Universität des Saarlandes
FR 6.2 Informatik

Kurt Mehlhorn und Corinna Coupette

WiSe 2021/22

Übungen zu Ideen der Informatik

https://www.mpi-inf.mpg.de/departments/algorithms-complexity/teaching/winter21/ideen/

Blatt 5: Websuche **Abgabeschluss: 29.11.2021**

Aufgabe 1 Vorkommenslisten *(10 Punkte)*

Betrachten Sie die folgenden Textdokumente:

- T_1: Der Bundestag hat das neue Infektionsschutzgesetz gebilligt.
- T_2: Der neue Bundestag hat mehr Abgeordnete als je zuvor.
- T_3: Was soll das neue Infektionsschutzgesetz ändern?

a) Erstellen Sie die Vorkommenslisten für die Wörter *Bundestag*, *Infektionsschutzgesetz*, *Abgeordnete*, *neue* und *ändern*. (*5 Punkte*)

Lösung:

- Bundestag: 1, 2
- Infektionsschutzgesetz: 1, 3
- Abgeordnete: 2
- neue: 1, 2, 3
- ändern: 3

b) Für einen Text T und ein Wort w schreiben wir $w \in T$, falls das Wort w in T vorkommt, anderenfalls $w \notin T$, was gleichbedeutend ist mit $\neg(w \in T)$. Wir können solche Aussagen mit logischen Operatoren verknüpfen. Für Wörter $w = \textit{Bundestag}$ und $v = \textit{Infektionsschutzgesetz}$ sowie Texte $T = T_1$ und $S = T_2$ gilt beispielsweise:

$$(w \in T) \wedge (v \notin S)\,.$$

Hinweis: Haben Sie die Symbole gerade nur überflogen? Lesen Sie den Einführungstext zu dieser Teilaufgabe nochmal und stellen Sie sicher, dass Sie die Notation verstanden haben.

Beantworten Sie die folgenden Fragen (jeweils mit Begründung):

(1) Für welche Texte $T \in \{T_1, T_2, T_3\}$ gilt $\textit{Bundestag} \in T$? (*1 Punkt*)
Lösung:
Für T_1 und T_2, siehe Aufgabenteil a).

(2) Für welche Wörter w der deutschen Sprache gilt $(w \in T_1) \wedge (w \in T_2)$? (*1 Punkt*)
Lösung:
Für *Der*, *Bundestag*, *neue* und *hat*, weil diese Wörter in beiden Texten vorkommen.

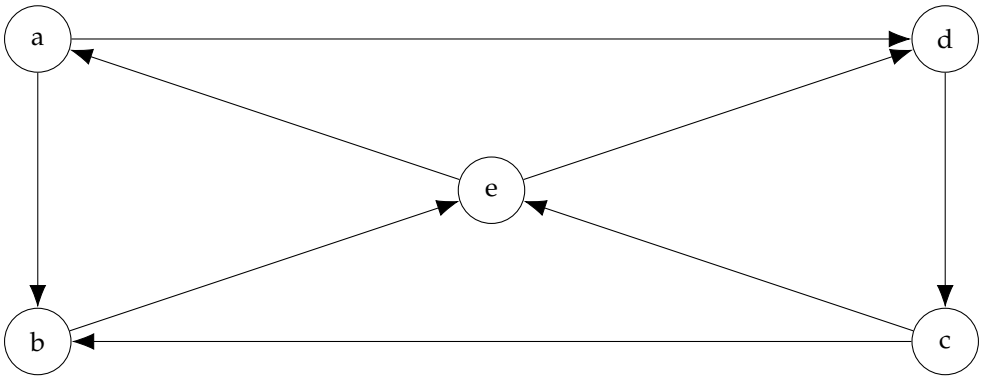
(3) Für welche Texte $T \in \{T_1, T_2, T_3\}$ gilt $(\textit{Abgeordnete} \in T) \vee (\textit{ändern} \in T)$? (*1 Punkt*)
Lösung:
Für T_2 und T_3, da in diesen Texten mindestens eines der genannten Wörter vorkommt, siehe Aufgabenteil a).

c) In der Vorlesung wurde dargestellt, wie Sie durch Mischen von Vorkommenslisten Ergebnisse zu einer Suchanfrage finden können, die zwei Wörter kombiniert (z.B. *Barack* und *Obama*). Stellen Sie dar, wie Sie nach einem ähnlichen Prinzip Ergebnisse zu einer Suchanfrage finden können, die ein Wort w_1 enthält, ein anderes Wort w_2 aber explizit *nicht*. (*2 Punkte*)

Lösung:
Sie wollen die Einträge bestimmen, die in einer sortierten Liste L_1, nicht aber einer sortierten Liste L_2 enthalten sind. Dies erreichen Sie, indem Sie die Einträge e in L_1 durchlaufen und in L_2 immer bis zur kleinsten Zahl z weitergehen, die mindestens so groß ist wie e. Nur wenn für diese Zahl $z > e$ gilt, gehört das Dokument mit der Nummer e zu den Treffern.

