

6 Ähnlichkeitsmaße

Ähnlichkeitsmaße dienen zur Bewertung von Ähnlichkeiten zwischen Medienobjekten. Nach einer Einführung im folgenden Abschnitt werden Ähnlichkeitswerte Distanzwerten gegenübergestellt. Generelle Grenzen von Ähnlichkeitsmaßen werden danach diskutiert. Im Anschluss daran werden konkrete Ähnlichkeitsmaße vorgestellt.

Häufig müssen Ähnlichkeitswerte aggregiert werden. Dazu werden spezielle Verfahren diskutiert. Sollen Ähnlichkeitswerte auf Distanzwerten basieren, ist eine Wertumwandlung und Normierung erforderlich. Ein Abschnitt über partielle Ähnlichkeit sowie Literaturempfehlungen schließen das Kapitel ab.

6.1 Einführung

Ein wichtiges Modell aus der Psychologie zur menschlich wahrgenommenen Ähnlichkeit geht davon aus, dass Objekte als ähnlich wahrgenommen werden, wenn sie bei Menschen zu ähnlichen Reizen (engl. stimuli) führen. Wir gehen hier vereinfachend davon aus, dass die relevanten Reize den extrahierten und aufbereiteten Feature-Werten entsprechen.

Feature-Werte als Stimuli

In Abbildung 3.5 auf Seite 108 wurde die Konstruktion eines RSV-Wertes eines Datenbankobjektes bezüglich einer komplexen Anfrage skizziert. In den ersten Schritten werden Feature-Werte des Datenbankobjektes und der Anfrage mittels Distanzfunktionen oder Ähnlichkeitsmaßen verglichen und in den letzten Schritten zu Ähnlichkeitswerten aggregiert.

Eine allgemein akzeptierte und exakte Definition des Begriffs Ähnlichkeit gibt es nicht und wird es wahrscheinlich auch in Zukunft nicht geben. In vielen Wissenschaftsgebieten, wie zum Beispiel Psychologie, Mathematik, Statistik, Bildverarbeitung und Mustererkennung wird zu diesem Thema geforscht und es wurden viele unterschiedliche Ähnlichkeitsmodelle entwickelt.

unterschiedliche Ähnlichkeitsmodelle

Definition 6.1

Ein Ähnlichkeitsmaß ist eine Funktion, die einem Paar von Objekten eine Zahl aus dem reellen Intervall $[0, 1]$ zuordnet. Dabei korrespondiert der Wert 1 zur maximalen Ähnlichkeit und der Wert 0 zur maximalen Unähnlichkeit.

Signatur eines Ähnlichkeitsmaß

Diese Definition ist insofern nicht ausreichend, da sie nur die Signatur des

Ähnlichkeitsmaßes festlegt, jedoch keine Aussagen über die konkrete Abbildung trifft.

In den beiden folgenden Unterabschnitten werden wir zunächst zwei Aspekte von Ähnlichkeitsmaßen diskutieren:

- Distanz versus Ähnlichkeit und
- Grenzen von Ähnlichkeitsmaßen.

Danach gehen wir auf konkrete Ähnlichkeitsmaße ein.

6.2 Distanz versus Ähnlichkeit

*Ähnlichkeit
basierend auf
Distanzen*

Viele Ansätze im Bereich Multimedia-Ähnlichkeit verwenden eine Distanzfunktion auf Feature-Werten als Grundlage eines Ähnlichkeitsmaßes, wobei die Distanzwerte auf das Ähnlichkeitsintervall $[0, 1]$ abgebildet werden. Eine generelle Frage ist, inwieweit eine Distanzfunktion tatsächlich als Grundlage geeignet ist.

