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

Übung zur Vorlesung
Einsatz und Realisierung von Datenbanksystemen im SoSe16

Moritz Kaufmann (moritz.kaufmann@tum.de)
http://db.in.tum.de/teaching/ss16/impldb/

Blatt Nr. 11

Hausaufgabe 1

Berechnen Sie für folgende drei Dokumente die TF-IDF-Werte:

1. „Beim Fußball dauert ein Spiel neunzig Minuten – und am Ende gewinnen die Deutschen“
2. „Beim Fußball muss das Runde (der Ball) in das Eckige (das Tor)“
3. „Nie war ein Tor so wertvoll wie jetzt“

Welches Ranking ergibt sich gemäß der Relevanzwerte für die Anfrage: „Fußball“ ∧ „Tor“. Zur Ermittlung des TF Wertes gehen sie davon aus, dass alle Wörter eines Dokuments *interessant* sind?
Zur Berechnung des Rankings reicht es nur die TF-IDF-Werte von *Fußball* und *Tor* zu berechnen.

Fußball IDF: 0.176

	Dokument 1	Dokument 2	Dokument 3
TF	0.077	0.083	0
TF-IDF	0.014	0.015	0

Tor IDF: 0.176

	Dokument 1	Dokument 2	Dokument 3
TF	0	0.083	0.125
TF-IDF	0	0.015	0.022

Ranking Dokument 2: 0.029
Dokument 3: 0.022
Dokument 1: 0.014

Hausaufgabe 2

Ihr System ist durch eine DDoS Attacke blockiert. Schreiben sie eine Streaming SQL Anfrage, die IP's identifiziert über die das System angegriffen wird. Als Metrik soll dabei das Verhältnis von neu geöffneten Verbindungen im Vergleich zur Anzahl der Datenpakete pro IP in einem 5 Minuten Fenster nicht über 20% liegen. Das System empfängt dabei folgenden Eventtyp:

```
Packet [IP,TYPE,DATE]
```

TYPE kann dabei einer der folgenden Werte sein:
NEW_CONNECTION, *DATA*, *CLOSE_CONNECTION*.

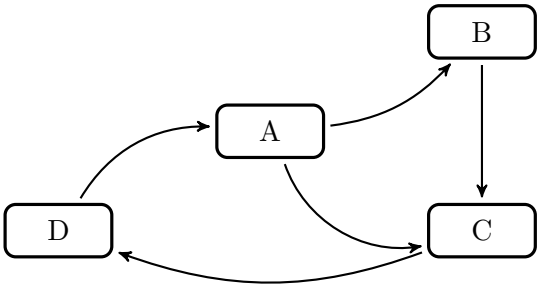

Abbildung 1: Ein kleiner Webgraph.

Lösung:

```
select CIP
from (select p1.IP as CIP, count(*) as connections
      from Packet p1 window(range 5 minutes)
      where p1.TYPE='NEW\_CONNECTION'
      group by p1.IP) c,
     LEFT JOIN (select p1.IP as DIP, count(*) as data
      from Packet p1 window(range 5 minutes)
      where p1.TYPE='DATA'
      group by p1.IP) d ON c.CIP = d.DIP
where data IS NULL OR (connections/data)>0.2
```

Hausaufgabe 3 In Abbildung 1 gezeigte Netzwerk von Web-Seiten wird ein kleines Beispiel für einen Webgraphen gezeigt. Lösen sie folgende Aufgaben.

1. Berechnen Sie, für das in Abbildung, den PageRank, sowie die HITS-Werte nach 2 Iterationen. Nutzen Sie $1/|V|$ als Anfangswert für den PageRank und 1 für HITS. $a = 0.1$
2. Formulieren sie eine Iteration des Pagerank Algorithmus in SQL. Der Graph ist dabei in der Tabelle *edges(VFrom,To)* gespeichert, die aktuelle PageRank Gewichtung in der Tabelle *pagerank(Vertex,Weight*. Sie können die Anzahl der Knoten als Konstante annehmen, z.B. 1000.
3. Formulieren sie die SQL Anfrage nun als rekursive SQL Anfrage um.

HITS: Iteration 1

	A	B	C	D
Hub	2	1	1	1
Auth (vorläufig)	1	2	3	1
Auth (normalisiert)	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{3}{3}$	$\frac{1}{3}$

HITS: Iteration 2

	A	B	C	D
Hub	$\frac{5}{3}$	1	$\frac{1}{3}$	$\frac{1}{3}$
Auth (vorläufig)	$\frac{1}{3}$	$\frac{5}{3}$	$\frac{8}{3}$	$\frac{1}{3}$
Auth (normalisiert)	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{8}{8}$	$\frac{1}{8}$

PageRank

1.

	A	B	C	D
PR Iter 1	$\frac{1}{4}$	$\frac{11}{80}$	$\frac{29}{80}$	$\frac{1}{4}$
PR Iter 2	$\frac{1}{4}$	$\frac{11}{80}$	0.2613	0.3513

2.
```
select To, ((0.1/(select count(*) from pagerank))+0.9*sum(Beitrag))
from (
    select e.To, p.Weight /
        (select count(*) from edges x where x.VFrom=e.VFrom) as Beitrag
    from edges e, pagerank p
    where e.VFrom=p.Vertex
    group by e.VFrom
) i
group by e.To
```

3.
```
with recursive pagerank(Vertex,Weight, Iteration) as (
    (select e.to, 0.001, 0 from edges union select e.VFrom, 0.001 from edges)
    union all
    (select To, (0.0001+0.9*sum(Beitrag)), Iteration + 1
    from (
        select e.To, p.Weight /
            (select count(*) from edges x where x.VFrom=e.VFrom) as Beitrag,
            p.Iteration as Iteration
        from edges e, pagerank p
        where e.VFrom=p.Vertex
    ) i
    group by i.To, i.Iteration)
```