Semestralklausur zur Vorlesung

Web Mining

Prof. J. Fürnkranz
Technische Universität Darmstadt — Sommersemester 2008
Termin: 4. 7. 2008

Name: **Vorname:** **Matrikelnummer:**

Fachrichtung:

Punkte: (1) .... (2) .... (3) .... (4) .... (5) .... **Summe:**

Wichtig!

- **Aufgaben:** Diese Klausur enthält auf den folgenden Seiten 5 Aufgaben zu insgesamt 100 Punkten. Jede Aufgabe steht auf einem eigenen Blatt. Kontrollieren Sie *sofort*, ob Sie alle sechs Blätter erhalten haben!
- **Zeiteinteilung:** Die Zeit ist knapp bemessen. Wir empfehlen Ihnen, daß Sie sich zuerst einen kurzen Überblick über die Aufgabenstellungen verschaffen, und dann mit den Aufgaben beginnen, die Ihnen am meisten liegen.
- **Papier:** Verwenden Sie nur Papier, das Sie von uns ausgeteilt bekommen. Sie können Ihre Lösungen beliebig auf die sechs Blätter verteilen, solange klar ersichtlich ist, welche Lösung zu welcher Aufgabe gehört. Sollten sich allerdings mehrere Lösungen zu derselben Aufgabe finden, suchen wir uns eine aus. Insbesondere können Sie auch auf den Rückseiten schreiben!
  Brauchen Sie zusätzlich Papier (auch Schmierpaper), bitte melden.
- **Hilfsmittel:** Als Hilfsmittel darf ein handbeschriebenes DIN-A4-Blatt benutzt werden. Ausländische Studenten dürfen darüber hinaus gedruckte Wörterbücher verwenden. Elektronische Wörterbücher sind nicht erlaubt.
- **Fragen:** Sollten Sie Teile der Aufgabenstellung nicht verstehen, bitte fragen Sie!
- **Abschreiben:** Sollte es sich (wie in den letzten Jahren leider immer wieder) herausstellen, daß Ihre Lösung und die eines Kommilitonen über das zu erwartende Maß hinaus übereinstimmen, werden beide Arbeiten negativ beurteilt (ganz egal wer von wem und egal in welchem Umfang abgeschrieben hat).
- **Ausweis:** Legen Sie Ihren *Studentenausweis* sichtbar auf Ihren Platz.
- **Aufräumen:** Sonst darf außer Schreibgerät, Essbarem, von uns ausgeteiltem Papier und eventuell Wörterbüchern nichts auf Ihrem Platz liegen. Taschen bitte unter den Tisch! Wer bei diesen Temperaturen einen Mantel mithat, wird gebeten, ihn anzubehalten.

Gutes Gelingen!

Aufgabe 1 20 Punkte (4/4/4/4/4)

Sie wollen einen Comparison Shopping Site für Studenten der TU Darmstadt aufbauen, der automatisch von verschiedenen Anbietern im Web Angebote zu Lehrbüchern sammelt, diese vergleicht, und die jeweils günstigen Angebote zur Verfügung stellt. Beantworten Sie die folgenden Fragen und begründen Sie mit mindestens einem Satz.

1–a Jeder Anbieter hat für jedes Buch eine eigene Seite, auf der sich unter anderem der Titel des Lehrbuchs, eine vom Verlag zur Verfügung gestellte kurze Inhaltsangabe, der Name des Autors und der Preis bei diesem Anbieter finden. Sie möchten diese Informationen in einer Datenbank-Relation `angebot(Autor,Titel,Inhalt,Anbieter,Preis)` sammeln. Wie würden Sie vorgehen?

1–b Leider stellen Sie fest, daß dieser Automatisierungsschritt nicht 100% verläßliche Ergebnisse bietet. Sie müssen nun erkennen, welche Seiten verschiedener Anbieter sich auf die gleichen Bücher beziehen, um die Daten angleichen zu können. Wie würden Sie dieses Problem lösen?

1–c Sie wollen das Angebot an Lehrbüchern strukturieren und versuchen, automatisch Gruppen von Fachbüchern zu finden, die ähnliche Themen behandeln. Welche Technik würden Sie hier einsetzen?