*Distanzeigenschaften
zu restriktiv*

Basiert ein Ähnlichkeitsmaß auf einer Distanzfunktion, wird damit automatisch vorausgesetzt, dass Ähnlichkeit den Gesetzen einer Distanzfunktion folgt. Zu dieser Fragestellung wurden in der Psychologie Untersuchungen vorgenommen, die belegen, dass die Distanzeigenschaften im Allgemeinen für das menschliche Ähnlichkeitsempfinden zu restriktiv sind. Das bedeutet nicht automatisch, dass Distanzfunktionen für Ähnlichkeitsmaße generell ungeeignet sind. Statt dessen sind sie nur nicht grundsätzlich für alle Anwendungen geeignet.

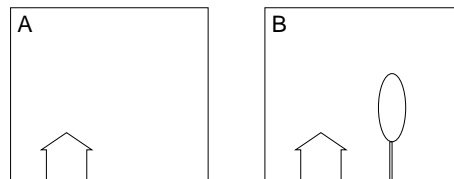
Folgende Probleme können mit den einzelnen Distanzeigenschaften auftreten:

- *Selbstidentität*: Diese Eigenschaft, formuliert als $d(A, A) = 0$ für ein beliebiges Objekt A , gilt nach Untersuchungen von Krumhansl in [112] nicht grundsätzlich.
- *Positivität*: Diese Eigenschaft wird von Tversky in [208] als allgemeine Bedingung für menschliches Ähnlichkeitsempfinden widerlegt.
- *Symmetrie*: Beim Ähnlichkeitsempfinden zwischen einem Suchbild und einem Datenbankbild macht es einen Unterschied, wenn die beiden Bilder ihre Rollen tauschen.
- *Dreiecksungleichung*: Menschen neigen oft dazu, Unterschiede zwischen Vergleichsobjekten zu hoch zu bewerten, wenn kein drittes Referenzobjekt zum Vergleich erfahrbar ist.

Die Verletzung der ersten beiden Eigenschaften ist nicht intuitiv verständlich. Wir verweisen dafür auf die entsprechende Literatur. Die Erfüllung der beiden anderen Eigenschaften kann jedoch anschaulich widerlegt werden.

Beispiel 6.1*fehlende Symmetrie*

Sucht man in einer Bilddatenbank ein Bild A mit einem Haus, dann akzeptiert man in der Regel ein Datenbankbild B, das neben dem Haus weitere Objekte zeigt. Sucht man jedoch anhand des Datenbankbildes B und erhält das ursprüngliche Suchbild A, wird der Baum vermisst. Dies resultiert in einer geringeren Ähnlichkeit und verletzt damit die Symmetrie. Dieses Szenario ist in Abbildung 6.1 dargestellt.

**Abb. 6.1:** Symmetrieprobleme

Der Grund für die mangelnde Symmetrie wird allgemein darin gesehen, dass sich Bilder danach unterscheiden, wie stark und wieviele ihrer Eigenschaften hervortreten (engl. salient feature). Allgemein gilt, wenn in Bild B die Eigenschaften mehr als in Bild A hervortreten, dann ist die Ähnlichkeit zwischen beiden Bildern größer, wenn A als Suchbild benutzt wird, als umgekehrt. Genaugenommen liegen implizit zwei unterschiedliche Anfragen vor: „Finde alle Bilder, auf denen mindestens ein Haus zu sehen ist“ versus „Finde alle Bilder, auf denen mindestens ein Haus und ein Baum abgebildet ist“.

*hervortretende
Eigenschaften*

Die Erfüllung der Eigenschaft der Dreiecksungleichung kann ebenfalls relativ anschaulich widerlegt werden. Am besten kann dieses Phänomen an einem Beispiel demonstriert werden.

Beispiel 6.2*verletzte
Dreiecksungleichung*

Vergleicht man in der Abbildung 6.2 die Grafiken A und B, stellt man kaum Gemeinsamkeiten fest. Jedoch sind Gemeinsamkeiten jeweils zum Objekt C vorhanden. Der Mensch neigt oft dazu, im Einzelvergleich die Unähnlichkeit zwischen A und B stärker als die Summe der jeweiligen Unähnlichkeiten zu Objekt C einzuschätzen ($d(A, B) > d(A, C) + d(B, C)$).

