Humboldt-Universität zu Berlin
Institut für Informatik
Lehrstuhl Logik in der Informatik
Prof. Dr. Nicole Schweikardt

11. Juni 2015

Big Data Analytics in Theorie und Praxis

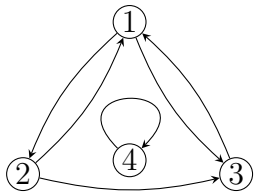
Prof. Johann-Christoph Freytag, Ph.D. und Prof. Dr. Nicole Schweikardt
Sommersemester 2015

Theorie-Übungsblatt 1

Zu bearbeiten bis 18. Juni 2015

Aufgabe 1: (6 + 6 + 6 + 6 = 24 Punkte)

Betrachten Sie den Web-Graph $G = (V, E)$, der aus den vier Webseiten 1, 2, 3 und 4 besteht, die wie in der nebenstehenden Abbildung miteinander verlinkt sind. Benutzen Sie für die folgenden Aufgaben den Dämpfungsfaktor $d := \frac{3}{4}$.



(a) Berechnen Sie die Page-Ranks PR_1, PR_2, PR_3 und PR_4 der vier Webseiten von G bezüglich des Dämpfungsfaktors d.

(b) Stellen Sie für den angegebenen Web-Graph G und den Dämpfungsfaktor d die Page-Rank-Matrix $P(G, d)$ auf.

(c) Sei P die Page-Rank-Matrix $P(G, d)$ aus Teilaufgabe (b). Wir nehmen an, der Zufalls-Surfer startet auf einer der vier Webseiten von G, wobei er jede Webseite gleichwahrscheinlich als Startpunkt wählen kann. Das bedeutet, dass die Anfangsverteilung für den Zufalls-Surfer durch $X^{(0)} := (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ beschrieben wird. Berechnen Sie die Wahrscheinlichkeitsverteilung des Zufalls-Surfers auf den Knoten von G nach einem Schritt (d.h. $X^{(1)}$), nach zwei Schritten (d.h. $X^{(2)}$) und nach drei Schritten (d.h. $X^{(3)}$). Dabei ist $X^{(1)} := X^{(0)} \cdot P$, $X^{(2)} := X^{(1)} \cdot P$ und $X^{(3)} := X^{(2)} \cdot P$.

(d) Gesucht ist ein Web-Graph $G' = (V', E')$ mit vier Webseiten, in dem jede Webseite auf mindestens eine Webseite verlinkt, die nicht sie selbst ist. Zusätzlich soll der Zufalls-Surfer mit der Anfangsverteilung $X^{(0)} := (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ nach einem Schritt in G' genau dieselbe Wahrscheinlichkeitsverteilung erreichen, es soll also $X^{(0)} \cdot P(G', d) = X^{(0)}$ gelten. Geben Sie einen solchen Graphen G' an und weisen Sie nach, dass $X^{(0)} \cdot P(G', d) = X^{(0)}$ gilt.

Aufgabe 2: (9 + 9 + 9 + 7 = 34 Punkte)

(a) Arbeiten Sie die Details dazu aus, wie das Vektor-Matrix-Produkt $Y' = Y \cdot P$ zur Berechnung des Page-Ranks ausgerechnet werden kann (siehe „Zweiter Ansatz“ im Vorlesungsskript). Gehen Sie hierbei insbesondere darauf ein, wie dies mit *MapReduce* gelöst werden kann und analysieren Sie die Güte Ihres Ansatzes.

(b) Das im Skript unter „Zweiter Ansatz“ beschriebene Verfahren schickt alle Teilsummen $Z_{1,t}, \ldots, Z_{\ell,t}$ an denselben Rechner, der dann die Gesamtsumme Z_t berechnen soll. Beschreiben Sie eine alternative Methode, die die Berechnung dieser Summe parallel auf mehreren Rechnern durchführt und dabei die im Rechencluster vorhanden Rechner so gut wie möglich nutzt. Geben Sie auch an, wie dies mit *MapReduce* umgesetzt werden kann und analysieren Sie die Güte Ihres Ansatzes.

(c) Gehen Sie von $n = 10^{12}$ Webseiten, durchschnittlich 20 Links pro Webseite und einem benötigten Speicherplatz von 4 Byte für jeden Eintrag im Eingabe-Vektor Y und jeden Eintrag der Page-Rank-Matrix P aus. Gehen Sie davon aus, dass jeder Clusterrechner einen zur Berechnung verfügbaren Arbeitsspeicher von 20 Gigabyte besitzt. Wie viele Rechner werden im Cluster benötigt, damit die jeweiligen Teilstücke von Y und P (dabei P in der in der Vorlesung behandelten kompakten Speicherung) in den Arbeitsspeicher des jeweiligen Clusterrechners passen? Eine gerundete Rechnung ist ausreichend, und es kann davon ausgegangen werden, dass die Werte in P nicht ungünstig verteilt sind.

