TU München, Fakultät für Informatik
Lehrstuhl III: Datenbanksysteme
Prof. Alfons Kemper, Ph.D.

Übung zur Vorlesung *Einsatz und Realisierung von Datenbanksystemen* im SoSe19

Maximilian {Bandle, Schüle} (i3erdb@in.tum.de)
http://db.in.tum.de/teaching/ss19/impldb/

Blatt Nr. 12

Hausaufgabe 1

Berechne für die folgenden drei Dokumente die TF-IDF Werte. Dabei sind alle Worte relevant.

1. ERDB macht echt viel Spaß
2. Die Klausur ist sicher machbar
3. Wir wünschen euch allen viel Erfolg bei der ERDB Klausur

Welches Ranking ergibt sich für die Anfrage: "ERDB Klausur"? Berechne die Werte auf 3 Nachkommastellen genau.

Hilfswerte (gerundet):

$$log(3) = 1/2$$
$$log(2,5) = 2/5$$
$$log(2) = 1/3$$
$$log(1,5) = 1/6$$
$$log(1) = 0$$

ERDB IDF = log(3/2) = 1/6

	D1	D2	D3
TF	1/5	0	1/10
TF-IDF	1/30	0	1/60

Klausur IDF = log(3/2) = 1/6

	D1	D2	D3
TF	0	1/5	1/10
TF-IDF	0	1/30	1/60

Ranking:
D1: 1/30 = 0,033
D2: 1/30 = 0,033
D3: 1/60 + 1/60 = 1/30 = 1/10 * 1/3 = 0.033

Hausaufgabe 2

In dem in Abbildung 1 gezeigten Netzwerk von Web-Seiten wird ein weiteres Beispiel für einen Webgraphen gezeigt. Berechnen Sie, für das in Abbildung gezeigte Netzwerk, den PageRank nach 2 Iterationen. Nutzen Sie $1/|V|$ als Anfangswert für den PageRank.

	A	B	C	D	E
PR Iter 0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
PR Iter 1	$\frac{1}{10}$	$\frac{1}{6}$	$\frac{4}{15}$	$\frac{4}{15}$	$\frac{1}{5}$
PR Iter 2	$\frac{1}{10}$	$\frac{2}{15}$	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{4}{15}$

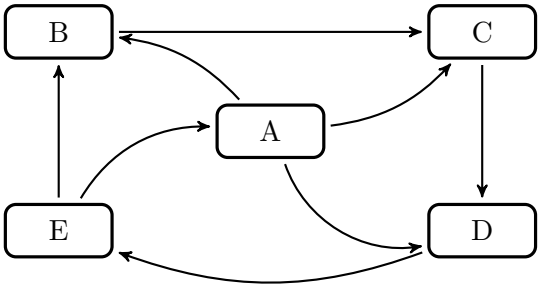

Abbildung 1: Ein weiterer Webgraph.

Hausaufgabe 3

In Abbildung 2 sind drei Graphen gegeben, ein sternförmiger, eine Clique und ein linear angeordneter.

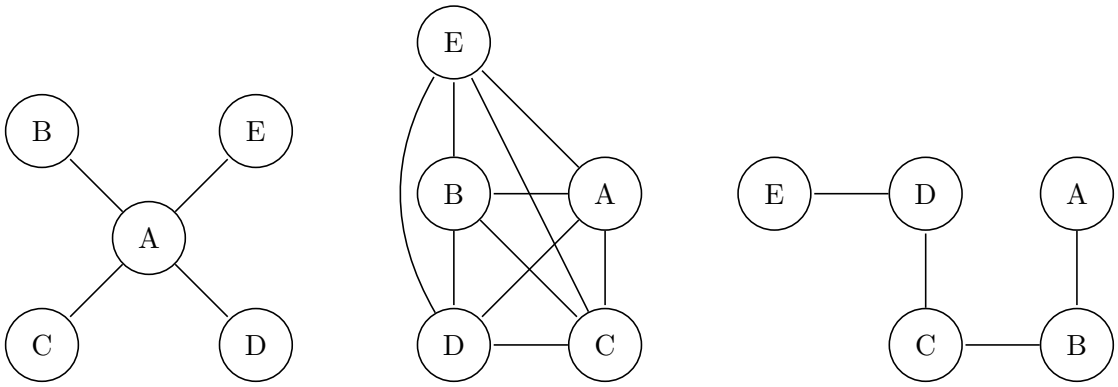

Abbildung 2: Star, Clique und Linie.