Die geschilderten Probleme zeigen, dass Distanzeigenschaften nicht generell auf Ähnlichkeitsmaße übertragen werden können. In vielen Anwendungen wurden trotzdem gute Erfahrungen mit Ähnlichkeitsberechnungen auf der Grundlage von Distanzfunktionen gemacht.

Da die Distanzeigenschaften im Allgemeinen als zu restriktiv erachtet werden, wird im Folgenden statt einer Distanz der Begriff des Ähnlichkeitsabstandes verwendet. Wir unterscheiden diesen Begriff vom Ähnlichkeitsmaß, da

Ähnlichkeitsabstand

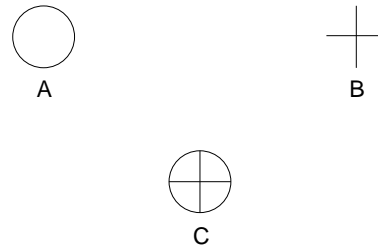


Abb. 6.2: Probleme mit der Dreiecksungleichung

der Ähnlichkeitsabstand ein Unähnlichkeitsmaß ist und erst durch Anwendung einer speziellen Umkehrfunktion auf ein Ähnlichkeitsmaß abgebildet werden kann.

*Eigenschaften eines
Ähnlichkeitsabstandes*

Ein wichtiger Beitrag über die Eigenschaften eines Ähnlichkeitsmaßes auf der Grundlage eines Ähnlichkeitsabstands wurde durch die Arbeiten von Tversky und Gati [207] geleistet. Sie fordern die folgenden drei Eigenschaften, die ein Ähnlichkeitsabstand d mindestens erfüllen muss. Dabei werden als Grundlage Feature-Werte als Elemente eines endlichen Vektorraums vorausgesetzt. Ohne Einschränkung der Allgemeinheit wird hier von einem zweidimensionalen Vektorraum ausgegangen:

1. **Dominanz:** Der Ähnlichkeitsabstand, der mehrere Dimensionen berücksichtigt, kann nicht kleiner als der maximale Ähnlichkeitsabstand aller einzelnen Dimensionen sein:

$$d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right) \geq \max\left\{d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_1 \\ y_2 \end{pmatrix}\right), d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_1 \end{pmatrix}\right)\right\}$$

2. **Konsistenz:** Die einzelnen Dimensionswerte wirken unabhängig voneinander:

$$\begin{aligned} d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_1 \end{pmatrix}\right) &> d\left(\begin{pmatrix} x_3 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_4 \\ y_1 \end{pmatrix}\right) \\ &\iff \\ d\left(\begin{pmatrix} x_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right) &> d\left(\begin{pmatrix} x_3 \\ y_2 \end{pmatrix}, \begin{pmatrix} x_4 \\ y_2 \end{pmatrix}\right) \end{aligned}$$

3. **Transitivität:** Mögliche Reihenfolgen von Objekten müssen pro Dimension transitiv wirken. Eine Reihenfolge von drei Objekten $x_1|x_2|x_3$ gilt, wenn x_2 zwischen x_1 und x_3 liegt:

$$d\left(\begin{pmatrix} x_1 \\ y \end{pmatrix}, \begin{pmatrix} x_3 \\ y \end{pmatrix}\right) > \max\left\{d\left(\begin{pmatrix} x_1 \\ y \end{pmatrix}, \begin{pmatrix} x_2 \\ y \end{pmatrix}\right), d\left(\begin{pmatrix} x_2 \\ y \end{pmatrix}, \begin{pmatrix} x_3 \\ y \end{pmatrix}\right)\right\}$$

Die Transitivität fordert, dass wenn $x_1|x_2|x_3$ und $x_2|x_3|x_4$ gelten, dass dann auch $x_1|x_2|x_4$ und $x_1|x_3|x_4$ erfüllt sein müssen.