1–d Viele Anbieter bieten den Benutzern die Möglichkeit, ihre Meinung zu den Büchern in Form eines kurzen Texts zu äußern. Bei einigen dieser Anbieter hat der Benutzer zusätzlich noch die Möglichkeit, diese Bewertung auf einer Skala von eins bis fünf zusammenzufassen. Wie würden Sie vorgehen, wenn Sie so eine numerische Zusammenfassung der schriftlichen Bewertungen auch für die restlichen Sites nachträglich automatisch ergänzen möchten?

1–e Sie wollen außerdem Ihren Kunden Bücher empfehlen, die sie vielleicht interessieren können. Sie haben aber über die Kunden keinerlei Informationen außer welche Bücher sie bisher bei Ihnen gekauft haben. Wie würden Sie vorgehen?

Aufgabe 2 24 Punkte (8/6/6/4)

Sie betreiben eine Web-Site, die Empfehlungen für bestimmte Produkte abgibt, basierend auf bisherigen Bewertungen dieser Produkte. Im Moment gibt es 5 Benutzer, die 8 Artikel auf einer Skala von 0 bis 3 bewertet haben (0 ist am schlechtesten, 3 ist am besten, bei leeren Einträgen liegt noch keine Bewertung vor).

Benutzer	P1	P2	P3	P4	P5	P6	P7	P8
Adalbert			1		3	1	3	2
Berta		0	2		2	3	3	
Cäsar	0		3				3	2
Dorian	2	2		2	0		**?**	
Emil	3	1		3		1	2	

2–a Schätzen Sie mit Hilfe dem in der Vorlesung besprochenen Verfahren die Bewertung, die Benutzer Dorian für das Produkt P7 abgeben würde.

i. ohne Berücksichtigung der Durchschnittsbewertungen $m(u)$

ii. mit Berücksichtigung der Durchschnittsbewertungen $m(u)$.

Berechnen Sie die Ähnlichkeit zwischen den Benutzern dabei nicht mit den in der Vorlesung angegebenen Korrelations bzw. Kosinus-Maßen, sondern ziehen Sie die durschnittliche Abweichungen aller gemeinsamen Empfehlungen von 3 ab, also (in der Terminologie der Vorlesungsfolien):

$$w(u_1, u_2) = 3 - \frac{1}{|I_{u_1} \cap I_{u_2}|} \sum_{i \in I_{u_1} \cap I_{u_2}} |v(u_1, i) - v(u_2, i)|$$

2–b Nehmen Sie an, daß alle Benutzer Produkte, die sie mit einer Bewertung > 0 bewertet haben, auch gekauft haben (und daß sie keine anderen Produkte gekauft haben). Wenn ein neuer Benutzer nun P6 kauft, welches andere Produkt würden Sie ihm aufgrund von Item-to-Item Korrelationen empfehlen? Stellen Sie die entsprechende Item-to-Item Matrix wie in der Vorlesung auf.

2–c Geben Sie an, welche Zerlegung in zwei Cluster die folgenden Methoden auf der oben angegebenen Matrix jeweils finden sollten (Sie brauchen sich dabei nicht auf einen konkreten Clustering-Algorithmus zu beziehen):

i. User-Clustering

ii. Item-Clustering

iii. Co-Clustering

2–d Geben Sie ein Beispiel für eine gute Assoziations-Regel, die man aus den Kaufdaten aus Aufgabe b) lernen könnte, und begründen Sie, warum Sie die Regel für gut halten.

Aufgabe 3 18 Punkte (6/4/5/3)

Gegeben seien die Web-Dokumente $d_1, d_2, d_3, d_4, d_5, d_6$ und folgende Hyperlinks zwischen diesen Dokumenten:

$d_2 \to d_1$ $d_2 \to d_5$ $d_3 \to d_5$ $d_4 \to d_3$ $d_4 \to d_2$ $d_5 \to d_1$ $d_5 \to d_4$ $d_6 \to d_4$

Sie erhalten als Antwort auf eine Query Q die Dokumente d_1 und d_2 zurück.

3–a Berechnen Sie die Hub und Authority Scores *für die Query Q* anhand des in der Vorlesung vorgestellten HITS-Algorithmus (Führen Sie *maximal* drei Iterationen durch).

3–b Berechnen Sie die erste Iteration des PageRank-Algorithmus auf allen Knoten. Verwenden Sie dabei einen Damping-Faktor von $\frac{2}{3}$ und initialisieren Sie die Werte gleichverteilt.

