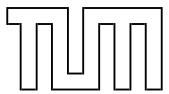
Fortgeschrittene Netzwerk- und Graph-Algorithmen

Prof. Dr. Hanjo Täubig

Lehrstuhl für Effiziente Algorithmen
(Prof. Dr. Ernst W. Mayr)
Institut für Informatik
Technische Universität München

Wintersemester 2010/11



Katz' Status Index

Modell:

- Einfluss entsteht aus Anhängerschaft bzw. aus indirekter Wahl

$\Rightarrow$ wenn A ein Anhänger von B ist und B ist Anhänger von C, dann ist A auch ein (indirekter) Anhänger bzw. Wähler von C

- mit zunehmender Anzahl von Zwischenschritten soll dieser Effekt jedoch abnehmen

$\Rightarrow$ Dämpfungsfaktor $\alpha > 0$

- ungewichteter, gerichteter Graph $G = (V, E)$ ohne Schleifen
- Adjazenzmatrix A, Anzahl Wege der Länge k von j nach i ist $(A^k)_{ji}$
- Status von Knoten i:

$$\mathbf{c}_K(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji}$$

falls die unendliche Summe konvergiert

Katz' Status Index

- in Matrixnotation:

$$\mathbf{c}_K = \sum_{k=1}^{\infty} \alpha^k (A^T)^k \mathbf{1}_n$$

  wobei $\mathbf{1}_n$ der n-dimensionale Vektor ist, dessen Einträge alle 1 sind
- für die Konvergenz muss α beschränkt werden

Satz

Sei A die Adjazenzmatrix eines Graphen G mit größtem Eigenwert λ_1 und $\alpha > 0$, dann gilt:

$$\lambda_1 < \frac{1}{\alpha} \quad \Longleftrightarrow \quad \sum_{k=1}^{\infty} \alpha^k A^k \text{ konvergiert}$$

Katz' Status Index

Unter der Annahme der Konvergenz erhält man die geschlossene Form:

$$\mathbf{c}_K = \sum_{k=1}^{\infty} \alpha^k (A^T)^k \mathbf{1}_n = \left(\left(I - \alpha A^T \right)^{-1} \right) \mathbf{1}_n$$

bzw.

$$\left(I - \alpha A^T \right) \mathbf{c}_K = \mathbf{1}_n$$

$\Rightarrow$ inhomogenes lineares Gleichungssystem

$\mathbf{c}_K(i)$ hängt von den anderen $\mathbf{c}_K(j)$ mit $j \neq i$ ab (Feedback-Zentralität)

Eigenvektor-Zentralität von Bonacich

Phillip Bonacich, 1972:

- gegeben: ungerichteter Graph
  (ungewichtet, zusammenhängend, einfach und ohne Schleifen)

- 3 Ansätze:
  a. Faktoranalyse
  b. Konvergenz einer unendlichen Reihe
  c. Lösung eines linearen Gleichungssystems

- Werte $\mathbf{s}^a$, $\mathbf{s}^b$ und $\mathbf{s}^c$ unterscheiden sich aber nur durch konstante Faktoren

Eigenvektor-Zentralität von Bonacich

Faktoranalyse:

- Interpretiere den Graph als Freundschaftsnetzwerk
- Eine Kante bedeutet Freundschaft zwischen den Endknoten
- gesucht: Vektor $\mathbf{s}^a \in \mathbb{R}^n$, so dass der i-te Eintrag $\mathbf{s}_i^a$ das "Freundschaftspotential" von Knoten i enthält.
- $\mathbf{s}_i^a \mathbf{s}_j^a$ soll möglichst nah bei a_{ij} liegen
- interpretiere das Problem als Minimierung des folgenden Ausdrucks (mit der Methode der kleinsten Quadrate):

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left(\mathbf{s}_i^a \mathbf{s}_j^a - a_{ij} \right)^2$$

Eigenvektor-Zentralität von Bonacich

Unendliche Reihe:

Für ein gegebenes $\lambda_1 \neq 0$ definiere

$$\mathbf{s}^{b_0} = \mathbf{1}_n \quad \text{und} \quad \mathbf{s}^{b_k} = A\frac{\mathbf{s}^{b_{k-1}}}{\lambda_1} = A^k\frac{\mathbf{s}^{b_0}}{\lambda_1^k}$$

Betrachte die Sequenz

$$\mathbf{s}^{b} = \lim_{k\to\infty} \mathbf{s}^{b_k} = \lim_{k\to\infty} A^k\frac{\mathbf{s}^{b_0}}{\lambda_1^k}$$

Satz

Sei $A \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix mit größtem Eigenwert λ_1, dann konvergiert

$$\lim_{k\to\infty} A^k\frac{\mathbf{s}^{b_0}}{\lambda_1^k}$$

gegen einen Eigenvektor von A mit zugehörigem Eigenwert λ_1.

