Das Vektorraummodell

im Information Retrieval

Die zunehmende Digitalisierung von Daten, wie Texte oder Bilder, schafft viele Vorteile zu deren Speicherung und Verbreitung. Nicht zuletzt ist die massive Anzahl an Daten, die im Internet zur Verfügung stehen, darauf zurückzuführen. Für den Verbraucher stellt die so entstandene Fülle eine zuverlässige Quelle für die eigene Recherche dar, deren Nutzung vergleichsweise wenig Aufwand erfordert. Dennoch benötigt eine solche Datenmenge auch eine gewisse Struktur und Ordnung, insbesondere wenn es darum geht, Daten zielgenau abzufragen. Dazu ist eine geeignete Charakterisierung notwendig, die jeder Datei ein einzigartiges Merkmal zuweist. Im Folgenden möchte ich anhand von Texten bzw. Dokumenten etwas näher auf die Frage der Struktur eingehen und mithilfe einer Begriffsabfrage erläutern, wie die dazu entsprechenden Dokumente gefunden werden können.
Zunächst einmal stellt sich die Frage nach einer geeigneten Charakterisierung und Indexierung der Daten, um sie wie erwähnt in eine strukturierte Umgebung einbetten zu können. Da der Inhalt der Dokumente dafür ausschlaggebend ist, legt das den Schluss nahe, übergreifende Schlagworte zu definieren und diese dem Dokument zuzuordnen. Auf diese Weise erhält jedes Dokument seine eigene einzigartige Sammlung an Schlagworten. Von zentraler Bedeutung ist nun die Frage nach einem Modell, das es ermöglicht, die so modifizierten Dokumente zu kategorisieren und diese mit anderen (bzw. im Falle einer Abfrage mit Suchworten) vergleichbar zu machen. Eine geeignete Möglichkeit, dies zu realisieren, bietet das Konzept des Vektorraums. Jede Koordinatenachse steht dabei für ein Schlagwort, das Dokument wird als Vektor dargestellt. Um Verwechslungen zweier inhaltlich ähnlicher Dokumente zu vermeiden und eine noch stärkere Spezifizierung zu ermöglichen, werden die Schlagworte in der Regel noch zusätzlich lokal gewichtet (Anteil an der Gesamtanzahl der Schlagworte eines Dokuments). Dabei bleibt die Summe der Quadrate der Einträge gleich, es wird jedoch innerhalb eines Dokuments (Vektors) klarer, wie viel Bedeutung den einzelnen Begriffen zukommt. Eine Suchanfrage (mit mehreren Begriffen) wird mithilfe dieses Modells folglich als Vektor abstrahiert, dessen Koordinatenwerte die Gewichtung der einzelnen Schlagworte und damit die eingegebenen Begriffe widerspiegeln. Bei dieser Form der Entstehung eines Vektors dienen die Begriffe somit direkt als Schlagwort, was angesichts Mehrdeutigkeit (Polysemie) und Synonymie von Wörtern zu Problemen führen kann. Um schließlich grundlegende Berechnungen für die Suche nach Dokumenten durchführen zu können, werden alle Dokumente, die anfangs die Datenbank darstellen sollen, als Spaltenvektoren zu einer Matrix (Dokumentenmatrix) zusammengefasst. Damit lassen sich nun elementare Vektoroperationen verwirklichen, wie die QR-Zerlegung oder die Singulärwertzerlegung, die vor allem für die Niedrig-Rang-Approximation der Dokumentenmatrix von Bedeutung sind. Diese Niedrig-Rang-Approximation verringert den Rang der Dokumentenmatrix und schließt dadurch Bereiche der Matrix aus, die auf Kosten eines gewissen Genauigkeitsverlusts als irrelevant betrachtet werden können. Der Vorteil liegt hierbei in der schnelleren Berechnung der gewünschten Dokumente. Geometrisch betrachtet handelt es sich dabei um die Vektoren, die am nächsten zu dem Vektor liegen, dessen Koordinatenwerte durch die Suchbegriffe gegeben sind. Da der Abstand von Vektoren verglichen wird, spielen Winkelberechnungen beim IR eine wichtige Rolle.

Erste Versuche, Dokumente nach Schlagworten zu ordnen, gehen bis in die 50iger Jahre des 20. Jahrhunderts zurück. Die ersten standardisierten Berechnungen zur Einordnung von Dokumenten wurden jedoch von einer 1992 gegründeten Konferenz, der Text Retrieval Conference (TREC), durchgeführt. Sie tritt jährlich zusammen und wird dank finanzieller Unterstützung der Defense Advanced Research Projects Agency (DARPA), einer Behörde des Verteidigungsministeriums der Vereinigten Staaten von Amerika, und dem National Institute of Standards and Technology (NIST) aufrecht erhalten.

Das Vektorraum-Modell

Die Dokumente stellen also, um Berechnungen anstellen zu können, die Spalten einer Matrix dar. Damit ergibt sich, dass die Schlagworte als Zeilenindizes verwendet werden. Es entsteht somit eine *sxd* - Matrix mit d Spalten (Anzahl der Dokumente) und s Zeilen (Anzahl der Schlagworte). Der semantische Inhalt der Datenbank befindet sich folglich im Spaltenraum der Matrix. Es lassen sich jedoch keine Linearkombinationen der Spalten bilden, sodass man nach beliebiger Kombination auf einen aussagekräftigen Vektor (Dokument) schließen könnte. Neben der lokalen Gewichtung der einzelnen Begriffe kann auch noch eine globale Komponente hinzugefügt werden, die das Verhältnis der Einträge zu allen Einträgen der Dokumente (Datenbank) berücksichtigt. Generell kann die Gewichtung von einfachen Binärwerten (0 und 1), über Normalisierung (des Vektors) bis zu komplexen Algorithmen reichen. Ein gewöhnliches Lexikon weist in der Regel deutlich mehr Schlagworte als Dokumente auf. Das Internet dagegen bietet wesentlich mehr Dokumente in Form von Webseiten. Eine *sxd*-Matrix, die den Inhalt des Internets darstellen soll, käme so auf ungefähr 300.000 x 6.000.000.000 Einträge, wobei die Anzahl der Zeilen den Schlagworten eines Wörterbuchs entspricht. Dabei ist zu beachten, dass bei Matrizen mit solch großem Umfang die meisten Einträge 0 sind, da nur ein Bruchteil der Begriffe eines Wörterbuchs wirklich verwendet wird. Folglich besteht eine Suchanfrage, die als Vektor dargestellt ist, auch weitestgehend aus Nullen. Um nun Begriffe aus einer Eingabe mit einer Datenbank abzugleichen, werden die Vektoren miteinander verglichen. Dabei wird jeweils der Kosinus der Winkel zwischen dem Eingabevektor und den Vektoren der Dokumente berechnet. Mit den Spalten a_j einer Matrix A und dem Eingabevektor q ergeben sich die d Winkel:

