossen Anzahl an Treffern kommen, was die Frage nach der Relevanz nac welcher die Terfferliste geordnet sein sollte aufwirft. Die hier vorgestellten Algorithmen, welche diesa AUfgabe übernehemen sind PageRank unf HITS. Zunächst ging Frau Bespala auf PageRank ein, welcher von den Google Gründern, Larry Page und Sergei Brin, 1996 entwickelte wurde. Er beruht auf der Verlinkungsstruktur im Hypertext und der Annahme daß ein Dokument je relevantere Information enthält, je mehr andere Dokumente darauf verlinken. Da aber das Gewicht bei der Berechung der Relevanz ist jedoch wiederum von der Relevanz der verlinkenden Quelle abhängig was zu einer rekursiven Definition von Relevanz führt. Das dabei aufkommende Problem der zirluklären Verlinkung und der Linksenke wir mit der Verwendung eines Quellenrangs gelöst welcher sich über die Anzahl aller Seiten verteilt. Dadurch kann man das Web als Matrix von Übergangswarscheinlichkeiten darstellen. Ein Nachteil dieses Verfahrens ist die kausale Loslösung der Query von dem Trefferset. Das Ranking wird nur von der internen Struktur des Netzes determiniert. Darüber hinaus besteht die Gefahr der Rankmanipulation, zb durch "unsichtbare" Links oder Bannerwerbung. Zuguterletzt ist die rekursive Berechnung über alle Knoten äusserst Zeit/Rechenintensiv, muss aber natürlich nicht erst bei der Query berechnet werden.

HITS betrachtet das Internet ebenfalls as Graphen betrachtet aber kleinere Bereiche davon. Jedem Dokument werden ein Hubwert und ein Authoritywert zugeschrieben. Ein hoher Hubwert kommt durch starke Verlinkung auf Dokumente mit hohem Authoritywert zustande und umgekehrt, diese können als Vektoren betrachtet werden, deren quadrierte Werte über die Anzahl der Vektorelemente normalisiert werden sodass eine Wahrscheinlichkeitsverteilung ensteht. Die Berechnung erfolgt bei der Query kann aber wegen der reduzierten Betrachtung relativ schnell durchgeführt werden. Auch hier besteht eine relative große Gefahr der Manipulation und die thematische Loslösung von der Query selbst

1.2 Das Zipfsche Gesetz

In Katja Bertholds Vortrag ging es um die Verteilung der Wörter in Texten und den Möglichkeiten der Abschätzungen die sich daraus ergeben. Zunächst ging Sie auf biographische Eckdaten zu George Kingsley Zipf ein, welcher ein amerikanischer Linguist war welcher sich vornehmlich mit der statistischen Betrachtung von Sprachen widmete und so einen naturwissenschaftlichen Ansatz zur Lingusitik betrieb. Sodann wies Frau Berthold auf die Grundannahme zur so begründeten

Zipfschen Verteilung hin, nämlich dass Lebewesen immmer den einfachsten Weg nehmen um ein Ziel zu erreichen. So ist zu erklären dass wenige Worte sehr oft vorkommen, da sie ausreichen um sich ausreichend verständlich zu machen. Dann leitet sie aus der Grundformel $n_r \sim 1/r$ die Berechung von $r * n_r \approx k$ ab und zeigt deren Gültigkeit anhand einer indizierten Frequenzliste, generiert aus dem Korpus *Projekt Deutscher Wortschatz*. Zunächst zeigt sie so den konstanten Zusammenhang zwischen der Position in der Frequenzliste und der absoluten Häufigkeit wie auch aus der relativen Häufigkeit. Dese beiden Konstanten benutzt Sie dann um Abschätzungen bezüglich der Mindestanzahl von Wörtern wie auch der Gesamgröße des in einem Korpus benutzen Vokabulars zu machen. Besonders folgenreich sind diese Beobachtungen wenn man die Kontexte von Wörtern betrachtet, so lassen sich verwertbare Aussagen erst ab einer Frequenz von 20 überhaupt erst machen. Die Bedeutung der Zipf-Verteilung lässt sich auch in anderen Domänen als der Linguistik zeigen, so findet man diese auch in der Einwohnerverteilung von Städten oder von Notenwerten in Musikkorpora. Eine Verbesserung der Zipfverteilung lieferte Mandelbrot, der durch zwei zusätzliche Parameter die Anpassung der Formel an Daten in den Grenzbereichen erlaubt.

1.3 Multiple Stringsuche: Verfahren von Aho-Corasick

In Elena Atanasovas Vortrag ging es die Verbesserung des Knut-Morris-Pratt Algorithmus zum Stringmatching für mehrere Suchstrings. Der Knut-Morris-Pratt Algorithmus reduziert die Laufzeit von $\mathcal{O}(n*m)$ auf $\mathcal{O}(n+m)$ für einen Suchstring indem Informationen über brereits durchsuchten Text gespeichert werden sodaß dieser nicht erneut durchsucht werden muss. Bei mehreren k Suchstrings erhöht sich aber der Auwand auch hier auf $\mathcal{O}(n + m * k)$ da der Text für jedes k neu durchlaufen wird. Die Idee des ACA Algorithmus besteht nun darin aus einer Menge von Suchstrings eine Trie zu bauen und diesen als erweiterten endlichen Automaten zu betrachten. Die Erweiterungen bestehen aus Fehlerfunktion f und Ausgabefunktion o . Die Berechnung der Fehlerfunktion erfolgt nach den Regeln: Knoten der Tiefe 1 verlinken auf das Rootelement im Trie, alle anderen verweisen auf einen Knoten welcher vom Rootknoten mit einem Suffix des zu betrachtenden Knotens aus erreichbar ist. Die Ausgabefunktion gibt zu jedem Zustand (Knoten) die Menge der gefundenen Suchstrings aus. Auf diese Weise lässte sich die Laufzeitkomplexität kann so auf $\mathcal{O}(n + m + k)$ reduziert werden. Verwendung findet dieses Verfahren bei *fgrep* in Unix, in der Bildbearbeitung, der Erkennung von bekannten Virenmustern oder auch der Durchsuchung von DNA in der Bioinformatik.

2 Repetitorium

2.1 Ausblick und Empfehlungen für die Zeit nach dem Bachelor

Herr Roth diskutierte verschiedene Optionen bezüglich Veröffentlichungsmöglichkeiten der fertigen Bachelorarbeit auf Kongressen, sowie LMU-interner Bewerbungen auf Wettbewerbe und Stipendien. Insbesondere zeigte er eine Alternative zum Master auf indem er auf Graduiertenprogramme in Deutschland und den USA hinwies, welche den Masterstudiengang verzahnen mit einem Promotionsstudiengang. Darüber hinaus gab er Empfehlungen und Tips zu Bewerbungen auf Internships bei renomierten Forschungszentren und Firmen sowie auf Anstellungen bzw Einsatzmöglichkeiten auf dem freien Markt.