

Aufgabe 31

Wie funktioniert Relevance Feedback? Überlegen Sie sich, in welchen Fällen das Relevance- Feedback-Verfahren nicht sinnvoll ist?

Relevance-Feedback funktioniert typischerweise wie folgt:

1. Der Anwender gibt eine Anfrage in das Information Retrieval-System ein.
2. Das System liefert eine Menge von Dokumente zurück.
3. Der Anwender markiert einige der Dokumente als relevant bzw. nichtrelevant.
4. Das System berechnet eine bessere Repräsentation des Information-Needs auf der Basis der Feedback-Daten. Daraus wird eine neue Anfrage abgeleitet, die relevante Dokumente verstärkt und die nichtrelevanten Dokumenten vermeidet.
5. Das System führt den Suchschritt mit der neuen Anfrage erneut aus und liefert eine neue Menge von Dokumenten zurück, die den Interessen des Nutzers besser entsprechen sollte.
6. Evtl. Rückkehr zu Schritt 3. So wird das Suchergebnis Schritt für Schritt verbessert.

Weitere Informationen finden Sie:

1. http://de.wikipedia.org/wiki/Relevance_Feedback
2. http://greententacle.techfak.uni-bielefeld.de/ir/folien/IR_07_RelevanceFeedback.pdf

Die problematischen Fälle für Relevance-Feedback-Verfahren sind:

1. Rechtschreibfehler. Wenn der Benutzer einen Suchbegriff falsch buchstabiert, wie es in einem Dokument in der Sammlung geschrieben ist, dann ist Relevanz-Feedback unwahrscheinlich, um wirksam zu sein.
2. Cross Language Information Retrieval: Basierend auf Term Distribution stehen die Dokumente in anderen Sprache in dem Vektorraum weiter voneinander als die Dokumente, die in gleichen Sprache sind.
3. Mismatch von Vokabular des Benutzers gegenüber des Dokumentes: Wenn der Benutzer nach „Laptop“ sucht, aber in allen Dokumente „Notebook Computer“ benutzt wird, dann scheitert die Suche

Weitere Informationen finden Sie: Manning Skript Kap. 9 S. 169

Aufgabe 32

Gegeben seien die folgenden Dokumente mit gewichteten Termen:

Doc1: {ROT:0.4, GRÜN:0.6, BLAU:0.3, GELB:0.5, SCHWARZ:0.7}

Doc2: {ORANGE:0.6, LILA:0.8, ROSA:0.7, BEIGE:0.6, BLAU:0.4, GELB:0.2}

Doc3: {BLAU:0.2, PINK:0.6, GRAU:0.3, BEIGE:0.5, SCHWARZ:0.4}

Doc4: {ROT:0.4, LILA:0.2, WEISS:0.7, SCHWARZ:0.4}

Doc5: {GELB:0.4, LILA:0.5, GRAU:0.2, ORANGE:0.7, SCHWARZ:0.6}

Doc6: {BLAU:0.6, GRÜN:0.4}

In dieser Menge werde mit der Anfrage

Gelb:1.0, Lila:0.2

nach dem Vektorraummodell gesucht. Dabei wird als Ähnlichkeitsmaß das Skalarprodukt verwendet.

Relevant markiert seien die Dokumente 2, 3 und 4.

Berechnen sie das Relevance Feedback mit der Methode von Rocchio mit Parametern $\alpha=0.7$, $\beta=0.2$ und $r=0.1$

Folgende Matrix gibt das Gewicht der Terme auf die Dokumente wieder:

	Rot	Grün	Blau	Gelb	Schwarz	Orange	Lila	Rosa	Beige	Pink	Weiss	Grau
Doc1	0.4	0.6	0.3	0.5	0.7	0	0	0	0	0	0	0
Doc2	0	0	0.4	0.2	0	0.6	0.8	0.7	0.6	0	0	0
Doc3	0	0	0.2	0	0.4	0	0	0	0.5	0.6	0	0.3
Doc4	0.4	0	0	0	0.4	0	0.2	0	0	0	0.7	0
Doc5	0	0	0	0.4	0.6	0.7	0.5	0	0	0	0	0.2
Doc6	0	0.4	0.6	0	0	0	0	0	0	0	0	0
Anfrage	0	0	0	1.0	0	0	0.2	0	0	0	0	0

Als Gewichte für das Feedback verwenden $\alpha=0.7$, $\beta=0.2$ und $r=0.1$. Die Dokumente 2,3 und 4 sind relevant. Der Vektor Q_m ergibt sich dann aus:

$$Q_m = 0.7 * \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1.0 \\ 0 \\ 0 \\ 0.2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + (0.2 / 3) * \begin{pmatrix} 0 & 0 & 0.4 \\ 0 & 0 & 0 \\ 0.4 & 0.2 & 0 \\ 0.2 & 0 & 0 \\ 0 & 0.4 & 0.4 \\ 0.6 & 0 & 0 \\ 0.8 & 0 & 0.2 \\ 0.7 & 0 & 0 \\ 0.6 & 0.5 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0 & 0.7 \\ 0 & 0.3 & 0 \end{pmatrix} - (0.1/3) * \begin{pmatrix} 0.4 & 0 & 0 \\ 0.6 & 0 & 0.4 \\ 0.3 & 0 & 0.6 \\ 0.5 & 0.4 & 0 \\ 0.7 & 0.6 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0.2 & 0 \end{pmatrix}$$

$$Q_m = \begin{pmatrix} 0.013 \\ -0.033 \\ 0.01 \\ 0.683 \\ 0.01 \\ 0.017 \\ 0.19 \\ 0.047 \\ 0.073 \\ 0.04 \\ 0.047 \\ 0.013 \end{pmatrix}$$