Eigenvektor-Zentralität von Bonacich

Lineares Gleichungssystem:

- Definiere Zentralität jedes Knotens als Summe der Zentralitäten seiner Nachbarn:

$$\mathbf{s}_i^c = \sum_{j=1}^{n} a_{ij}\mathbf{s}_j^c \quad \text{bzw.} \quad \mathbf{s}^c = A\mathbf{s}^c$$

$\Rightarrow$ immer mögliche Lösung: alle Zentralitäten gleich Null

$\Rightarrow$ hat nur eine sinnvolle Lösung, falls $\det(A - I) = 0$ bzw. wenn 1 ein Eigenwert von A ist

- relaxiere das Problem zum Eigenwertproblem:

$$\lambda\mathbf{s} = A\mathbf{s} \quad \text{bzw.} \quad \mathbf{s} = \frac{1}{\lambda}A\mathbf{s}$$

Eigenvektor-Zentralität von Bonacich

Satz

Sei $A \in \mathbb{R}^{n \times n}$ die Adjazenzmatrix eines ungerichteten zusammenhängenden Graphen. Dann

- *ist der größte Eigenwert λ_1 von A ein einfacher Eigenwert,*
- *sind alle Einträge des zu λ_1 gehörenden Eigenvektors ungleich Null und haben das gleiche Vorzeichen.*

(siehe auch Satz von Perron und Frobenius)

Eigenvektor-Zentralität von Bonacich

- Alle 3 Varianten unterscheiden sich nur durch konstante Faktoren.

$\Rightarrow$ Normierte Eigenvektor-Zentralität:

$$\mathbf{c}_{EV} = \frac{|\mathbf{s}|}{||\mathbf{s}||}$$

- im Fall unzusammenhängender Graphen kann man den Ansatz auf jede Zusammenhangskomponente anwenden

Hubbell Index

Charles Hubbell, 1965:

- gegeben: einfacher gewichteter gerichteter Graph (darf Schleifen enthalten, aber keine Multikanten)
- Adjazenzmatrix ist hier u.U. asymmetrisch
- Idee: (ähnlich wie bei Bonacich)
  Wert eines Knotens v hängt von der Summe der Werte seiner Nachbarn w, gewichtet mit dem jeweiligen Kantengewicht ab
- Es soll also gelten:

$$\mathbf{e} = W\mathbf{e}$$

- "Exogener Input" E (externe Information zu jedem Knoten) wird eingeführt, um das System lösbar zu machen
- im Falle des Fehlens: $\mathbf{E} = \mathbf{1}$

$\Rightarrow$ Gleichung

$$\mathbf{s} = \mathbf{E} + W\mathbf{s}$$

Hubbell Index

- umgeformt:

$$\mathbf{s} = (I - W)^{-1}\mathbf{E}$$

- System hat eine Lösung, falls $(I - W)$ invertierbar ist
- Da

$$(I - W)^{-1} = \sum_{k=0}^{\infty} W^k$$

$\Rightarrow$ äquivalent zur Konvergenz der geometrischen Reihe

- Reihe konvergiert gegen $(I - W)^{-1}$ gdw. der größte Eigenwert λ_1 von W kleiner als 1 ist
- Lösung des Gleichungssystems heißt Hubbell-Index

Bonacich's Verhandlungszentralität

Phillipp Bonacich, 1987: Verhandlungszentralität (bargaining centrality)

- bisherige Feedback-Zentralitäten basieren auf positiver Rückkopplung: die Zentralität eines Knotens ist umso höher, je höher die Zentralität seiner Nachbarn ist
- Folgender Ansatz ermöglicht sowohl eine Modellierung dieses positiven Einflusses als auch eine Modellierung mit negativem Einfluss wie z.B. in Verhandlungssituationen
- Verhandlungssituation:
  Ein Knoten ist stark, wenn er mit Knoten verbunden ist, die keine anderen Optionen haben und deshalb schwach sind.

Bonacich's Verhandlungszentralität

- gegeben: ungewichteter gerichteter Graph $G = (V, E)$ ohne Schleifen

$\Rightarrow$ Adjazenzmatrix ist im Allgemeinen nicht symmetrisch und enthält nur Nullen und Einsen

- Definition:

$$c_{\alpha,\beta}(i) = \sum_{j=1}^{n} (\alpha + \beta \cdot c_{\alpha,\beta}(j)) a_{ij}$$

- in Matrixnotation:

$$\mathbf{c}_{\alpha,\beta} = \alpha(I - \beta A)^{-1} A\mathbf{1}$$

$\Rightarrow$ α ist nur ein Skalierungsfaktor
Bonacich schägt vor, α so zu wählen, dass $\sum_{i=1}^{n} c_{\alpha,\beta}(i)^2 = n$

- β: positiver oder negativer Einfluss der Werte der Nachbarn
  $\beta = 0$: Zentralität proportional zum Grad
  $\beta < 0$ kann zu negativen Zentralitätswerten führen
- je größer $|\beta|$, desto größer der Einfluss der Netzwerkstruktur

Bonacich's Verhandlungszentralität

- Gleichungssystem ist lösbar, falls $(I - \beta A)$ invertierbar ist
- Nach folgendem Theorem existiert dieses Inverse, falls kein Eigenwert von A den Wert $1/\beta$ hat.

Satz

Seien $\lambda_1, \ldots, \lambda_n$ die Eigenwerte einer Matrix $M \in \mathbb{R}^{n \times n}$.