3–c Um eine endgültige Reihung der Suchresultate vorzunehmen, kombiniert Google den PageRank noch mit Merkmalen, die sich auf den Inhalt der Seiten beziehen.

1. Nennen Sie drei verschiedene Arten von Informationen, die hier berücksichtigt werden.
2. Was würde passieren, wenn man nur den PageRank einer Seite zur Reihung der Suchresultate heranziehen würde?

3–d Nehmen Sie an, daß die Dokumente d_2, d_4, d_6 für die Query Q relevant gewesen wären. Berechnen Sie den F1-Wert für das oben angegebene Query-Resultat.

Aufgabe 4 16 Punkte (2/3/4/3/4)

Textklassifikation

4–a Der Term-Frequency-Vektor eines Dokumentes bezieht sich auf eine bestimmte Termmenge. In die Termmenge sollte man nur Wörter aufnehmen, die vorkommen in

- ☐ der Trainingsmenge
- ☐ der Testmenge
- ☐ der Trainingsmenge und in der Testmenge
- ☐ keine der beiden Mengen

4–b Welche der folgenden Techniken werden eingesetzt, um die Grundmenge der Terme zu verkleinern?

- ☐ Shingling
- ☐ Stemming
- ☐ Sparse Encoding
- ☐ Delta Encoding
- ☐ Case Folding

4–c Wir sagen, ein binärer Klassifierer C ist *bilinear*, wenn es einen Vektor v und eine Zahl b gibt, so dass die Vorhersage $C(x) \in \{-1, +1\}$ sich für jeden Eingabvektor x schreiben lässt als $C(x) = \text{sign}(\langle v, x\rangle + b)$.

Welche der folgenden Lernverfahren erzeugen bilineare Klassifizierer, wenn sie auf ein binäres Klassifikationsproblem angewandt werden?

- ☐ Rocchio
- ☐ Naive-Bayes
- ☐ k-NN
- ☐ lineare SVM

4–d Erklären Sie die Grundidee der TF-IDF-Gewichtung.

4–e Bestimmen Sie, wie viele binäre arithmetische Rechenoperationen ein Naive-Bayes Klassifizierer benötigt, um die Klassenwahrscheinlichkeiten für ein Testdokument zu berechnen, das als TF-Vektor (nicht sparse) gegeben ist, wenn die Termmenge die Größe m hat und K Klassen zu Auswahl stehen.

Aufgabe 5 22 Punkte (5/6/5/6)

Gegeben sei folgende (symmetrische) Matrix von paarweisen Distanzen zwischen Dokumenten. Die letzte Zeile gibt dabei die Klassenzugehörigkeit der Dokumente an.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
d_1	—	27	43	41	25	91	88	77	70
d_2		—	23	68	42	118	115	104	96
d_3			—	81	66	133	129	119	105
d_4				—	31	54	48	39	34
d_5					—	71	70	58	64
d_6						—	12	14	52
d_7							—	13	41
d_8								—	42
d_9									—
Klasse	A	A	A	B	B	C	C	C	C

Hinweis: Sie können sich vorstellen, daß es sich dabei um euklidische Abstände handelt. Ein kleiner Wert bedeutet demnach eine hohe Ähnlichkeit und ein großer Wert steht für unähnliche Dokumente.

5–a Klassifizieren Sie jedes Dokument anhand eines k-NN Klassifizierers, der jeweils auf den restlichen Dokumenten trainiert wurde. Trainieren Sie z.B. um d_5 zu klassifizieren auf $\{d_1, d_2, d_3, d_4, d_6, d_7, d_8, d_9\}$. Geben Sie die Vorhersage für jedes Dokument an und die dafür verantwortlichen Nachbarn. Verwenden Sie $k = 1$.

5–b Erstellen Sie für die Resultate aus a) die Multiclass-Konfusionsmatrix. Berechnen Sie anhand dieser die Gesamt-Accuracy der Vorhersagen und für jede Klasse getrennt jeweils Precision und Recall.

5–c Berechnen Sie die durchschnittlichen Recall- und Precision-Werte der Vorhersagen unter Verwendung von Micro-Averaging und Macro-Averaging.

5–d Beim Micro-Averaging von Mehr-Klassen-Konfusionsmatrizen sind Recall und Precision allgemein gleich. Wie läßt sich das erklären?