Aufgabe 2 Pagerank *(10 Punkte)*

Betrachten Sie den folgenden Graphen. Die Knoten stehen für Webseiten und die Kanten für Verweise zwischen diesen Webseiten. Wir möchten für alle Knoten im Graphen ihre Relevanz bestimmen.



a) Welche Probleme träten bei der Bestimmung der Relevanz nach dem Simulationsprinzip auf, wenn Sie den Pfeil (d,c) umdrehten (mit Begründung)? (*1 Punkt*)

Lösung:
Der Knoten d würde alle Relevanzpunkte „aufsaugen“ (nur eingehende Pfeile) und der Knoten c erhielte 0 Relevanzpunkte (nur ausgehende Pfeile).

b) Stellen Sie das Relevanzgleichungssystem nach dem Prinzip aus der Veranstaltung auf und lösen Sie es. (*6 Punkte*)

Lösung:
Wir beginnen mit der Aufstellung des Relevanzgleichungssystems:

$$
\begin{aligned}
a &= \frac{1}{2}e \\
b &= \frac{1}{2}a + \frac{1}{2}c \\
c &= \frac{1}{1}d = d \\
d &= \frac{1}{2}a + \frac{1}{2}e \\
e &= \frac{1}{1}b + \frac{1}{2}c = b + \frac{1}{2}c \\
1 &= a + b + c + d + e
\end{aligned}
$$

Das ist *überbestimmt*, d.h. wir haben mehr Gleichungen als Variablen. Das liegt daran, dass sich eine der Gleichungen für die Knoten aus den anderen Gleichungen für die Knoten ableiten lässt, und gibt uns etwas Spielraum für „strategisches“ Vorgehen.

Wir können das $\frac{1}{2}e$ in der Gleichung für d durch a ersetzen und haben dann bereits drei Variablen als Ausdrücke von a (nämlich a, c und d):

$$\begin{aligned} a &= \frac{1}{2}e \\ b &= \frac{1}{2}a + \frac{1}{2}c \\ c &= \frac{1}{1}d = d = \frac{3}{2}a \\ d &= \frac{1}{2}a + a = \frac{3}{2}a \\ e &= \frac{1}{1}b + \frac{1}{2}c = b + \frac{1}{2}c \\ 1 &= a + b + c + d + e \end{aligned}$$

Jetzt können wir den Ausdruck für c in die Gleichungen für b und e einsetzen:

$$\begin{aligned} a &= \frac{1}{2}e \\ b &= \frac{1}{2}a + \frac{3}{4}a = \frac{5}{4}a \\ c &= \frac{1}{1}d = d = \frac{3}{2}a \\ d &= \frac{1}{2}a + a = \frac{3}{2}a \\ e &= \frac{1}{1}b + \frac{1}{2}c = b + \frac{3}{4}a \\ 1 &= a + b + c + d + e \end{aligned}$$

Nun setzen wir den Ausdruck für b in die Gleichung für e ein:

$$\begin{aligned} a &= \frac{1}{2}e \\ b &= \frac{1}{2}a + \frac{3}{4}a = \frac{5}{4}a \\ c &= \frac{1}{1}d = d = \frac{3}{2}a \\ d &= \frac{1}{2}a + a = \frac{3}{2}a \\ e &= \frac{1}{1}b + \frac{1}{2}c = \frac{5}{4}a + \frac{3}{4}a = \frac{8}{4}a = 2a \\ 1 &= a + b + c + d + e \end{aligned}$$

Folglich ist

$$1 = a + \frac{5}{4}a + \frac{3}{2}a + \frac{3}{2}a + 2a \Leftrightarrow 1 = \frac{29}{4}a \Leftrightarrow a = \frac{4}{29}\,,$$

und Einsetzen dieses Wertes für a in die Gleichungen für die anderen Knoten liefert $b = \frac{5}{29}$, $c = d = \frac{6}{29}$ sowie $e = \frac{8}{29}$, sodass sich die Werte der einzelnen Knoten – wie erhofft – zu 1 aufaddieren.

c) Sie betreiben neben dem Studium einen Online-Shop für bunte Socken und ärgern sich darüber, dass dieser in den Suchergebnissen einer prominenten Suchmaschine nur weit unten auftaucht. Nehmen Sie an, dass die Suchmaschine das Ranking von Suchergebnissen für bunte Socken allein über die in der Vorlesung eingeführte Variante des Pagerank bestimmt.

