

Information Retrieval - Sommer 2006
Dipl.-Inform. Sascha Kriewel, LF 137
kriewel@is.informatik.uni-duisburg.de

Übungsblatt 2

Aufgabe 4: Mittelwertberechnung für Recall und Precision

In der Vorlesung hast Du gehört, dass es zwei verschiedene Möglichkeiten zur Mittelwertbildung gibt:

Makrobewertung

$$p_M = \frac{1}{N} \sum_{i=1}^N \frac{|REL_i \cap GEF_i|}{|GEF_i|}$$

und Mikrobewertung.

$$p_\mu = \frac{\sum_{i=1}^N |REL_i \cap GEF_i|}{\sum_{i=1}^N |GEF_i|}$$

Zu welchem Zweck ist die Mittelwertbestimmung bei Recall und Precision nötig oder sinnvoll? Warum benötigt man die zwei Ansätze? Stelle die analogen Formeln für den Recall auf.

Betrachte nun die Ergebnisse der folgenden Anfragen. Bestimme zunächst jeweils Precision und Recall für die einzelnen Anfragen. Was bedeuten diese Zahlen auf ein konkretes Ergebnis bezogen?

Anfrage	$ GEF $	$ REL $	$ REL \cap GEF $
q_1	100	25	15
q_2	75	15	12
q_3	120	45	20
q_4	50	35	25

Wie sieht der Mittelwert von Precision und Recall für die beiden Ansätze Makro- und Mikrobewertung aus?

Aufgabe 5: Praktische Suchaufgabe
Abgabe bis 27. April 2006, 12 Uhr:
iruebg-abgaben@is.informatik.uni-duisburg.de

Ende Dezember letzten Jahres hat das Europäische Parlament eine Richtlinie zur Vorratsdatenspeicherung der Kommunikationsdaten von Unionsbürgern mit grosser Mehrheit beschlossen. Diese Richtlinie wurde und wird kontrovers diskutiert.

Um Dich umfassend zu informieren, führe eine Onlinesuche durch, bei der Du zum einen einen Einblick in den Inhalt der Richtlinie gewinnen möchtest, zum anderen an Argumentationen für Pro- und Kontrapositionen interessiert bist (je drei bis vier „gute“ Artikel für Pro und Kontra sollten genügen). Versuche Dokumente aus unterschiedlichen Quellenarten zu finden, um ein möglichst breites Spektrum an Information abzudecken:

- Offizielle Dokumente
- Zeitungsmeldungen, Magazinartikel
- Pressemitteilungen
- Meinungsartikel, Blog-Einträge
- ...

Beschreibe das Vorgehen und die Resultate Deiner Suche. Konntest Du alle Dokumente mit einer Suchmaschine finden, oder hast Du verschiedene Systeme und Quellen benutzt?

Beschränke Dich in Deiner Suche nicht nur auf eine Iteration, sondern versuche, Dich durch Veränderungen der Anfrage einem besonders guten Ergebnis anzunähern. Dazu muß zunächst die Relevanz der gefundenen Dokumente bestimmt werden.

- Durch Analyse der Fehltreffer läßt sich dann die Precision verbessern — warum?
- Durch Analyse der Volltreffer läßt sich der Recall erhöhen — warum?

Gib in der Abgabe Deine Suchschritte mit den benutzten Suchbegriffen an, und beschreibe wie (wenn überhaupt) Du diese aufgrund der zunächst erhaltenen Ergebnisse veränderst hast.

20 Punkte

Aufgabe 6: Methoden zur Recallabschätzung

Bewerte das Resultat Deiner Recherche bezüglich der Abdeckung der relevanten Dokumente. Nimm dazu eine Deiner Anfragen mit ihren Ergebnissen heran. Kannst Du sinnvoll den Recall der Recherche abschätzen? Wenn ja, wie? Wenn nein, warum nicht?

Wie würde die Durchführung der in der Vorlesung vorgestellten Methoden an diesem konkreten Beispiel aussehen? Angenommen, Deine Anfrage lieferte 240 Dokumente. Kommt man um die Relevanzbeurteilung jedes einzelnen Dokumentes herum?