Setzt man die negativen Einträge auf 0, so erhält man folgenden neuen Anfragevektor:

$$Q_{neu} = \begin{pmatrix} 0.013 \\ 0 \\ 0.01 \\ 0.683 \\ 0.01 \\ 0.017 \\ 0.19 \\ 0.047 \\ 0.073 \\ 0.04 \\ 0.047 \\ 0.013 \end{pmatrix}$$

Dieser modifizierte Anfragevektor führt zu einem deutlich anderen Ergebnis. Die Werte der Ähnlichkeitsmaß (das Skalarprodukt) und auch die Reihenfolge sehen nun so aus:

- 1 → Doc5: 0.3887
- 2 → Doc2: 0.3795
- 3 → Doc1: 0.3567
- 4 → Doc4: 0.0801
- 5 → Doc3: 0.0704
- 6 → Doc6: 0.006

Aufgabe 33

- 1) Beschreiben Sie das Prinzip der automatischen Anfrageerweiterung (Pseudo Relevance Feedback), wie sie von vielen automatischen Systemen (unter anderen SMART) in TREC benutzt wurde.
- 2) Welche Annahmen liegen diesem Anfrageerweiterungsverfahren zugrunde?
- 3) Wie unterscheiden sich diese Pseudo Relevance Feedback Verfahren vom Rocchio Verfahren?

- 1) Das Pseudo Relevance Verfahren ist eine abgeänderte Form des Relevance Feedback, um dem Benutzer die Bewertungsentscheidung abzunehmen. Das Suchsystem assoziiert automatisch eine Relevanz für die jeweiligen Ergebnisse Dokumente, woraufhin dann die Suchanfrage per Query Expansion automatisch erweitert wird, und eine neue Ergebnisliste mit der erweiterten Suchanfrage generiert wird. Es funktioniert wie folgt:
 - Die Top-k (mit k zwischen 10 bis 50 in den meisten Experimenten) Ergebnisse von erster Anfrage werden als relevante Dokumente zurückgeliefert.
 - Top 20-30 (indikative Anzahl) Terme aus diesen Dokumenten werden dann gewählt beispielsweise unter Verwendung von TF-IDF Gewichte. Solche Terme werden zur Anfrage hinzugefügt.
 - Das System führt den Suchschritt mit der neuen Anfrage erneut aus. Schließlich werden die wichtigsten Dokumente zurückgeliefert.
- 2) Pseudo Relevance Feedback kann erfolgreich eingesetzt werden nur wenn die erste Anfrage bereits gute Ergebnisse liefert, sonst schlechte Retrievalqualität.
- 3) Nach dem Rocchio Verfahren sind die Menge der relevanten und nichtrelevanten Dokumente bekannt. Zum Vergleich werden nur die relevanten Dokumente im Pseudo Relevance Feedback Verfahren berücksichtigt. Weitere Informationen finden Sie: http://en.wikipedia.org/wiki/Relevance_feedback

Aufgabe 34

In der Aufgabe 32 war Relevance Feedback nach dem Rocchio Verfahren berechnet worden. SMART verwendet in TREC modifizierte Relevance Feedback Verfahren zur Query Expansion.

- 1) Nutzen Sie die drei bestplatzierten Dokumente zur Expansion und expandieren Sie die Query um 3 Terme, wobei die Terme nach der Summe ihrer Gewichte in den bestplatzierten Dokumenten ausgewählt werden. Verwenden Sie ansonsten die Ähnlichkeitsmaße und Gewichte aus Aufgabe 32
 - 2) Wie können dabei Gewichte für die neuen Queryterme bestimmt werden?
- 1) Nach der Berechnung in der Aufgabe 32 sind 1, 2 und 5 Dokumente die drei besten Treffer als positives Feedback mit Gewichte: Doc5: 0.3887, Doc2: 0.3795, Doc1: 0.3567. Nach der Summe der Termgewichte werden die Terme „Gelb, Lila, Schwarz, Orange, Blau“ als neue Anfrage erweitert.

	Rot	Grün	Blau	Gelb	Schwarz	Orange	Lila	Rosa	Beige	Pink	Weiss	Grau
Doc1	0.4	0.6	0.3	0.5	0.7	0	0	0	0	0	0	0
Doc2	0	0	0.4	0.2	0	0.6	0.8	0.7	0.6	0	0	0
Doc5	0	0	0	0.4	0.6	0.7	0.5	0	0	0	0	0.2
Summe	0.4	0.6	0.7	1.1	1.3	1.3	1.3	0.7	0.6	0	0	0.2

- 2)
Die Termgewichte für die neue Queryterme (TF-IDF Werte, Log basiert auf 2):
 $W(\text{schwarz}) = 1 * \log 6/4 = 0.585$
 $W(\text{orange}) = 1 * \log 6/2 = 1.585$
 $W(\text{blau}) = 1 * \log 6/4 = 0.585$