Binary Independence Retrieval (BIR) und Robertson-Sparck Jones Gewichtung (RSJ)

(13.4.2007, Martin Braschler)

(Die vorliegende Darstellung ist in der Notation stark vereinfacht, sollte aber genügen, die Einschränkungen/Annahmen, welche dem BIR, resp. der RSJ-Gewichtung unterliegen, darzustellen. Für genauere Herleitungen siehe N. Fuhr, P. Schäuble, R. Ferber und andere)

Das Probability Ranking Principle fordert eine Rangierung gemäss der Wahrscheinlichkeit, dass ein Dokument zu einer Anfrage relevant ist.

D.h., wir suchen eine Schätzung für

$$P(R|q,d)=P(R|q\cap d)$$

(die beiden Schreibweisen sind im folgenden äquivalent)

Es kommt jede Funktion in Frage, welche den Dokumenten Werte zuweist der Form (sogenannte RSV-Funktion):

$$RSV(q,d)=f(P(R|q,d))+g(q)$$

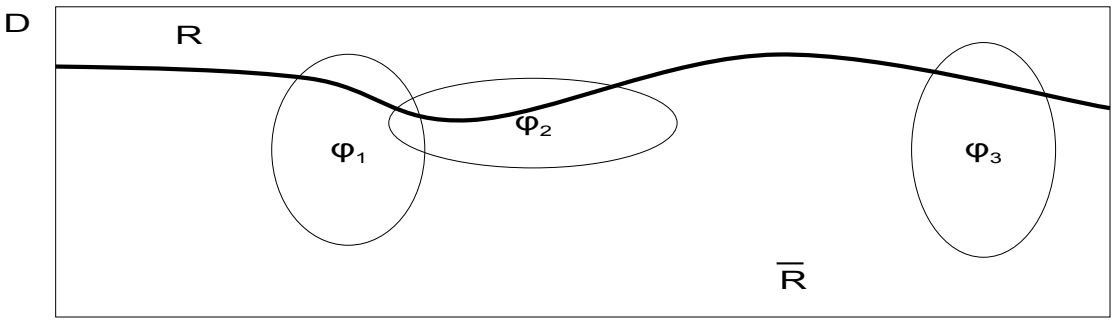
Dabei muss f eine ordnungserhaltende Funktion sein, d.h., die Anwendung von f auf die Wahrscheinlichkeiten führt zu keiner Umsortierung der Rangliste. Sonstige Summanden beeinflussen die Reihenfolge in der Rangliste ebenfalls nicht, solange sie rein von der Anfrage abhängig sind, nicht aber vom Dokument (wie oben angedeutet).

Als Rüstzeug für die folgende Betrachtung brauchen wir ein paar Äquivalenzen aus der Wahrscheinlichkeitsrechnung:

$P(A|B)=\frac{P(A\cap B)}{P(B)}$, $P(A\cap B)=P(A|B)\cdot P(B)=P(B|A)\cdot P(A)$ (Definition der bedingten Wahrscheinlichkeit)

$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)}$ (Bayes'sche Regel)

Wir können ja mit Hilfe der grafischen Betrachtung erahnen, wie wir Wahrscheinlichkeiten, dass das gewisse Terme in relevanten oder irrelevanten Dokumenten auftreten, berechnen können. D.h., wir müssen die gesuchte Schätzung der W'keit, dass ein Dokument zu einer Anfrage relevant ist, auf eine Frage der Wahrscheinlichkeiten, dass die Terme des Dokuments in relevanten bzw. irrelevanten Dokumenten auftreten würden, reduzieren.



Es gilt nun mit Hilfe unseres Rüstzeugs von oben:

$P(R|q,d)=P(R|q\cap d)=\frac{P(R\cap q\cap d)}{P(q\cap d)}$ und wenn wir d nun sozusagen „ausklammern“ erhalten wir $\frac{P(d|R\cap q)\cdot P(R\cap q)}{P(d|q)\cdot P(q)}=\frac{P(d|R\cap q)\cdot P(R|q)}{P(d|q)}=\frac{P(d|R,q)\cdot P(R|q)}{P(d|q)}$

Es bleiben nun also drei Wahrscheinlichkeiten zum Abschätzen. Der Faktor $P(d|R,q)$ ist nun eben die Wahrscheinlichkeit, dass ein beliebiges relevantes Dokument die Gestalt von d hat, und damit das Ziel der „Begierde“.

