Wie Google Webseiten bewertet

François Bry

Heutige Vorlesung

1. Einleitung
2. Graphen und Matrizen
3. Erste Idee: Ranking als Eigenvektor
4. Fragen: Exisitiert der Eigenvektor? Usw.
5. Zweite Idee: Die Google-Matrix
6. Dritte Idee: Die Power-Methode
7. Schlussbemerkungen

1. Einleitung

Die Suche im Web benötigt **Suchmaschinen**, weil:

- das Web keine Verwaltung, folglich keine bekannte Struktur hat;
- das Web für sehr unterschiedlichen Zwecke verwendet wird;
- viele Daten im Web kurzlebig sind;
- Es Sichtbarkeitsverzerrungen durch "*link farms*", "*Google bombs*", "*spamdexing*", etc. gibt;
- das Web riesig ist: über 12 Milliarden Dokumenten in 2009.

1. Einleitung

1997 wurde klar: bekannte Ansätze der IR (Information Retrieval) reichen fürs Web nicht aus.

Zwei Ansätze zum **Ranking von Webseiten** wurden vorgeschlagen:

- **HITS** (Hypertext Induced Topic Search) durch **Jon Kleinberg**
- **PageRank** durch **Sergey Brin** and **Larry Page**

2. Graphen und Matrizen

Gerichteter Graph

Adjazenzmatrix A eines gerichteten Graphen

- Summe der Zeile i: Anzahl der ausgehenden Kanten des Knotens i
- Summe der Spalte i: Anzahl der eingehenden Kanten des Knotens i
- $A \cdot 1$: Komponente i ist die Anzahl der *aus*gehenden Kanten des Knotens i

2. Graphen und Matrizen

Transponierte A^T der Adjazenzmatrix A eines gerichteten Graphen

- $A^T \cdot 1$: Komponente i ist die Anzahl der *ein*gehenden Kanten des Knotens i

H **Hyperlink-Matrix des Webs:**

Transponierte der Adjazenzmatrix des Hyperlink-Graphens, wobei Kanten einer Seite zu sich selbst nicht berücksichtigt werden

2. Graphen und Matrizen

H‘ **veränderte Hyerlink-Matrix des Webs:**

Die Komponenten einer Spalte, die nicht nur Nullen enthält, werden durch die Spaltensumme dividiert.

Die Summe einer Spalte ist also 0 oder 1.

Bedeutung:

Verlinkt eine Webseite auf n weiteren Seiten, so gibt sie jeder dieser n Webseiten 1/n ihrer Wichtigkeit ab.

H‘.1: Komponente i ist die von der Webseite i über die Links vererbte Wichtigkeit

3. Erste Idee: Ranking als Eigenvektor

Das gesuchte Ranking ist ein Vektor mit realen und positiven Komponenten und ein **Eigenvektor**, d.h. ein Vektor v, so dass:

$$H' \cdot v = v$$

Begründung: v gibt die Wichtigkeiten der Webseiten richtig an, weil nichts mehr vererbt werden kann:

$$H' \cdot H' \cdot v = H' \cdot v = v$$

Bemerkung: ist v Lösung, so auch k . v. Man kann sich auf Vektoren der Länge 1 einschränken.

3. Erste Idee: Ranking als Eigenvektor

Auslegung:

- Eine Webseite ist desto wichtiger, dass sie von wichtigen Webseiten angezeigt wird.
- Die Wichtigkeit einer Webseite S ist die Summe der Wichtigkeiten der Webseiten, die auf S zeigen.
- Wenn eine Seite S auf mehrere weiteren Seiten zeigt, dann wird die Wichtigkeit von S unter den Webseiten (in gleichen Teilen) geteilt, worauf S zeigt.

3. Erste Idee: Ranking als Eigenvektor

Diese erste Idee war nicht ganz neu:

- **Input-output-Analyse** von Wassily Leontiev (Nobel-Preis von 1973): Matrix aber kein Eigenvektor
- **Kreditrisikoschätzung** bei Banken: Eigenvektor

4. Existiert der Eigenvektor?

Präzisierung der Frage:

- Gibt es eine Lösung v der Gleichung

$$H' \cdot v = v$$

mit realen und positiven Komponenten?

- Falls ja ist diese Lösung eindeutig?

Sonst wären die Lösungen nutzlos.

4. Existiert der Eigenvektor?

Satz von Perron-Frobenius:

Wenn A die Adjazenzmatrix von einem **stark-verbundenen** Graph ist, dann hat die Gleichung

$$A \cdot v = v$$

eine eindeutige Lösung v mit realen und positiven Komponenten.

Diese Lösung heißt **Perron-Vektor** von A.

4. Existiert der Eigenvektor?

Offensichtlich ist der Hyperlink-Graph des Webs nicht stark-verbunden:

- Einige Webseiten sind gar nicht angelinked. Eine Suchmaschine kann die meisten davon ignorieren.
- Einige Webseiten sind **Senken**, d.h. haben keine ausgehende Links.

In H' entsprechen die Senken Spalten, die nur Nullen beinhalten.

5. Zweite Idee : Die Google-Matrix

Zur Beseitigung der Senken:

Ersetze jede Null in einer Spalte von H', die nur Nullen enthält, durch 1/n (wobei n die Anzahl der Webseiten ist).

Zur Starkverbundenheit des Graphen:

$$G = c\,H' + (1 - c)\,E$$

mit $0 < c < 1$ und E Matrix mit identischen Spalten $(p_1, \ldots, p_n)$ mit $0 < p_i < 1$ und Summe der $p_i = 1$.

5. Zweite Idee : Die Google-Matrix

Auslegung der Google-Matrix als zufällige Wanderung durch das Web

- Mit Wahrscheinlichkeit c wird ein Link von der Webseite verfolgt, wo man sich befindet.
- Mit Wahrscheinlichkeit (1 - c) wird auf irgendeine Webseite gesprungen.

5. Zweite Idee : Die Google-Matrix

p_i ist der **Personalisierungswert** der Webseite i:

- hoch für *whitehouse.gov* oder *lmu.de*
- niedrig für *myhomepage.de*

6. Dritte Idee : Die Power-Methode

Sind A die Adjazenzmatrix eines stark-verbundenen Graphen und w ein Vektor w mit realen und positiven Komponenten, so ist der Limes von $G^k.w$ für k gegen unendlich der Perron-Vektor von A.

Effizient wenn A schwach besetzt ist. G ist es nicht aber $G^k.w = c\ H'^k.w + (1 - c)\ E^k.w$ und die letzte Komponente der Summe ist leicht zu berechnen.

6. Dritte Idee : Die Power-Methode

Die Power-Methode so implementieren, dass

- nur Vektoren aber keine Matrizen Zwischenergebnisse sind;
- parallel berechnet wird.

7. Schlussbemerkungen

Was sind Vektoren und Matrizen? Wie kann man damit rechnen?

Antwort in der Vorlesung *Lineare Algebra*

Was kann man mit Matrizen und Vektoren noch tun?

Sehr viel:

- Algebraisierung der Geometrie
- Clustering, Ranking in Data Mining, IR
- Social Network Analysis
- Risikoschäzung
- etc.

Siehe meine Vorlesung *Web-Informationssysteme*