$$\cos \alpha_j = \frac{a_j^t q}{||a_j||_2 ||q||_2} = \frac{\sum_{i=1}^{s} a_{ij} q_i}{\sqrt{\sum_{i=1}^{s} a_{ij}^2} \sqrt{\sum_{i=1}^{s} q_i^2}}$$

für $j = 1, \ldots, d$ und $dim(a_j) = s$.
Dabei müssen die Vektornormen $||a_j||_2$ nur einmal pro *sxd*-Matrix berechnet werden.

Anhand eines Beispiels soll nun verdeutlicht werden, wie grundlegende Operationen an einer solchen *sxd*-Matrix A durchgeführt werden. Zur Veranschaulichung werden folgende 6 Schlagworte gewählt:

S1: fahren
S2: Zug
S3: Auto
S4: Flugzeug
S5: Fahrrad
S6: Bus

Dazu 5 Dokumente mit folgendem Inhalt:

D1: Auto fahren ist flexibler als mit dem Zug.
D2: Er verkauft sein Fahrrad.
D3: Ein Zug entsteht durch ein offenes Fenster.
D4: Reisen Sie mit dem Auto, Zug, Bus oder Flugzeug, oder wollen Sie doch mit dem Fahrrad fahren?
D5: Das Fahrrad lässt sich im Zug transportieren.

In Matrixschreibweise ergibt sich für die Dokumente:

$$\hat{A} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$A = \begin{pmatrix} 0.5774 & 0 & 0 & 0.4082 & 0 \\ 0.5774 & 0 & 1.0000 & 0.4082 & 0.7071 \\ 0.5774 & 0 & 0 & 0.4082 & 0 \\ 0 & 0 & 0 & 0.4082 & 0 \\ 0 & 1.0000 & 0 & 0.4082 & 0.7071 \\ 0 & 0 & 0 & 0.4082 & 0 \end{pmatrix}$$

In diesem Beispiel wurde als Gewichtung die Normalisierung der Spaltenvektoren gewählt. Die einzelnen Einträge der Vektoren sind somit durch die relative Häufigkeit (innerhalb des Vektors) bestimmt und nicht durch die absolute Anzahl, wie in $\hat{A}$ angegeben. Die Euklidische Norm jeder Spalte ergibt somit $||aj||_2 = 1, \quad j = 1, \ldots, d$. Für eine erfolgversprechende Suche ist die Wahl der Schlagwörter mitentscheidend. Damit nicht unnötig viele Wörter verglichen werden müssen (durch die Kosinusberechnung), werden häufig auch sogenannte Stoppwörter verwendet, die bei der Berechnung keine Rolle mehr spielen. Bei diesen handelt es sich um Artikel, Konjunktionen oder auch Präpositionen, die für das Ergebnis irrelevant sind. Ein weiteres wichtiges Konzept ist das des Stemmings, bei dem ein Wort immer dem Wortstamm zugeordnet wird, um eine bessere Treffergenauigkeit zu erzielen. Dadurch lässt sich Speicherplatz sparen, da weniger Wörter gelistet werden müssen.
Eine erste Suche könnte sich auf den Begriff "Auto fahren" beziehen. Diese Anfrage würde als Vektor folgendermaßen dargestellt:

$$q^{(1)} = (\ 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0\ )^t$$

Nun wird dieser Vektor mit jedem einzelnen Spaltenvektor der Matrix A verglichen, indem jeweils der Kosinus der Winkel ausgerechnet wird. Ein Dokument wird dabei als relevant erachtet, wenn der daraus berechnete Wert des Kosinus einen bestimmten Schwellenwert erreicht. Dieser kann in der Praxis bei ca. 0.9 liegen, zur Vereinfachung wird im Folgenden aber ein Wert von 0.5 angenommen. Für $q^{(1)}$ ergeben sich die Kosinuswerte:

$$q^{(1)}: \quad \alpha_1 = 0.8165 \quad \alpha_2 = 0 \quad \alpha_3 = 0 \quad \alpha_4 = 0.5774 \quad \alpha_5 = 0$$

Die Dokumente eins und vier wurden in diesem Fall als relevant erachtet. Eine weitere Abfrage mit dem Begriff "fahren" würde folgenden Vektor

$$q^{(2)} = (\ 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\ )^t$$

und folgende Ergebnisse liefern:

$$q^{(2)}: \quad \alpha_1 = 0.5774 \quad \alpha_2 = 0 \quad \alpha_3 = 0 \quad \alpha_4 = 0.4082 \quad \alpha_5 = 0$$

Hierbei wird deutlich, dass die einzigen Dokumente, die einen Kosinuswert $\neq 0$ ergeben, wieder die beiden des ersten Beispiels sind. Diesmal wird jedoch das vierte Dokument nicht als relevant erachtet (< 0.5) und damit auch nicht zurückgegeben, obwohl es vermutlich inhaltlich bezüglich des Begriffs "fahren" geeigneter wäre. Eine Suche mit mehreren passenden Begriffen kann daher Vorteile bringen, da die relative Häufigkeit gesuchter Begriffe an der Gesamtzahl der Begriffe des Dokuments zunimmt.