1. Berechnen Sie den Grad der Knoten für jeden der Graphen.
   Stern: $C_D(A) = 4, C_D(B) = 1, C_D(C) = 1, C_D(D) = 1, C_D(E) = 1$
   Clique: $C_D(A) = 4, C_D(B) = 4, C_D(C) = 4, C_D(D) = 4, C_D(E) = 4$
   linear: $C_D(A) = 1, C_D(B) = 2, C_D(C) = 2, C_D(D) = 2, C_D(E) = 1$
2. Berechnen Sie die Verbindungszentralität $C_D(G)$ der drei Graphen, sowie deren normierte Verbindungszentralität $C'_D(G)$.

$$
\begin{aligned}
C_D(G*) &= \sum_{v \in V} [C_D(v*) - C_D(v)] = \sum_{v \in V} [C_D(A) - C_D(v)] = (|V| - 2)(|V| - 1) \\
&= (4-4) + (4-1) + (4-1) + (4-1) + (4-1) = 12 = (5-2)(5-1) \\
C'_D(G*) &= \frac{C_D(G*)}{C_D(G*)} = 1 \\
C_D(G_{Clique}) &= \sum_{v \in V} [C_D(v*) - C_D(v)] = 0 \\
C'_D(G_{Clique}) &= \frac{C_D(G_{Clique})}{C_D(G*)} = 0/12 = 0 \\
C_D(G_{lin}) &= \sum_{v \in V} [C_D(v*) - C_D(v)] = \sum_{v \in V} [C_D(B) - C_D(v)] = \\
&= (2-1) + (2-2) + (2-2) + (2-2) + (2-1) = 2 \\
C'_D(G_{lin}) &= \frac{C_D(G_{lin})}{C_D(G*)} = 2/12
\end{aligned}
$$

3. Berechnen Sie die Nähe-Zentralität $H_G(v)$ für jeden Knoten der drei Graphen.

Für $G*$:

$$\begin{aligned}
H_{G*}(A) &= \sum_{y \neq A \in V} [\frac{1}{d(y,A)}] = \frac{1}{d(B,A)} + \frac{1}{d(C,A)} + \frac{1}{d(D,A)} + \frac{1}{d(E,A)} \\
&= \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} = 4 \\
H_{G*}(B) &= \sum_{y \neq B \in V} [\frac{1}{d(y,B}] = \frac{1}{d(A,B)} + \frac{1}{d(C,B)} + \frac{1}{d(D,B)} + \frac{1}{d(E,B)} \\
&= \frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2.5 \\
H_{G*}(C) &= H_{G*}(D) = H_{G*}(E) = 2.5 \text{ analog.}
\end{aligned}$$

Für G_{Clique}:

$$\begin{aligned}
H_{G_{Clique}}(A) &= \sum_{y \neq A \in V} [\frac{1}{d(y,A)}] = \frac{1}{d(B,A)} + \frac{1}{d(C,A)} + \frac{1}{d(D,A)} + \frac{1}{d(E,A)} \\
&= \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} = 4 \\
H_{G_{Clique}}(B) &= H_{G_{Clique}}(C) = H_{G_{Clique}}(D) = H_{G_{Clique}}(E) = 4 \text{ analog.}
\end{aligned}$$