Dann gilt:

$$(I - M) \text{ ist invertierbar} \iff \forall i \in \{1, \ldots, n\} : \lambda_i \neq 1$$

Webgraph

- gerichteter Graph $G = (V, E)$
- Webseiten entsprechen den Knoten
- Links zwischen Webseiten entsprechen den gerichteten Kanten

Random Surfer Model

- Modellierung des Verhaltens eines Websurfers als Random Walk auf dem Webgraph:

$$\Pr[X_{t+1} = v \mid X_t = u] = \begin{cases} \frac{1}{d^+(u)}, & \text{falls } (u,v) \in E \\ 0, & \text{sonst} \end{cases}$$

- in jedem Schritt überschreitet der Random Walk zufällig eine der ausgehenden Kanten des aktuellen Knotens
- nur wohldefiniert, falls $\forall v \in V: \ d^+(v) \geq 1$
- in diesem Fall ist die Transitionsmatrix die stochastische $n \times n$-Matrix $T = (t_{i,j})$ mit $t_{i,j} = 1/d^+(i)$ falls $(i,j) \in E$ und $t_{i,j} = 0$ sonst

Random Surfer Model

- Webgraph ist nicht stark zusammenhängend

$\Rightarrow$ zugrundliegende Transitionsmatrix ist nicht irreduzibel

- Es existieren Senken
  (Webseiten ohne Links / Knoten ohne ausgehende Kanten)

$\Rightarrow$ Transitionsmatrix ist nicht einmal stochastisch

$\Rightarrow$ Matrix muss modifiziert werden, damit die entsprechende Markov-Kette in eine stationäre Verteilung konvergiert

- Um die Transitionsmatrix T stochastisch zu machen:
  Annahme, dass im Fall von Senken der Surfer zu einer zufälligen Seite springt:

$$t'_{i.j} = \begin{cases} \frac{1}{d^+(i)}, & \text{falls } (i,j) \in E \\ \frac{1}{n}, & \text{falls } d^+(i) = 0 \\ 0, & \text{sonst} \end{cases}$$

Random Surfer Model

$\Rightarrow$ Transitionsmatrix T' ist stochastisch, aber nicht irreduzibel, Berechnung einer stationären Verteilung u.U. nicht möglich

- Sei $E = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ die Random-Jump-Matrix, in der alle Einträge $\frac{1}{n}$ sind.
- Diese wird einfach in gewichteter Form zu T' addiert:

$$T'' = \alpha T' + (1 - \alpha)E$$

- α wird aus dem Intervall $[0, 1)$ gewählt und entspricht der Wahrscheinlichkeit, entweder einem Link auf der Seite zu folgen (T') oder zu einer zufälligen Seite zu springen (E)
- T'' ist stochastisch
  (denn T' und E sind stochastisch und in jeder Zeile bzw. Spalte wird ein α-Anteil mit einem $(1 - \alpha)$-Anteil der Gesamtwahrscheinlichkeit 1 addiert)
- T'' ist irreduzibel (alle Wahrscheinlichkeiten > 0)

Power Iteration

Algorithmus 4 : Power Method

Input : Matrix $A \in \mathbb{R}^{n \times n}$ und Vektor $||\vec{q}^{(0)}||_2 = 1$
Output : Betragsmäßig größter Eigenwert $\lambda^{(k)}$
und korrespondierender Eigenvektor $\vec{q}^{(k)}$

$k := 1$;
repeat
  $\vec{z}^{(k)} := A\vec{q}^{(k-1)}$;
  $\vec{q}^{(k)} := \vec{z}^{(k)} / ||\vec{z}^{(k)}||_2$;
  $\lambda^{(k)} := (\vec{q}^{(k)})^T A\vec{q}^{(k)}$;
  $k := k + 1$;
until $\lambda^{(k)}$ *und* $\vec{q}^{(k)}$ *sind akzeptabel approximiert* ;

Power Iteration

- konvergiert garantiert, wenn A einen dominanten Eigenwert λ_1 hat, d.h. $\forall i \in \{2, \ldots, n\} : \ |\lambda_1| > |\lambda_i|$.
- konvergiert auch, wenn die Matrix symmetrisch ist
- Konvergenzgeschwindigkeit hängt vom Verhältnis $\frac{|\lambda_2|}{|\lambda_1|}$ ab
- Approximationsfehler sinkt mit

$$\mathcal{O}\left(\left(\frac{|\lambda_2|}{|\lambda_1|}\right)^k\right)$$

- erfordert lediglich Matrix-Vektor-Multiplikationen

$\Rightarrow$ ist besonders für große Matrizen geeignet
Für eine Iteration muss nur ein Scan über die Matrix laufen

$\Rightarrow$ auch effizient, falls nicht die ganze Matrix in den Hauptspeicher passt

Stationäre Verteilung im Random Surfer Model

- stationäre Verteilung π'' für Transitionsmatrix T'' kann mit Power Iteration berechnet werden
- Durch Modifikation von E können die Zufallssprünge in Richtung personalisierter Surfergewichtungen verändert werden.