Approximation mit der QR-Zerlegung:

Die QR-Zelegung $A = QR$ existiert für jede Matrix A. Dabei ist Q eine orthogonale sxs-Matrix und R eine sxd-Matrix. Die Zeilen und Spalten von Q sind jeweils orthonormal, daher gilt $Q^tQ = QQ^t = I$. Weiterhin bildet eine Teilmenge der Spalten von Q eine Basis des Spaltenraums von A. Die eindeutige QR-Zerlegung der oben erwähnten Matrix A lautet folgendermaßen:

$$Q = \left(\begin{array}{rrrr|rr} -0.5774 & 0 & -0.4082 & 0 & -0.7071 & 0 \\ -0.5774 & 0 & 0.8165 & 0 & 0 & 0 \\ -0.5774 & 0 & -0.4082 & 0 & 0.7071 & 0 \\ 0 & 0 & 0 & -0.7071 & 0 & -0.7071 \\ 0 & -1.0000 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.7071 & 0 & 0.7071 \end{array}\right)$$

$$R = \left(\begin{array}{rrrrr} -1.0001 & 0 & -0.5774 & -0.7070 & -0.4082 \\ 0 & -1.0000 & 0 & -0.4082 & -0.7071 \\ 0 & 0 & 0.8165 & 0 & 0.5774 \\ 0 & 0 & 0 & -0.5774 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

Beide Matrizen sind zusätzlich noch durch einen Balken in 2 Bereiche aufgeteilt. Dabei wurden bei der oberen Dreiecksmatrix R die Zeilen getrennt, die keine Einträge aufweisen. Daraus lässt sich schließen, dass die letzten beiden Spalten der Orthogonalmatrix Q für die Werteberechnung der Matrix A gar nicht relevant sind. Folglich sind auch diese beiden Spalten aus Gründen der Übersichtlichkeit optisch abgegrenzt. Diese Maßnahme lässt nun folgende Schreibweise zu:

$$A = \begin{pmatrix} Q_A & Q_A^{\perp} \end{pmatrix} \begin{pmatrix} R_A \\ 0 \end{pmatrix} = Q_A R_A + Q_A^{\perp} \cdot 0 = Q_A R_A$$

Aus dieser wird nun ersichtlich, dass nur die Untermatrizen Q_A und R_A zu den Werten von A beitragen. Die Spalten von Q_A bilden daher eine Basis des Spaltenraums von A, während die Einträge der Zeilen von R_A die Koeffizienten der Linearkombinationen dieser Basisvektoren darstellen. Die verbleibenden Spalten von Q ($Q_A^{\perp}$) bilden eine Basis des orthogonalen Komplements des Spaltenraums von A. Weiterhin gilt, dass A und R_A den gleichen Rang haben, welcher sich bei R_A als Anzahl der Zeilen leicht ablesen lässt.

$$r_A := rang(A) = rang(R_A)$$

Zusätzlich ist zu beachten, dass bei der QR-Zerlegung üblicherweise mit Spaltenpivotisierung gearbeitet wird, um eine gewisse Stabilität des Verfahrens zu ermöglichen. Da dabei Spaltenvertauschungen auftreten, müsste die eigentliche Gleichung $AP = QR$ mit P als einer Permutationsmatrix lauten. Diese Spaltenvertauschungen durch die Matrix P ändern aber nichts am Inhalt der Matrix A, da lediglich die Reihenfolge der Spalten (Dokumentenvektoren) getauscht wird. Zur Vereinfachung erfolgt deshalb auf den kommenden Seiten eine Beschränkung auf die Matrix A.
Will man nun die j-te Spalte von A mit Q und R bestimmen, dann ergibt sich mit r_j als j-te Spalte von R:

$$a_j = Ae_j = Q_A r_j$$

Auf die Winkelberechnung hat die QR-Zerlegung nun folgende Auswirkungen:

$$\cos \alpha_j = \frac{a_j^t q}{||a_j||_2||q||_2} = \frac{(Q_A r_j)^t q}{||Q_A r_j||_2||q||_2} = \frac{r_j^t (Q_A^t q)}{||r_j||_2||q||_2}$$

Für $j = 1, \ldots, d$
Dabei wurde berücksichtigt, dass $||Q_A r_j||_2 = ||r_j||_2$, da Q_A orthonormale Spalten besitzt.
Für den Anfragevektor $q^{(1)}$ "Auto fahren" ergeben sich die selben Kosinuswerte wie zuvor, wobei diesmal mit den Matrizen Q und R gerechnet wurde. Durch die QR-Zerlegung findet also kein Informationsverlust statt.
Die Matrix Q lässt sich folgendermaßen in ihre Basen (des orthogonalen Komplements) des Spaltenraums von A aufteilen:

$$I = QQ^t = \begin{pmatrix} Q_A & Q_A^\perp \end{pmatrix} \begin{pmatrix} Q_A & Q_A^\perp \end{pmatrix}^t = Q_A Q_A^t + Q_A^\perp (Q_A^\perp)^t$$

Der Eingabevektor q lässt sich dann als Summe der Komponenten von diesen beiden zueinander orthogonalen Räumen von A schreiben:

$$q = Iq = QQ^t q = (Q_A Q_A^t + Q_A^\perp (Q_A^\perp)^t)q = Q_A Q_A^t q + Q_A^\perp (Q_A^\perp)^t q = q_A + q_A^\perp$$

Dabei ist der Ausdruck $q_A = Q_A Q_A^t q$ die orthogonale Projektion von q in den Spaltenraum von Q_A. Weiterhin ist q_A sogar die beste Annäherung von q im Spaltenraum von A.

$$||q - q_A||_2 \; = \; min \; \{||q - x||_2, \; x \; aus \; Spaltenraum \; A\}$$

$\Rightarrow q_a$ Bestapproximation zu x im Spaltenraum von A.