Wir können uns einen der Faktoren schenken, wenn wir statt $P(R|q,dj)$ die „Odds“, d.h., das Verhältnis zwischen Auftreten und Nicht-Auftreten eines Ereignisses berechnen. Die Odds sind ordnungserhaltend, also für eine RSV-Funktion geeignet. Die Definition ist:

$O(x)=\frac{P(x)}{P(\overline{x})}=\frac{P(x)}{1-P(x)}$

Wir suchen also im Folgenden nun:

$$O(R|q,d)=\frac{P(R|q,d)}{P(\overline{R}|q,d)}=\frac{\frac{P(d|R,q)\cdot P(R|q)}{P(d|q)}}{\frac{P(d|\overline{R},q)\cdot P(\overline{R}|q)}{P(d|q)}}=\frac{P(d|R,q)}{P(d|\overline{R},q)}\cdot O(R|q)$$

Wie oben erwähnt, ist $P(d|R,q)$ die W'keit, dass ein beliebiges relevantes Dokument die Gestalt von d hat, und analog ist $P(d|\overline{R},q)$ die W'keit, dass ein beliebiges irrelevantes Dokument die Gestalt von d hat.

Wir wollen nun den Schritt von den Dokumenten zu den einzelnen Termen machen. Dazu benötigen wir ein paar Annahmen:

1. Dokumente, die aus dem gleichen „Bag of Words“, d.h. der gleichen ungeordneten Menge von Termen bestehen, sind identisch (d.h., Dokumente werden nur durch ihre Terme repräsentiert)
2. Einzelne Terme sind voneinander unabhängig (!)

Das Dokument d kann also als Menge seiner Terme dargestellt werden.

$d=\Phi(d)=\{\phi_1,\phi_2,..,\phi_n\}$

Man definiert dann basierend auf diesen Mengen Vektoren der Länge n=Anzahl Terme in der Kollektion, welche die Dokumente und Anfragen repräsentieren (daher der Name „Binary“, aus Annahme 2 von oben folgt das „Independence“).

$\vec{x}=(x_1,..,x_n)$, wobei $x_i=1\ falls\ \phi_i\in\Phi(d)$, $x_i=0\ sonst$

Die Unabhängigkeitsannahme führt dazu, dass wir gleichsetzen können:

$$O(R|q,\vec{x}) = \frac{P(\vec{x}|R,q)}{P(\vec{x}|\overline{R},q)} \cdot O(R|q) = \frac{\prod_{i=1}^{n} P(x_i|R,q)}{\prod_{i=1}^{n} P(x_i|\overline{R},q)} \cdot O(R|q)$$

Man kann das entstandene Produkt dann aufteilen, je nach Auftreten der Terme im betrachteten Dokument:

$$O(R|q,\vec{x}) = O(R|q) \cdot \prod_{x_i=1} \frac{P(x_i=1|R,q)}{P(x_i=1|\overline{R},q)} \cdot \prod_{x_i=0} \frac{P(x_i=0|R,q)}{P(x_i=0|\overline{R},q)}$$

Wir wollen nun die Schreibweise vereinfachen:

$p_i = P(x_i=1|R,q)$ (W'keit, dass ϕ_i in bel. relevantem Dokument auftritt).

$q_i = P(x_i=1|\overline{R},q)$ (W'keit, dass ϕ_i in bel. irrelevantem Dokument auftritt).

Sowie $1-p_i = P(x_i=0|R,q)$ und $1-q_i = P(x_i=0|\overline{R},q)$

Wir betrachten nun die Terme im Wechselspiel zwischen Dokument und Anfrage. Ein Term ist entweder:

1. Nur im Dokument vorhanden, aber nicht in der Anfrage, d.h. $\phi_i \in \Phi(d) \setminus \Phi(q)$
2. Nur in der Anfrage vorhanden, nicht aber im Dokument, d.h. $\phi_i \in \Phi(q) \setminus \Phi(d)$
3. In Anfrage und Dokument vorhanden, d.h. $\phi_i \in \Phi(q) \cap \Phi(d)$
4. Weder in Anfrage noch Dokument vorhanden

Wir nehmen im folgenden an, dass Terme, die nicht der Anfrage vorkommen, in relevanten und irrelevanten Dokumenten gleich häufig vorkommen (d.h., diese Terme sind nicht geeignet, um bezüglich der Anfrage Relevanz und Irrelevanz zu entscheiden). D.h., für $\phi_i \notin \Phi(q)$ gilt $p_i = q_i$. Dies führt zu einer erheblichen Vereinfachung: die entsprechenden Faktoren fallen weg (da je = 1).