Für G_{linear}:

$$\begin{aligned}
H_{G_{linear}}(A) &= \sum_{y \neq A \in V} [\frac{1}{d(y,A)}] = \frac{1}{d(B,A)} + \frac{1}{d(C,A)} + \frac{1}{d(D,A)} + \frac{1}{d(E,A)} \\
&= \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{25}{12} \\
H_{G_{linear}}(B) &= \sum_{y \neq B \in V} [\frac{1}{d(y,B)}] = \frac{1}{d(A,B)} + \frac{1}{d(C,B)} + \frac{1}{d(D,B)} + \frac{1}{d(E,B)} \\
&= \frac{1}{1} + \frac{1}{1} + \frac{1}{2} + \frac{1}{3} = \frac{17}{6} \\
H_{G_{linear}}(C) &= \sum_{y \neq C \in V} [\frac{1}{d(y,C)}] = \frac{1}{d(A,C)} + \frac{1}{d(B,C)} + \frac{1}{d(D,C)} + \frac{1}{d(E,C)} \\
&= \frac{1}{2} + \frac{1}{1} + \frac{1}{1} + \frac{1}{2} = 3 \\
H_{G_{linear}}(D) &= \sum_{y \neq D \in V} [\frac{1}{d(y,D)}] = \frac{1}{d(A,D)} + \frac{1}{d(B,D)} + \frac{1}{d(C,D)} + \frac{1}{d(E,D)} \\
&= \frac{1}{3} + \frac{1}{2} + \frac{1}{1} + \frac{1}{1} = \frac{17}{6} \\
H_{G_{linear}}(E) &= \sum_{y \neq E \in V} [\frac{1}{d(y,E)}] = \frac{1}{d(A,E)} + \frac{1}{d(B,E)} + \frac{1}{d(C,E)} + \frac{1}{d(D,A)} \\
&= \frac{1}{4} + \frac{1}{3} + \frac{1}{2} + \frac{1}{1} = \frac{25}{12}
\end{aligned}$$

4. Berechnen Sie die Pfad-Zentralität $H_G(v)$ für jeden Knoten der drei Graphen.

Für $G*$:

$$\begin{aligned}
C_{G*}(A) &= \sum_{s\neq A\neq t\in V} [\frac{\sigma_{st}(v)}{\sigma_{st}}] \\
&= \frac{\sigma_{BC}(v)}{\sigma_{BC}} + \frac{\sigma_{BD}(v)}{\sigma_{BD}} + \frac{\sigma_{BE}(v)}{\sigma_{BE}} + \frac{\sigma_{CD}(v)}{\sigma_{CD}} + \frac{\sigma_{CE}(v)}{\sigma_{CE}} + \frac{\sigma_{DE}(v)}{\sigma_{DE}} \\
&= \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} = 6 \\
C_{G*}(B) &= \sum_{s\neq B\neq t\in V} [\frac{\sigma_{st}(v)}{\sigma_{st}}] \\
&= \frac{\sigma_{AC}(v)}{\sigma_{AC}} + \frac{\sigma_{AD}(v)}{\sigma_{AD}} + \frac{\sigma_{AE}(v)}{\sigma_{AE}} + \frac{\sigma_{CD}(v)}{\sigma_{CD}} + \frac{\sigma_{CE}(v)}{\sigma_{CE}} + \frac{\sigma_{DE}(v)}{\sigma_{DE}} \\
&= \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} = 0 \\
C_{G*}(C) &= C_{G*}(D) = C_{G*}(E) = 0 \text{ analog.}
\end{aligned}$$

Für G_{Clique}:

$$\begin{aligned}
C_{G*}(A) &= \sum_{s\neq A\neq t\in V} [\frac{\sigma_{st}(v)}{\sigma_{st}}] \\
&= \frac{\sigma_{BC}(v)}{\sigma_{BC}} + \frac{\sigma_{BD}(v)}{\sigma_{BD}} + \frac{\sigma_{BE}(v)}{\sigma_{BE}} + \frac{\sigma_{CD}(v)}{\sigma_{CD}} + \frac{\sigma_{CE}(v)}{\sigma_{CE}} + \frac{\sigma_{DE}(v)}{\sigma_{DE}} \\
&= \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} = 0 \\
C_{G*}(B) &= C_{G*}(C) = C_{G*}(D) = C_{G*}(E) = 0 \text{ analog.}
\end{aligned}$$