Beweis mit dem Satz des Pythagoras:

$$||q - x||_2^2 = ||q - q_A + q_A - x||_2^2 = ||q - q_A||_2^2 + ||q_A - x||_2^2 \geq ||q - q_A||_2^2$$

q kann nun in Komponentenschreibweise in die Formel zur Berechnung des Kosinuswerts eingesetzt werden. Dann ergibt sich:

$$\cos \alpha_j = \frac{a_j^t q_A + a_j^t q_A^\perp}{||a_j||_2||q||_2} = \frac{a_j^t q_A + a_j^t Q_A^\perp (Q_A^\perp)^t q}{||a_j||_2||q||_2}$$

Da $Q_A^\perp$ orthogonal zu den Spalten von Q_A und damit auch orthogonal zu den Spalten von A liegt (Spalten von Q_a sind Basis von A), ist $Q_A^\perp$ orthogonal zu einer beliebigen Spalte a_j. Dies impliziert $a_j^t Q_A^\perp = 0$. Dadurch lässt sich die Winkelberechnung vereinfachen auf:

$$cos\alpha_j = \frac{a_j^t q_A + 0 \cdot (Q_A^{\perp})^t q}{||a_j||_2||q||_2} = \frac{a_j^t q_A}{||a_j||_2||q||_2}$$

Dieses Ergebnis entstand allein durch Einsetzen der Komponentenschreibweise von q in die Formel zur Winkelberechnung. Daraus könnte man ableiten, dass bei der Dokumentensuche die Terme automatisch durch eine beste Annäherung aus dem Inhalt der Datenbank ersetzt werden. Man könnte die Formel sogar ein weiteres Mal verbessern, indem ausschließlich die orthogonale Projektion des Eingabevektors q verwendet wird.

$$\cos \alpha_j' = \frac{a_j^t q_A}{||a_j||_2||q_A||_2}$$

Es gelten zudem die folgenden Beziehungen:

$$\cos \alpha_j = \cos \alpha_j' \frac{||q_A||_2}{||q||_2} = \cos \alpha_j' \frac{||q_A||_2}{\sqrt{||q_A||_2^2 + ||q_A^{\perp}||_2^2}}$$

Der Bruch auf der rechten Seite ist immer kleiner gleich 1 und echt kleiner 1, wenn $q_A^{\perp} \neq 0$. Dadurch werden die Kosinuswerte $\cos \alpha'$, denen die Bestapproximation q_A bei der Berechnung zugrunde liegt, größer oder gleich den Kosinuswerten $\cos \alpha$, die aus dem unmodifizierten Eingabevektor q entstanden sind. Die orthogonale Projektion q_A ist somit geometrisch betrachtet näher an den Dokumentenvektoren als q. Dies kann zu einer wachsenden Anzahl zurückgegebener relevanter Dokumente führen, erhöht aber auch die Wahrscheinlichkeit, dass nicht relevante Dokumente zurückgegeben werden.

Niedrigrang-Approximation:

Die QR-Zerlegung ist im Information Retrieval nicht nur bei der Winkelberechnung hilfreich, sie spielt auch bei der Niedrigrang-Approximation eine wichtige Rolle. Niedrigrang-Approximationen dienen dazu, belanglose Informationen im Inhalt der Datenbank zu entfernen, um dadurch die Effektivität der Berechnung zu steigern. Zusätzlich können Ungenauigkeiten, die beim Messen der Daten oder bei der Indexierung entstanden sind, vermieden werden. Eine geeignetere Matrix wäre somit die Summe der Matrix A mit einer Matrix E, die die Ungenauigkeiten berücksichtigt. Die Matrix $A + E$ hätte dann einen niedrigeren Rang als Matrix A.
Um Aussagen über das Maß des Inhalts einer Matrix treffen zu können, benötigt man noch eine weitere Größe. Diese erhält man, indem man die Euklidische Norm auf Matrizen verallgemeinert. Die daraus entstandene Norm wird Frobeniusnorm genannt und ist für eine sxd-Matrix definiert als:

$$||X||_F = \sqrt{\sum_{i=1}^{s} \sum_{j=1}^{d} x_{ij}^2}$$

Es soll nun eine Niedrigrang-Approximation an der Matrix A durchgeführt werden, wodurch sich der Rang von A verringert. Dazu betrachtet man zunächst die obere Dreiecksmatrix R, da sie den gleichen Rang wie A besitzt. Die Spaltenpivotisierung während der QR-Zerlegung erweist sich an dieser Stelle als überaus nützlich, da während dieser große Einträge eher in den linken oberen Bereich, kleine eher in den rechten unteren Bereich platziert wurden. Folglich kann nun ein Bereich mit relativ kleinen Einträgen markiert werden, indem Zeilen im unteren Bereich mit Einträgen $\neq 0$ von den darüber liegenden Zeilen getrennt werden.

$$R = \left(\begin{array}{rrr|rr} -1.0001 & 0 & -0.5774 & -0.7070 & -0.4082 \\ 0 & -1.0000 & 0 & -0.4082 & -0.7071 \\ 0 & 0 & 0.8165 & 0 & 0.5774 \\ \hline 0 & 0 & 0 & -0.5774 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right) = \left(\begin{array}{cc} R_{11} & R_{12} \\ 0 & R_{22} \end{array}\right)$$

