Errata in R. Ferber: Information Retrieval

Stand: 1. Oktober 2003

Durch einen vom Autor zu verantwortenden Fehler im Konvertierungsprogramm, mit dem die SGML-Version nach LaTeX konvertiert wurde, fehlen an folgenden fünf Stellen Wurzelzeichen:

Seite 71 unten (Abschnitt 3.6.5 Lokale Gewichtungseinflüsse)

Eine komplexere Formel, die für das experimentelle System SMART (Salton und McGill, 1983 [103]) entwickelt wurde, lautet z. B.:

$$\tilde{w_{i,j}} = \frac{1}{2}\left(1 + \frac{h\left(i,j\right)}{max_{k\in\{1,\ldots,n\}}\left\{h\left(i,k\right)\right\}}\right) ln\left(\frac{m}{d\left(j\right)}\right)$$

bzw. als normierte Version:

$$w_{i,j} = \frac{\tilde{w_{i,j}}}{\sqrt{\sum_{k=1}^{n} \tilde{w_{i,k}}^2}}$$

Seite 74 unten und 75 mitte (Abschnitt: Das Cosinus-Maß)

Ein Ähnlichkeitsmaß, bei dem die Länge der Vektoren keinen direkten Einfluss auf die Ähnlichkeit hat, ist das Cosinus-Maß (der Cosinus im $\mathcal{R}^n$):

$$cos\left(w_i, q\right) = \frac{\sum_{k=1}^{n} w_{i,k} q_k}{\sqrt{\sum_{k=1}^{n} w_{i,k}^2} \cdot \sqrt{\sum_{k=1}^{n} q_k^2}}$$

Wie der Cosinus in der Ebene liegen die Ähnlichkeitswerte zwischen zwei Vektoren bei diesem Maß immer im Intervall $[-1, 1]$. Die beiden Wurzeln im Nenner sind gerade die euklidischen Längen der Vektoren w_i und q. Man kann sich die Formel also auch als einfaches Skalarprodukt der normierten Vektoren

$$\frac{w_i}{\sqrt{\sum_{k=1}^{n} w_{i,k}^2}} \text{ und } \frac{q}{\sqrt{\sum_{k=1}^{n} q_k^2}}$$

vorstellen.

Seite 184 (Abschnitt 9.1.3: Das P-Norm-Modell)

Schließlich beschreiben Fox, Betrabet, Koushik und Lee (1992) [37] noch das P-Norm-Modell. Hier sind auch im Anfragevektor Einträge zwischen 0 und 1 zugelassen. Als Ähnlichkeitsmaß für eine OR-Anfrage wird die Formel

$$s\left(w_i, q\right) = \frac{\sqrt[p]{\sum_{j=1}^{n} \left(w_{i,j} q_j\right)^p}}{\sqrt[p]{\sum_{j=1}^{n} q_j^p}}$$

mit $p \in \{1, \ldots, \infty\}$ verwendet. Für $p = 1$ und Anfragevektoren, die nur aus den Werten 0 und 1 bestehen, handelt es sich also um die gleiche Formel (9.2) wie bei der Paice-Ähnlichkeit mit $c = 1$. Für größere Werte von p beschreibt der Zähler die P-Norm-Länge der Projektion des Dokumentvektors auf den Unterraum, der von den Anfragetermen aufgespannt wird, für $p = 2$ also die euklidische Länge dieses Vektors. Dieser Wert wird mit der p-Norm-Länge des Anfragevektors normiert. Falls im Anfragevektor auch Werte zwischen 0 und 1 zugelassen sind, wird der Unterraum entsprechend afin abgebildet, d. h., die Skalen seiner Achsen werden verändert.

Fox, Betrabet, Koushik und Lee (1992) [37] geben als Heuristik für die Wahl des Maßes den Abstand zum Ursprung des Unterraums an. Als Ähnlichkeitsmaß für eine AND-Anfrage wählen sie die Formel

$$s\left(w_i, q\right) = 1 - \frac{\sqrt[p]{\sum_{j=1}^{n}\left(\left(1 - w_{i,j}\right) q_j\right)^p}}{\sqrt[p]{\sum_{j=1}^{n} q_j^p}}$$

mit $p \in \{1, \ldots, \infty\}$ und interpretieren diesen Wert als Abstand der Projektion des Dokumentvektors vom Punkt $(1, \ldots, 1)$ im Unterraum, der von den Termen der Anfrage aufgespannt wird.

Seite 208 (Abschnitt 12.1 Die TREC-3-Ergebnisse von Smart)

Die Gewichtung der Terme in einer Anfrage wurde mit der Formel

$$w_{i,k} = \frac{\left(ln\left(h\left(i,k\right)\right) + 1,0\right) \cdot ln\left(m/d\left(k\right)\right)}{\sqrt{\sum_{j=1}^{n}\left[\left(ln\left(h\left(i,j\right)\right) + 1,0\right) \cdot ln\left(m/d\left(j\right)\right)\right]^2}}$$

berechnet, wobei $h(i,k)$ die Häufigkeit des Terms t_k im Dokument bzw. der Anfrage i bezeichnet, m die Anzahl der Dokumente, n die Anzahl der Terme und $d(k)$ die Anzahl der Dokumente, die den Term t_k enthalten. Für die Termpaare wurde die gleiche Formel verwendet, wobei im Nenner aber wieder nur über die Terme summiert wurde.

Seite 229 (Abschnitt 13.2.4: Häufigkeit der Terme)

$$cos\left(i,j\right) = \frac{h\left(i,j\right)}{\sqrt{h\left(i\right)h\left(j\right)}}$$

(...) Vergleicht man diesen Quotienten mit den Ähnlichkeitsmaßen, so zeigt sich, dass alle drei Maße häufige Terme stärker begünstigen als dieser. Der $cos(i,j)$ z. B. unterscheidet sich von $U(i,j)$ durch den Faktor $F_c = \sqrt{h(i)h(j)}$:

$$cos\left(i,j\right) = \frac{h\left(i,j\right)}{\sqrt{h\left(i\right)h\left(j\right)}} = \frac{h\left(i,j\right)}{h\left(i\right)h\left(j\right)} \cdot \sqrt{h\left(i\right)h\left(j\right)} = U\left(i,j\right) \cdot \sqrt{h\left(i\right)h\left(j\right)}$$

Je häufiger ein Term vorkommt, desto größer ist der Faktor F_c.