Es bleibt übrig:

$$O(R|q,\Phi(d)) = O(R|q,\vec{x}) = O(R|q) \cdot \prod_{\phi_i \in \Phi(q) \cap \Phi(d)} \frac{p_i}{q_i} \cdot \prod_{\phi_i \in \Phi(q) \setminus \Phi(d)} \frac{1-p_i}{1-q_i}$$

Dieses Produkt lässt sich nach einem kleinen Trick (Multiplikation der Faktoren für $\phi_i \in \Phi(q) \cap \Phi(d)$ mit dem Faktor $\frac{(1-p_i)\cdot(1-q_i)}{(1-q_i)\cdot(1-p_i)}$; dieser kürzt sich problemlos auf 1) umgruppieren:

$$O(R|q,\Phi(d)) = O(R|q) \cdot \prod_{\phi_i \in \Phi(q) \cap \Phi(d)} \frac{p_i \cdot (1-q_i)}{q_i \cdot (1-p_i)} \cdot \prod_{\phi_i \in \Phi(q)} \frac{1-p_i}{1-q_i}$$

Wir sind am Ziel. Von den verbleibenden drei Faktoren ist nämlich nur noch der Mittlere vom Dokument abhängig. Die anderen beiden Faktoren sind nur von der Anfrage abhängig, und daher über die Dokumente konstant, d.h., für die Rangierung unnötig.

Der Logarithmus ist ordnungserhaltend, und erlaubt es, Produkte in Summen zu „verwandeln“. Wir können also eine RSV-Funktion wie folgt definieren:

$$RSV(q,d)=\sum_{\phi_i\in\Phi(q)\cap\Phi(d)}\log\left(\frac{p_i\cdot(1-q_i)}{q_i\cdot(1-p_i)}\right)$$

Man nennt dies die Robertson-Sparck Jones-Gewichtung.

Aufmerksame Leser werden einwenden, dass die Werte p_i (die W'keit, dass ϕ_i in einem bel. relevanten Dokument auftritt) und q_i (die W'keit, dass ϕ_i in einem bel. irrelevanten Dokument auftritt) ja noch nicht bestimmt sind.

Hätten wir Relevanzbeurteilungen für alle Dokumente in Hinsicht auf die Anfrage, so könnten wir folgendermassen vorgehen:

$$p_i=\frac{Anzahl\ rel.\ Dokumente\ mit\ \phi_i}{Anzahl\ rel.\ Dokumente}\quad,\quad 1-p_i=\frac{Anzahl\ rel.\ Dokumente\ ohne\ \phi_i}{Anzahl\ rel.\ Dokumente}$$

$$q_i=\frac{Anzahl\ irrel.\ Dokumente\ mit\ \phi_i}{Anzahl\ irrel.\ Dokumente}\quad,\quad 1-q_i=\frac{Anzahl\ irrel.\ Dokumente\ ohne\ \phi_i}{Anzahl\ irrel.\ Dokumente}$$

Es ist dann:

$$\log\left(\frac{p_i\cdot(1-q_i)}{q_i\cdot(1-p_i)}\right)=\log\left(\frac{p_i}{1-p_i}\right)-\log\left(\frac{q_i}{1-q_i}\right)\approx$$
$$\log\left(\frac{1/2+Anzahl\ rel.\ Dokumente\ mit\ \phi_i}{1/2+Anzahl\ rel.\ Dokumente\ ohne\ \phi_i}\right)-\log\left(\frac{1/2+Anzahl\ irrel.\ Dokumente\ mit\ \phi_i}{1/2+Anzahl\ irrel.\ Dokumente\ ohne\ \phi_i}\right)$$

Wobei der Faktor 1/2 unangenehme Divisionen verhindert. Dies können wir durchaus so handhaben im Falle, dass wir während eines Relevance Feedback-Vorganges Relevanzinformation haben.

Wie schätzt man nun aber in anderen Fällen die unbekannten Dokumentenanzahlen? Es drängt sich auf anzunehmen, dass deutlich mehr Dokumente zu einer Anfrage irrelevant als relevant sind. Eine Näherung ist es anzunehmen, dass gar kein Dokument relevant ist, und dass alle Dokumente irrelevant sind.

Es sind dann also $Anzahl\ rel.\ Dokumente\ mit\ \phi_i=0$ für alle ϕ_i und
$Anzahl\ irrel.\ Dokumente\ mit\ \phi_i=df(\phi_i)$ für alle ϕ_i

Mit n=Anzahl Dokumente=Anzahl irrel. Dokumente bekommen wir dann:

$$\log\left(\frac{p_i\cdot(1-q_i)}{q_i\cdot(1-p_i)}\right)\approx\log\left(\frac{1/2+0}{1/2+0}\right)-\log\left(\frac{1/2+df(\phi_i)}{1/2+n-df(\phi_i)}\right)=\log\left(\frac{1/2+n-df(\phi_i)}{1/2+df(\phi_i)}\right)$$

Was nichts anderes ist als Gewichte, welche der inversen Dokumentenhäufigkeit
$idf(\phi_i)=\log\left(\frac{1+n}{1+df(\phi_i)}\right)$ sehr ähnlich sind! Damit wäre also der Wert von idf-Gewichtungen auch theoretisch untermauert.