Die Untermatrix R_{22} kann nun mit Hilfe der Frobeniusnorm mit der ganzen Matrix verglichen werden. Es ergibt sich $||R_{22}||_F = 0.5774$, $||R||_F = 2.2361$ und ein Verhältnis von $||R_{22}||_F/||R||_F = 0.2582$.
Durch Nullsetzen der Einträge in R_{22} kann eine neue obere Dreiecksmatrix $\hat{R}$ konstruiert werden, die Rang 3 besitzt. Damit ist auch die neu aus A entstandene Matrix $A + E = Q\hat{R}$ von Rang 3. Die Matrix E ist gegeben durch:

$$E = (A + E) - A = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix} - Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix} = Q \begin{pmatrix} 0 & 0 \\ 0 & -R_{22} \end{pmatrix}$$

Durch die Rangreduzierung wurde die Matrix R um ca. 26% ($||R_{22}||_F/||R||_F = 0.2582$) verändert. Diese Veränderung gilt auch für A, da:

$$||A||_F = ||R||_F \text{ und } ||E||_F/||A||_F = ||R_{22}||_F/||R||_F = 0.2582$$

Somit reicht eine 26%-ige Veränderung (tritt in A und in B auf), um den Rang um eins zu erniedrigen (in A und B). Zur Berechung müssen nur noch die ersten drei Spalten von Q und $\hat{R}$ verwendet werden. Für die Kosinuswerte ergeben sich mit der Matrix $A + E$, deren Rang um eins erniedrigt wurde, und den Anfragevektoren $q^{(1)}$, $q^{(2)}$ diese neuen Werte:

cos	α_1	α_2	α_3	α_4	α_5
$q^{(1)}(neu)$	0.8165	0	0	0.7071	0
$q^{(1)}(alt)$	0.8165	0	0	0.5774	0

cos	α_1	α_2	α_3	α_4	α_5
$q^{(2)}(neu)$	0.5774	0	0	0.5	0
$q^{(2)}(alt)$	0.5774	0	0	0.4082	0

In beiden Fällen hat die Rangreduzierung das Ergebnis verbessern können; speziell für den zweiten Anfragevektor wird nun ein Dokument als relevant erachtet, dessen Wert zuvor noch unter dem Schwellenwert gelegen hat. Die Rangreduzierung soll nun einen Schritt weiter gehen, sodass auch Zeile 3 und Spalte 3 zur Untermatrix R_{22} gehören und die veränderte obere Dreiecksmatrix (und $A + E$) Rang 2 besitzt. Es ergibt sich eine relative Änderung der Matrix R von $||R_{22}||_F/||R||_F = 0.5146 \Rightarrow$ ca. 52%. Die Kosinuswerte betragen für $q^{(1)}$, $q^{(2)}$:

$$\begin{array}{rccccc} q^{(1)} = & 0.8165 & 0 & 0.8165 & 0.7071 & 0.4082 \\ q^{(2)} = & 0.5774 & 0 & 0.5774 & 0.5000 & 0.2887 \end{array}$$

Bei beiden Anfragen werden nun auch irrelevante Dokumente als wichtig erachtet. Die 52%ige Veränderung bzw. die Reduzierung des Rangs auf 2 hat das Ergebnis deutlich verschlechtert, was darauf schließen lässt, dass Reduzierungen auf einen Rang < 3 in diesem Fall nicht mehr sinnvoll sind. Im Allgemeinen lässt sich nicht vorhersagen, ob und bis zu welchem Rang eine Rangreduzierung sinnvoll ist. Sie kann aber, wie beschrieben, Vorteile bringen. Die dazu im Beispiel benötigte 26%ige Änderung der Matrix fällt im Vergleich zu wissenschaftlichen Anwendungen jedoch sehr hoch aus, in denen oft nur Änderungen von 0.1% eine Rolle spielen.

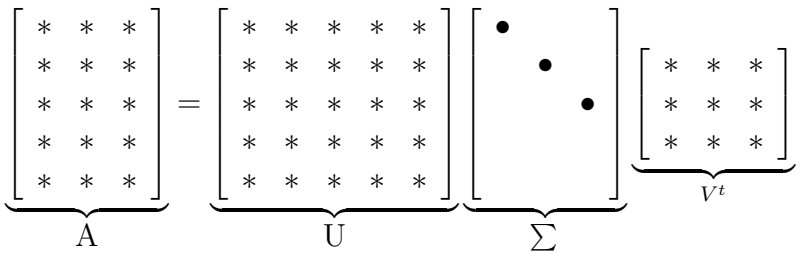
Approximation mit der Singulärwertzerlegung:

Unter Hinzunahme der QR-Zerlegung kann der Rang der Matrix A oder genauer der Rang des Spaltenraums der Matrix A verringert werden. Der Zeilenraum bleibt dabei unverändert. Mit Hilfe der Singulärwertzerlegung dagegen kann zusätzlich auch der Zeilenraum verringert werden. Die Berechnungen für dieses Verfahren sind jedoch etwas aufwändiger. Die Singulärwertzerlegung einer Matrix A liefert folgenden neuen Ausdruck:

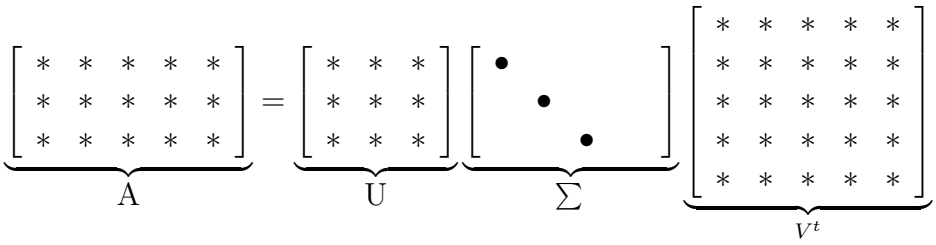
$$A = U \sum V^t$$

Für jede Matrix A existiert die Singulärwertzerlegung. Die Singulärwerte sind dabei eindeutig. Die sxs-Matrix U ist orthogonal, genauso wie die dxd-Matrix V. $\sum$ ist eine sxd-Diagonalmatrix, auf deren Diagonalen die Singulärwerte der Größe nach geordnet stehen: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{min_{(s,d)}}$.