(1) Würde es der Sichtbarkeit Ihres Online-Shops helfen, wenn die Universität des Saarlandes auf Ihrer Webseite einen Link zu Ihrem Online-Shop einrichtet (mit Begründung)? (*1 Punkt*)

Lösung:
Ja, denn die Universität des Saarlandes wird auch von anderen Webseiten im WWW verlinkt und es ist davon auszugehen, dass sie positive Relevanz hat, von der sie über den Pagerank einen Teil an Ihren Online-Shop weitergibt.

(2) Ein findiger Kommilitone schlägt vor, Sie könnten in großem Stil Webseiten kaufen und mit diesen Webseiten auf Ihren Online-Shop verlinken. Kann der Vorschlag des Kommilitonen funktionieren (mit Begründung)? (*1 Punkt*)

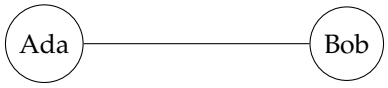
Lösung:
Nein, denn die verlinkenden Webseiten hätten alle eine Relevanz von 0 und würden demnach nicht zur Relevanz Ihres Online-Shops beitragen.

(3) Würde sich Ihre Antwort zur vorigen Teilaufgabe ändern, wenn der Kommilitone vorschlüge, die gekauften Webseiten auch untereinander zu verlinken (mit Begründung)? (*1 Punkt*)

Lösung:
Nein, solange diese Webseiten von keinen anderen als Ihren eigenen Webseiten verlinkt werden und Sie mit Ihrem Web-Shop auf mindestens eine andere Webseite verlinken, denn dann fließt nach wie vor alle Relevanz über Ihren Online-Shop in den Rest des WWW ab.

Aufgabe 3 Soziale Netzwerke *(10 Punkte)*

In der Vorlesung haben wir besprochen, wie man das WWW als Graphen modellieren kann: mit einzelnen Webseiten als Knoten und Links zwischen den Webseiten als Kanten. Ähnlich kann man auch soziale Netzwerke (z.B. LinkedIn) modellieren: Hier sind die Personen die Knoten und eine Kante signalisiert, dass zwei Personen miteinander bekannt sind. Im Unterschied zum WWW haben die Kanten hier keine (öffentlich sichtbare) Richtung: Man sieht nur, dass zwei Personen sich kennen, aber nicht, wer wem eine Kontaktanfrage geschickt hat. Daher zeichnen wir die Kanten ohne Pfeile, etwa so (am Beispiel der Beziehung $\{Ada, Bob\}$):

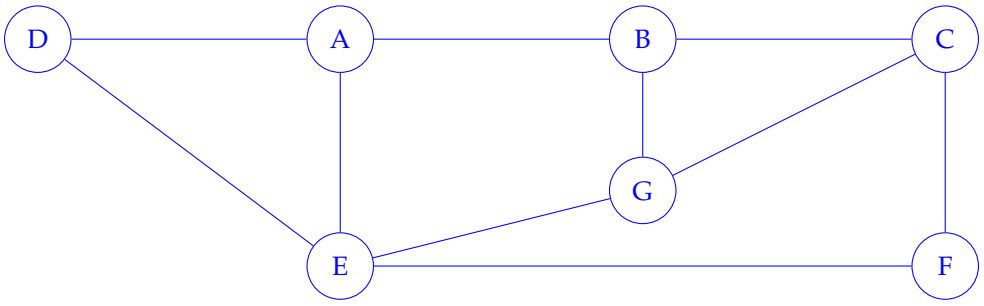


Hinweis: Wenn eine Kante zwischen den Knoten u und v eine Richtung hat, notieren wir sie als (u, v), wenn sie keine Richtung hat, als $\{u, v\}$. Das erste Konstrukt nennt sich Tupel, *hier spielt die Reihenfolge der Elemente eine Rolle, also $(u, v) \neq (v, u)$. Das zweite Konstrukt nennt sich* Menge, *hier spielt die Reihenfolge der Elemente* keine *Rolle, d.h. $\{u, v\} = \{v, u\}$.*

