Semestralklausur zur Vorlesung

Web Mining

Prof. J. Fürnkranz
Technische Universität Darmstadt — Sommersemester 2004
Termin: 22. 7. 2004

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

Fachrichtung:

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

Aufgabe 1 14 Punkte (3/3/8)

Recall und Precision

1–a Was versteht man unter dem Recall und Precision Trade-Off?

1–b Was ist der Breakeven Point und warum wird er betrachtet?

1–c Sie haben zwei Suchmaschinen A und B. Für eine Query Q wissen Sie, daß 100 Dokumente relevant sind, und daß A einen Recall von 60% und eine Precision von 75% hat, während B eine Precision von 60% und einen Recall von 75% hat.

□ Wie viele Dokumente retourniert jede der beiden Suchmaschinen?

□ Wie viele relevante Dokumente retourniert jede der beiden Suchmaschinen?

□ Was können Sie über die Anzahl der relevanten Dokumente unter den ersten 10 retournierten Dokumenten sagen?

Aufgabe 2 30 Punkte (6x5)

Sie haben die Aufgabe, ein Feedback-System für die Informatik-Lehrveranstaltungen an der TU Darmstadt zu installieren. Das Ziel ist, höhersemestrigen Studenten Vorlesungen zu empfehlen.

2–a Als ersten Schritt müssen Sie alle Home-pages von Lehrveranstaltungen in der Domain informatik.tu-darmstadt.de erkennen. Wie würden Sie hier vorgehen?

2–b In diesen Seiten müssen Sie nun den Titel der Vorlesung identifizieren. Welche Techniken würden Sie nun anwenden?

2–c Danach müssen wollen Sie die Vorlesungsnummer und den Titel aller Informatik-Vorlesungen an der TU Darmstadt in einer Datenbank sammeln. Welchen Wrapper würden Sie einsetzen, wenn Sie für Studenten jedes Semesters je eine Seite gegeben haben, die die jeweiligen Versanstaltungen in einer Tabelle aufbereitet?

2–d Sie wollen nun die im Schritt 2. gewonnene Datenbank `<Titel,URL>` und die im Schritt 3. gewonnene Datenbank `<Lehrveranstaltungsnummer,Titel>` zusammenführen. Welche Probleme können dabei auftreten und wie könnten sie gelöst werden?

2–e Sie erhalten nun vom Prüfungs-Sekretariat eine Datenbank, die angibt, welche Studenten (gegeben durch Matrikelnummern) welche Vorlesungen (gegeben durch Lehrveranstaltungsnummern) besucht haben, und mit welcher Note sie die jeweiligen Vorlesungen abgeschlossen haben. Wie können Sie diese Information nützen um Studenten Empfehlungen für Vorlesungsbesuche abzugeben?

2–f Welche Techniken könnten Sie einsetzen, wenn Sie wissen wollen, welche Vorlesungen sich aus der Sicht der Studenten gut ergänzen?

Begründen Sie Ihre Antworten in allen Punkten!

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

Sie erhalten die Aufgabe, einen Crawler zu bauen.

3–a Nennen Sie drei Probleme, die ein für den praktischen Einsatz tauglicher Crawler lösen muß.

3–b Wie speichern Sie die gefundenen Dokumente ab, sodaß eine effiziente Beantwortung von Boole'schen Queries gewährleistet ist?

3–c Sie sollen die Terme mittels TF-IDF gewichten. Was ist die Grundidee hinter dieser Vorgangsweise?

3–d Bei ersten Tests ihrer Query Engine stellt sich dennoch heraus, daß die Benutzer mit den Resultaten nicht sehr zufrieden sind. Wie können Sie das Ranking mit Hilfe der Benutzer verbessern? Was sind die Nachteile dieser Methode?

Aufgabe 4 22 Punkte (3/6/5/5/3)

Lernverfahren

4–a Erklären Sie die Begriffe supervised, semi-supervised und unsupervised Learning.

4–b Nennen Sie drei in der Vorlesung kennen gelernte Lernverfahren die eine Ähnlichkeitsfunktion verwenden. Beschreiben Sie kurz, in welchem Teil des Algorithmus die Ähnlichkeitsfunktion eingesetzt wird.

4–c Naive Bayes Klassifizierer

□ Was ist am Naive Bayes Klassifizierer "naiv"?

□ Warum schätzt man die Wahrscheinlichkeit des Auftretens eines Wortes üblicherweise nicht mit seiner relativen Häufigkeit in den Trainings-Daten ab?

4–d Was ist Co-Training und warum ist es gerade für Web-Klassifizierungsaufgaben wichtig?

4–e Nennen Sie drei praktische Anwendungen für Clustering-Verfahren.

Aufgabe 5 16 Punkte (4/3/3/6)

Vernetzung

5–a Erklären Sie den PageRank anhand des Random Surfer Modells. Was mißt der PageRank einer Seite?

5–b Eignet sich das Hubs und Authorities-Verfahren für den praktischen Einsatz in Suchmaschinen? Begründung?

5–c Nennen Sie drei Gründe, warum die Klassifikation von Web-Seiten aufgrund des Textes auf Vorgänger-Seiten besser funktionieren kann, als die Klassifikation mit dem Text auf der Seite selbst.

5–d Sie haben 100 Trainingsdokumente, und für jedes dieser Trainingsdokumente kennen Sie 10 Vorgänger-Seiten. Sie möchten gerne Hyperlink Ensembles zur Klassifikation einsetzen.

- □ Wie viele Klassifikatoren müssen Sie dafür trainieren?
- □ An wie vielen Beispielen können Sie diese trainieren?
- □ Wie funktioniert die Klassifikation eines Test-Beispiels, von dem Sie nur 5 Vorgänger-Seiten kennen?