Für $s > d$ ergibt die Singulärwertzerlegung:



Für $s < d$ ergibt sich:



Der Rang r_A einer Matrix A gibt die Anzahl der Singulärwerte $\neq 0$ an. Es gilt:

$$||A||_F = ||U \sum V^t||_F = || \sum V^t||_F = || \sum ||_F = \sqrt{\sum_{j=1}^{r_A} \sigma_j^2}$$

Genauso wie die ersten r_A Spalten von Q (bei der QR-Zerlegung) eine Basis des Spaltenraums von A bildeten, so bilden auch die ersten r_A Spalten von U eine Basis des Spaltenraums von A. Für eine Rang-k-Approximation A_k gilt (wie bei der QR-Zerlegung), dass anstelle aller Zeilen außer die ersten k, alle Singulärwerte außer die ersten (bzw. größten) k auf Null gesetzt werden. A_k lässt sich somit durch die Gleichung $A_k = U_k \sum_k V_k$ berechnen, wobei U_k die Matrix aus den ersten k Spalten von U, V_k die Matrix aus den ersten k Spalten von V und $\sum_k$ eine kxk-Diagonalmatrix mit den ersten k Singulärwerten darstellt. Diese Verkleinerung der Matrizen U, $\sum$ und V rührt (wie bei der QR-Zerlegung) daher, dass durch das Nullsetzen aller Singulärwerte außer der ersten k, alle Spalten u_j von U und v_j von V mit $j > k$ für die Werteberechnung von A_k nicht mehr relevant sind. Die Singulärwertzerlegung für die Beispielsmatrix A ergibt:

$$\mathrm{U}=\begin{pmatrix} 0.2670 & -0.2567 & 0.5308 & -0.2847 & -0.7071 & 0 \\ 0.7479 & -0.3981 & -0.5249 & 0.0816 & 0 & 0 \\ 0.2670 & -0.2567 & 0.5308 & -0.2847 & 0.7071 & 0 \\ 0.1182 & -0.0127 & 0.2774 & 0.6394 & 0 & -0.7071 \\ 0.5198 & 0.8423 & 0.0838 & -0.1158 & 0 & 0 \\ 0.1182 & -0.0127 & 0.2774 & 0.6394 & 0 & 0.7071 \end{pmatrix}$$

$$\sum = \begin{pmatrix} 1.6950 & 0 & 0 & 0 & \\ 0 & 1.1158 & 0 & 0 & 0 \\ 0 & 0 & 0.8403 & 0 & 0 \\ 0 & 0 & 0 & 0.4195 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 0.4366 & -0.4717 & 0.3688 & -0.6715 & 0 \\ 0.3067 & 0.7549 & 0.0998 & -0.2760 & -0.5000 \\ 0.4412 & -0.3568 & -0.6247 & 0.1945 & -0.5000 \\ 0.4909 & -0.0346 & 0.5711 & 0.6571 & 0 \\ 0.5288 & 0.2815 & -0.3712 & -0.0577 & 0.7071 \end{pmatrix}$$

Es wird deutlich, dass die Matrix A mit Rang(A)=4 vier Singulärwerte $\neq 0$ hat. Außerdem lassen die beiden unteren Null-Zeilen von $\sum$ darauf schließen, dass die ersten vier Spalten von U eine Basis des Spaltenraums von A bilden.
Der Unterschied der Matrix A zu seiner Approximation A_k lässt sich durch Umformen auf einen Term bringen, der nur die Singulärwerte beinhaltet:

$$||A - A_k||_F = \min_{rank(X)\leq k} ||A - X||_F = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_{r_A}^2}$$

Hierbei bleiben genau die Singulärwerte, die für eine beliebige Approximation A_k von A auf Null gesetzt werden mussten, als Größe für die Differenz übrig. Der Unterschied kann somit als die Euklidische Norm dieser Singulärwerte, aufgefasst als Vektor, betrachtet werden. Unter Verwendung dieser Gleichung ergibt sich für den Unterschied zwischen A und einer A_3 Approximation: $||A - A_3||_F = \sigma_4 = 0.4195$. Und mit $||A||_F = 2.2361$ das Verhältnis: $||A - A_3||_F/||A||_F \approx 0.1876$. Mit A_2 ergibt sich: $||A - A_2||_F/||A||_F \approx 0.4195$. Es ist also eine Veränderung von ca. 19% nötig, um den Rang von vier auf drei zu verringern. Für die Verringerung auf Rang zwei wäre eine Veränderung von 42% notwendig. Um den Rang über die QR-Zerlegung auf drei zu reduzieren, war eine Veränderung von 26% notwendig; mit der Singulärwertzerlegung sind es nun 19%. Die Verringerung auf Rang 2 machte mit dem QR-Verfahren eine Veränderung von 52% nötig, mit der Singulärwertzerlegung lässt sich dies auch mit einer um 10% geringeren Änderung durchführen. Die Singulärwertzerlegung ist deshalb, was die Genauigkeit anbelangt, das geschicktere Verfahren. Schaut man sich nun die entsprechenden Rang-3-Approximationen an,

Originalmatrix A:

$$A = \begin{pmatrix} 0.5774 & 0 & 0 & 0.4082 & 0 \\ 0.5774 & 0 & 1.0000 & 0.4082 & 0.7071 \\ 0.5774 & 0 & 0 & 0.4082 & 0 \\ 0 & 0 & 0 & 0.4082 & 0 \\ 0 & 1.0000 & 0 & 0.4082 & 0.7071 \\ 0 & 0 & 0 & 0.4082 & 0 \end{pmatrix}$$