Für G_{linear}:

$$\begin{aligned}
C_{G*}(A) &= \sum_{s\neq A\neq t\in V} [\frac{\sigma_{st}(v)}{\sigma_{st}}] \\
&= \frac{\sigma_{BC}(v)}{\sigma_{BC}} + \frac{\sigma_{BD}(v)}{\sigma_{BD}} + \frac{\sigma_{BE}(v)}{\sigma_{BE}} + \frac{\sigma_{CD}(v)}{\sigma_{CD}} + \frac{\sigma_{CE}(v)}{\sigma_{CE}} + \frac{\sigma_{DE}(v)}{\sigma_{DE}} \\
&= \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} = 0 \\
C_{G*}(B) &= \sum_{s\neq B\neq t\in V} [\frac{\sigma_{st}(v)}{\sigma_{st}}] \\
&= \frac{\sigma_{AC}(v)}{\sigma_{AC}} + \frac{\sigma_{AD}(v)}{\sigma_{AD}} + \frac{\sigma_{AE}(v)}{\sigma_{AE}} + \frac{\sigma_{CD}(v)}{\sigma_{CD}} + \frac{\sigma_{CE}(v)}{\sigma_{CE}} + \frac{\sigma_{DE}(v)}{\sigma_{DE}} \\
&= \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{1} = 3 \\
C_{G*}(C) &= \sum_{s\neq C\neq t\in V} [\frac{\sigma_{st}(v)}{\sigma_{st}}] \\
&= \frac{\sigma_{AB}(v)}{\sigma_{AB}} + \frac{\sigma_{AD}(v)}{\sigma_{AD}} + \frac{\sigma_{AE}(v)}{\sigma_{AE}} + \frac{\sigma_{BD}(v)}{\sigma_{BD}} + \frac{\sigma_{BE}(v)}{\sigma_{BE}} + \frac{\sigma_{DE}(v)}{\sigma_{DE}} \\
&= \frac{0}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{0}{1} = 4 \\
C_{G*}(D) &= \sum_{s\neq D\neq t\in V} [\frac{\sigma_{st}(v)}{\sigma_{st}}] \\
&= \frac{\sigma_{AB}(v)}{\sigma_{AB}} + \frac{\sigma_{AC}(v)}{\sigma_{AC}} + \frac{\sigma_{AE}(v)}{\sigma_{AE}} + \frac{\sigma_{BC}(v)}{\sigma_{BC}} + \frac{\sigma_{BE}(v)}{\sigma_{BE}} + \frac{\sigma_{CE}(v)}{\sigma_{CE}} \\
&= \frac{0}{1} + \frac{0}{1} + \frac{1}{1} + \frac{0}{1} + \frac{1}{1} + \frac{1}{1} = 3
\end{aligned}$$

Hausaufgabe 4

Beantworten Sie folgende Fragen auf Basis des Schemas des TPC-H-Benchmarks.

1. Bestimmen Sie das Handelsvolumen des US-Automobilmarktes. Hierbei ist `l_quantity * l_extendedprice` das Handelsvolumen pro Einzelposten in `lineitem`. Das Handelsvolumen des US-Automobilmarktes ist die Summe des Handelsvolumens aller Einzelposten, bei denen der Kunde aus den USA stammt und dessen Marktsektor (`mktsegment`) die Automobilindustrie (`'AUTOMOBILE'`) ist.

```
select sum(l_quantity * l_extendedprice)
from lineitem, orders, customer, nation
where l_orderkey = o_orderkey and o_custkey = c_custkey and
      c_nationkey = n_nationkey and c_mktsegment = 'AUTOMOBILE' and
      n_name = 'UNITED␣STATES'
```

2. Bestimmen Sie das Handelsvolumen der inländischen Kunden. (Nation des Zulieferes gleich des des Kunden).

```
select sum(l_quantity * l_extendedprice)
from lineitem, orders, customer, supplier
where l_orderkey = o_orderkey and o_custkey = c_custkey and
      l_suppkey = s_suppkey and s_nationkey = c_nationkey
```