(d) Auf Basis der Werte aus Aufgabenteil (c): Vergleichen Sie die Methode zur Berechnung der Summen Z_t aus dem Skript mit Ihrer in Aufgabenteil (b) entwickelten Methode.

Aufgabe 3: (5 + 6 + 6 = 17 Punkte)

Für eine Zahl $n \geq 1$ sei der Web-Graph G mit $n + 2$ Knoten und Knotenmenge

$$V := \{A_1, A_2, B_1, \ldots, B_n\}$$

gegeben, bei dem A_1 und A_2 gegenseitig aufeinander verlinken und jede Seite B_j mit $j \in \{1, \ldots, n\}$ einen Link auf Seite A_1 besitzt.

(a) Welche Werte haben die Page-Ranks der Webseiten von G für den Dämpfungsfaktor $d = 1$?

(b) Welche Werte haben die Page-Ranks der Webseiten von G für Dämpfungsfaktoren $d > 0$?

(c) Welchen Wert nimmt PR_{A_1} in Aufgabenteil (b) für $n \longrightarrow \infty$ an?

Aufgabe 4: ((5 + 5 + 5) + (5 + 5) = 25 Punkte)

Wir nehmen an, das morgige Wetter ließe sich allein aus der Kenntnis des heutigen Wetters vorhersagen. Unter dieser Annahme kann der Wetterverlauf als Markov-Kette modelliert werden. Der Einfachheit halber unterscheiden wir im Folgenden nur die beiden Wetterbedingungen *Regen* und *Sonnenschein*. Das Wetter formt dann eine Markov-Kette mit der Zustandsmenge $Z = \{z_1, z_2\}$, wobei z_1 den Regen und z_2 den Sonnenschein bezeichnet, und der Übergangsmatrix

$$P = \begin{pmatrix} p_{z_1,z_1} & p_{z_1,z_2} \\ p_{z_2,z_1} & p_{z_2,z_2} \end{pmatrix}.$$

Dabei gibt der Wert p_{z_i,z_j} die Wahrscheinlichkeit dafür an, dass auf Wetter im Zustand z_i am folgenden Tag Wetter im Zustand z_j folgt.

Ist die Verteilung des Wetters $X^{(k)} = (X^{(k)}_{z_1}, X^{(k)}_{z_2})$ für einen Tag $k \geq 0$ bekannt, so kann die Verteilung des Wetters am Tag $k + 1$ berechnet werden als $X^{(k+1)} = X^{(k)} \cdot P$.

(a) Für das Berliner Wetter wird oft behauptet, die beste Art der Wettervorhersage bestehe einfach darin, das morgige Wetter als identisch mit dem heutigen zu prognostizieren. Wenn diese Vorhersagemethode mit einer Wahrscheinlichkeit von 3/4 richtig liegt (unabhängig davon, ob aktuell Regen oder Sonnenschein herrscht), dann ergibt sich für die Markov-Kette des Berliner Wetters die Übergangsmatrix

$$P_B = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix}.$$

Wir nehmen an, dass die Markov-Kette für das Berliner Wetter an einem regnerischen Tag beginnt, d. h. es gilt $X_B^{(0)} = (1, 0)$.

(i) Berechnen Sie die Verteilung des Berliner Wetters an Tag drei, d. h. berechnen Sie $X_B^{(3)}$.

(ii) Beweisen Sie durch vollständige Induktion, dass $X_B^{(k)} = \left(\frac{1}{2}(1 + 2^{-k}), \frac{1}{2}(1 - 2^{-k})\right)$ für jedes $k \in \mathbb{N}$ gilt.

(iii) Wie verhält sich $X_B^{(k)}$, wenn k gegen unendlich geht?

(b) Wir betrachten Los Angeles als Beispiel für einen Ort, an dem der Wetterverlauf ein anderer ist als in Berlin. Sei die Übergangsmatrix für das Wetter in Los Angeles gegeben durch

$$P_{LA} = \begin{pmatrix} 1/2 & 1/2 \\ 1/10 & 9/10 \end{pmatrix}.$$

(i) Zeigen Sie, dass die Verteilung $X_{LA} = (1/6, 5/6)$ eine stationäre Verteilung für das Wetter in Los Angeles ist, d. h. zeigen Sie, dass $X_{LA} = X_{LA} \cdot P_{LA}$ ist.

(ii) Geben Sie eine stationäre Verteilung für das Berliner Wetter in Teilaufgabe (a) an, d. h. geben Sie eine Verteilung X_B an mit $X_B \cdot P_B = X_B$.