Approximation mit der QR-Zerlegung:

$$\tilde{A} = \begin{pmatrix} 0.5774 & 0 & 0 & 0.4082 & 0 \\ 0.5774 & 0 & 1.0000 & 0.4082 & 0.7071 \\ 0.5774 & 0 & 0 & 0.4082 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1.0000 & 0 & 0.4082 & 0.7071 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Approximation mit der Singulärwertzerlegung:

$$A_3 = \begin{pmatrix} 0.4971 & -0.0330 & 0.0232 & 0.4867 & -0.0069 \\ 0.6003 & 0.0094 & 0.9933 & 0.3858 & 0.7091 \\ 0.4971 & -0.0330 & 0.0232 & 0.4867 & -0.0069 \\ 0.1801 & 0.0740 & -0.0522 & 0.2320 & 0.0155 \\ -0.0326 & 0.9866 & 0.0094 & 0.4402 & 0.7043 \\ 0.1801 & 0.0740 & -0.0522 & 0.2320 & 0.0155 \end{pmatrix}$$

lässt sich erkennen, dass eine bessere Approximation nicht zwangsläufig der Originalmatrix ähnlicher sein muss. In diesem Fall gleicht die Rang-3-Approximation mit der QR-Zerlegung eher der Originalmatrix A als die eigentlich geeignetere Approximation A_3, die auf der Singulärwertzerlegung basiert.
Für die Winkelberechnungen zwischen einem Anfragevektor q und den Dokumentenvektoren der approximierten Matrix A_k mit j-ter Spalte $A_k e_j$ ergibt sich:

$$\cos \alpha_j = \frac{(A_k e_j)^t q}{||A_k e_j||_2 ||q||_2} = \frac{(U_k \sum_k V_k^t e_j)^t q}{||U_k \sum_k V_k^t e_j||_2 ||q||_2} = \frac{e_j^t V_k \sum_k (U_k^t q)}{||\sum_k V_k^t e_j||_2 ||q||_2}$$

für $j = 1, \ldots, d$.
Anstatt aufwändig eine Rang-k-Approximation A_k zu berechnen, müssen nun nur noch die ersten k Spalten von V,$\sum$ und U betrachtet werden.
Mit $s_j := \sum_k V_k^t e_j$ vereinfacht sich die Gleichung zu:

$$\cos \alpha_j = \frac{s_j^t (U_k^t q)}{||s_j||_2 ||q||_2}, \qquad j = 1, \ldots, d \qquad (*)$$

Diese s_j können nun einmal pro sxd-Matrix bestimmt werden, sodass nicht ständig $\sum_k V_k^t e_j$ bei jedem Vektorenvergleich berechnet werden muss. Weiterhin sind die k Einträge des Vektors s_j die Koordinaten der j-ten Spalte von A_k, ausgedrückt in der Basis, die die ersten k Spalten von U (U_k) bilden. Auch die k Einträge des Vektors $U_k^t q$ sind Koordinaten zur selben Basis U_k und ergeben mit den Basisvektoren den Vektor $U_k U_k^t q$. Dieser ist, wie bei der QR-Zerlegung gesehen, die orthogonale Projektion von q und damit die beste Annäherung von q im Spaltenraum von A_k. Es kann nun eine Winkelberechnung durchgeführt werden, die ausschließlich die näher an den Dokumentenvektoren befindliche orthogonale Projektion von q verwendet:

$$\cos \alpha'_j = \frac{s_j^t (U_k^t q)}{||s_j||_2 ||U_k^t q||_2}, \qquad j = 1, \ldots, d$$

Nach einmaliger Berechnung von $U_k^t q$ treten dabei für die weitere Berechnung nur noch k-dimensionale Vektoren auf, was die Berechnungszeit ein wenig reduziert. Für die Eingabevektoren $q^{(1)}$ und $q^{(2)}$ ergeben sich mit einer Rang-3-Approximation, basierend auf der Singulärwertzerlegung, und der Berechnungsmethode $(*)$ folgende Ergebnisse:

$$q^{(1)}: \quad \alpha_1 = 0.7327 \quad \alpha_2 = -0.0469 \quad \alpha_3 = 0.0330 \quad \alpha_4 = 0.7161 \quad \alpha_5 = -0.0097$$

Das erste und vierte Dokument wurden wieder richtig identifiziert, dieses Mal mit einem fast identischen Wert. Die übrigen Werte sind nun nicht mehr Null, haben aber dennoch einen vernachlässigbar kleinen Wert.

$$q^{(2)}: \quad \alpha_1 = 0.5181 \quad \alpha_2 = -0.0332 \quad \alpha_3 = 0.0233 \quad \alpha_4 = 0.5064 \quad \alpha_5 = -0.0069$$

Bei Anfrage von $q^{(2)}$ werden wie bei der Rang-3-Approximation unter Verwendung der QR-Zerlegung das erste und vierte Dokument zurückgegeben. Im Vergleich dazu aber nun mit einer geringeren Differenz zwischen den Werten der beiden Vektoren.
Mit einer Rang-2-Approximation und derselben Methode ergeben sich für $q^{(1)}$ diese Werte:

$$q^{(1)}: \quad \alpha_1 = 0.5181 \quad \alpha_2 = -0.1107 \quad \alpha_3 = 0.5038 \quad \alpha_4 = 0.3940 \quad \alpha_5 = 0.2362$$

Nun wird das dritte Dokument, das vermutlich am wenigsten gewünscht wurde, zusammen mit dem ersten als relevant erachtet, während der Wert des vierten Dokuments unterhalb des Schwellenwerts bleibt. Die Werte von $q^{(2)}$ liegen alle unterhalb des Schwellenwerts.