a) Stellen Sie die folgende Liste von Kontakten aus einem sozialen Netzwerk als Graph dar (Sie können die Namen der Personen mit ihren Anfangsbuchstaben abkürzen). (*2 Punkte*)

$$\begin{aligned}\text{Kontakte} = \big[&\{Ana, Bob\}, \{Ana, David\}, \{Ana, Ela\}, \{Bob, Carl\}, \{Bob, Gina\},\\ &\{Carl, Fred\}, \{Carl, Gina\}, \{David, Ela\}, \{Ela, Fred\}, \{Ela, Gina\}\big]\end{aligned}$$

Lösung:



b) Welcher Kontakt ist von *David* am weitesten entfernt (mit Begründung)? (*1 Punkt*)

Lösung:
Carl, weil *Carl* für *David* eine Bekanntschaft dritten Grades ist und es keine Bekanntschaft höheren Grades im Netzwerk gibt.

c) Welchen neuen Kontakt würden Sie *Fred* vorschlagen, wenn Sie ihm jemanden nennen wollen, den er voraussichtlich kennt (mit Begründung)? (*1 Punkt*)

Lösung:
Gina, weil *Gina* die meisten Kontakte mit *Fred* gemeinsam hat.

d) Jede Person im Netzwerk kann die Posts und Re-Posts der anderen Personen sehen, mit denen sie verbunden ist.

(1) *David* hat etwas gepostet und hofft darauf, dass sein Post von allen Personen im Netzwerk gesehen wird. Wie häufig muss der Post mindestens geteilt werden, damit jede Person im Netzwerk die Chance hat, den Post zu sehen – und welche Person(en) müssten in diesem Fall den Post teilen? (*2 Punkte*)

Lösung:
Zweimal: erst von *Ela* und dann von *Gina*, denn jede Person im Netzwerk ist mit mindestens einer Person aus der Menge $\{David, Ela, Gina\}$ verbunden und der Post kann jede Person in der Menge erreichen.

(2) Wie häufig müsste ein Post von *Gina* mindestens geteilt werden, damit jede Person im Netzwerk die Chance hat, den Post zu sehen – und welche Person(en) müssten in diesem Fall den Post teilen? (*1 Punkt*)

Lösung:
Einmal: von *Ela*, denn jede Person im Netzwerk ist mit mindestens einer Person aus der Menge $\{Gina, Ela\}$ verbunden und der Post kann jede Person in der Menge erreichen.

(3) Nun abstrakt formuliert: Wenn in *irgendeinem* sozialen Netzwerk *irgendwer* postet. . .

(i) . . . welche Personen haben eine Chance, den Post zu sehen (d.h. in wessen Newsfeeds taucht der Post auf)? (*1 Punkt*)
Lösung:
Alle Personen, die über einen Pfad (d.h. eine Sequenz von Kanten, in der benachbarte Kanten jeweils einen Endpunkt teilen) vom Ursprung des Posts aus erreichbar sind, der nur Personen enthält, die den Post geteilt haben.

(ii) . . . welche Personen haben *keine* Chance, den Post zu sehen (d.h. in wessen Newsfeeds taucht der Post *nicht* auf)? (*1 Punkt*)
Lösung:
Alle Personen, die *nicht* über einen Pfad vom Ursprung des Posts aus erreichbar sind, der nur Personen enthält, die den Post geteilt haben.

(iii) . . . was muss demnach gegeben sein, damit alle Personen im Netzwerk eine Chance haben, den Post zu sehen? (*1 Punkt*)
Lösung:
Jede Person muss entweder den Post geteilt haben oder mit einer Person bekannt sein, die den Post geteilt hat.

Ich habe für die Videos, die Nachbereitung und das Übungsblatt etwa ☐ Stunden gebraucht.
(Ann-Sophie fertigt aus diesen Zahlen eine Statistik an. Kurt und Corinna sehen nur diese Statistik. Wir möchten wissen, ob der Schwierigkeitsgrad in etwa richtig ist.)

Websuche war spannend ☐ okay ☐ langweilig ☐
schwierig ☐ okay ☐ einfach ☐