Der Term-Term Vergleich:

Wie anfangs erwähnt kann über die Singulärwertzerlegung nicht nur eine Approximation des Spaltenraums, sondern auch eine des Zeilenraums gebildet werden. Damit lassen sich nun Vergleiche der Schlagworte (Term-Term-Vergleiche) durchführen, die auf dieselbe Weise wie bei den vorangegangenen Verfahren auf (diesmal: Zeilenrang-)Approximationen beruhen.
Die aus einer Begriffssuche resultierenden Dokumente können nach weiteren Begriffen, die in engem Zusammenhang zum Suchbegriff stehen, durchsucht werden. Dabei kann der Zusammenhang aus der bloßen Anzahl der Erwähnungen in den gefundenen Dokumenten bestehen. Aus den Dokumenten, dem Suchbegriff und den weiteren Begriffen kann wieder eine sxd -Matrix nach Prinzip der Beispielsmatrix A aufgestellt werden.
Ein Beispiel für den eingegebenen Begriff "Laufen" wären

die 5 Dokumente:

D1: Laufen, Schwimmen und Gehen ist ein gutes Training.
D2: Die Maschinen laufen einwandfrei.
D3: Das Training für heute sieht Laufen vor.
D4: Sie werden Film und Musik laufen lassen.
D5: Durch Training können Fähigkeiten, wie das Laufen, verbessert werden.

und die daraus entnommenen Begriffe:

S1: Laufen
S2: Schwimmen
S3: Gehen
S4: Training
S5: Film
S6: Musik
S7: Maschinen

Daraus ergibt sich folgende Matrix G:

$$G = \begin{pmatrix} 0.5000 & 0.7071 & 0.7071 & 0.5774 & 0.7071 \\ 0.5000 & 0 & 0 & 0 & 0 \\ 0.5000 & 0 & 0 & 0 & 0 \\ 0.5000 & 0 & 0.7071 & 0 & 0.7071 \\ 0 & 0 & 0 & 0.5774 & 0 \\ 0 & 0 & 0 & 0.5774 & 0 \\ 0 & 0.7071 & 0 & 0 & 0 \end{pmatrix}$$

Der eigentliche Term-Term-Vergleich findet nun zwischen den Zeilen dieser Matrix statt, die jeweils einen Begriff darstellen. Wiederum werden die Winkel bzw. die Kosinuswerte zwischen den Vektoren, in diesem Fall Zeilenvektoren, ausgerechnet:

$$\cos \omega_{ij} = \frac{(e_i^t G)(G^t e_j)}{||G^t e_i||_2 ||G^t e_j||_2}, \qquad \text{für } i, j = 1, \ldots, s$$

Mit diesen Werten lässt sich eine Tabelle K mit den Einträgen $K_{ij} = \cos \omega_{ij}$ aufstellen, aus der sich folgern lässt, welcher Begriff welchem am nächsten ist. Hohe Werte nahe der 1 deuten auf einen engen Zusammenhang hin, während Werte bei Null auf orthogonale Vektoren schließen lassen und sich die Begriffe damit nicht ähneln.

$K =$

1.0000	0.3464	0.3464	0.7746	0.4000	0.4000	0.4899
	1.0000	1.0000	0.4472	0	0	0
		1.0000	0.4472	0	0	0
			1.0000	0	0	0
				1.0000	1.0000	0
					1.0000	0
						1.0000

(Aus Gründen der Übersichtlichkeit sind nur die oberen Einträge der symmetrischen Tabelle K eingetragen.)

Aus der Tabelle lässt sich entnehmen, dass die Begriffe (ausgenommen des Suchbegriffs) untereinander Gruppen bilden, in denen sich wieder jeweils Begriffe in eigenem Zusammenhang befinden. So ergeben z.B. die Zeilenvektoren mit den Begriffen "Training" und "Musik" den Wert 0, während die Vektoren von "Training" und "Schwimmen" den Wert 0.4472 ergeben. Dieser Vorgang des Einteilens in Gruppen mit spezifischem Inhalt wird auch Clustern genannt.

Die Kosinusberechnung mit einer Niedrigrang-Approximation des Zeilenraums von G erfolgt durch Einsetzen von $G = U_k \sum_k V_k^t$ nach folgender Gleichung:

$$\cos \omega_{ij} = \frac{(e_i^t U_k \sum_k V_k^t)(V_k \sum_k U_k^t e_j)}{||V_k \sum_k U_k^t e_i||_2 ||V_k \sum_k U_k^t e_j||_2} = \frac{(e_i^t U_k \sum_k)(\sum_k U_k^t e_j)}{||\sum_k U_k^t e_i||_2 ||\sum_k U_k^t e_j||_2},$$

für i=1, ...,s und j=1, ...,d

Auf die gleiche Weise wie die Spalten von U_k eine Basis des Spaltenraums von A_k darstellten, bilden die Zeilen von V_k eine Basis des Zeilenraums von A_k (in diesem Fall: G_k).
Mit der Vereinfachung $b_j = \sum_k U_k^t e_j$ ergibt sich:

$$\cos \omega_{ij} = \frac{b_i^t b_j}{||b_i||_2 ||b_j||_2}, \qquad \text{für } i = 1, \ldots, s \text{ und } j = 1, \ldots, d$$

Das Prinzip des Clusterns kommt vor allem bei der Polysemie zum Tragen. So könnte bei einer Suchanfrage nach Identifizierung des Nutzers festgestellt werden, nach welcher Auffassung des mehrdeutigen Begriffs gesucht wird.

Neben dem Vektorraummodell spielen in der Praxis hauptsächlich noch das Boolesche Modell und das probabilistische Modell eine wichtige Rolle. Das Boolesche Modell zeichnet sich im Vergleich zu den anderen beiden durch das Prinzip des "exact match" aus, der exakten Übereinstimmung zwischen Anfrage und Dokument. Zur Formulierung der Suchanfrage stehen dabei die drei booleschen Operatoren AND, OR und NOT zur Verfügung. Das probabilistische Modell arbeitet auf der Grundlage von Wahrscheinlichkeiten und berechnet diese aus der Häufigkeit der Suchbegriffe im Dokument.
Heutige Suchmaschinen nutzen in der Regel Verfahren, die Eigenarten der verschiedenen Modelle kombinieren. Das Vektorraummodell ist in der Hinsicht insofern von Bedeutung, als dass es durch die Winkelberechnung ein Ranking der relevanten Dokumente liefert.