Man kann nachweisen, dass diese Eigenschaften allgemeiner als die Distanzeigenschaften sind. Jede Distanzfunktion erfüllt also diese Eigenschaften, aber nicht umgekehrt. Die Symmetrieeigenschaft etwa wird nicht für einen Ähnlichkeitsabstand gefordert.

Die drei Eigenschaften besitzen eine wichtige Eigenart: Wendet man auf den Werten eines Abstandsmaßes, welches die drei geforderten Eigenschaften erfüllt, eine monoton steigende Funktion an, bleiben dadurch die Eigenschaften erhalten. Dies ist wichtig, da in vielen Ähnlichkeitsmodellen Funktionsanwendungen auf Ähnlichkeitswerten, etwa zur Skalierung, notwendig sind.

Erhaltung der Eigenschaften

6.3 Grenzen von Ähnlichkeitsmaßen

Maschinell berechenbaren Ähnlichkeitswerten sind im Vergleich zum menschlichen Ähnlichkeitsempfinden enge Grenzen gesetzt. Diese Grenzen sollen hier diskutiert werden.

Ähnlichkeitswerte in Multimedia-Systemen werden aus Feature-Werten berechnet. Im Gegensatz zu den Daten in einem klassischen Datenbanksystem werden Feature-Werte in der Regel automatisch aus Multimedia-Rohdaten extrahiert, also weder vom Menschen interpretiert noch entsprechend kodiert. Ein exaktes „Matching“ in Form von SQL-Bedingungen ist daher nicht möglich.

automatisch extrahierte Feature-Werte

kein exaktes „Matching“

Weiterhin sind semantische Eigenschaften in Feature-Daten in der Regel holistischer Natur. Dies bedeutet, dass die interessierenden Eigenschaften nicht exakt bestimmten Feature-Werten zugeordnet werden können, sondern durch die Gesamtheit der Daten ausgedrückt werden. Dies ist auch ein wichtiger Grund, warum eine Objekterkennung im allgemeinen Fall nicht funktioniert.

holistische Natur

Beispiel 6.3

Ein Multimedia-System soll auf beliebigen Rasterbildern Hunde erkennen. Dies ist jedoch nicht zuverlässig möglich, da die Eigenschaft „Hund“ weder an einzelne Pixel noch an bestimmte Feature-Werte exakt gebunden werden kann, sondern durch die Gesamtheit der Daten ausgedrückt wird.

Bei der menschlichen Ähnlichkeitsempfindung spielt bewusst oder unbewusst immer ein bestimmtes Weltwissen eine Rolle.

Weltwissen

Beispiel 6.4

Wird einem geschichteinteressierten Deutschen ein Photo von Erich Honecker und eine Abbildung von einem Trabi präsentiert, ergibt sich für ihn eine gewisse Ähnlichkeit aufgrund der DDR-Zugehörigkeit.

Weltwissen beim Ähnlichkeitsempfinden

Ohne Weltwissen ist ein solcher Zusammenhang nicht erkennbar. In Abhän-

Subjektivität

gigkeit vom Weltwissen entstehen unterschiedlich wahrgenommene Ähnlichkeiten. Ein großes Problem ist daher die Subjektivität menschlich wahrgenommener Ähnlichkeit, also die Abhängigkeit von Personen und deren Vorwissen, Interessen und Absichten.

Subjektivität von Ähnlichkeit

Beispiel 6.5

Ein weit verbreiteter „Volkssport“ ist das Vergleichen von Gesichtszügen eines Kindes mit denen der Eltern. Kommt das Kind mehr nach dem Vater oder nach der Mutter? Dabei stellt sich meistens heraus, dass die Meinungen weit auseinanderliegen. Oft ist das Ergebnis solcher Ähnlichkeitstests abhängig davon, wie gut derjenige die Elternteile kennt. Weiterhin werden oft unterschiedliche Merkmale für den Vergleich herangezogen und unterschiedlich gewichtet.

nicht modellierbares Weltwissen

Das von Menschen beim Ähnlichkeitsvergleich verwendete Weltwissen kann im Allgemeinen (noch) nicht im Computer abgelegt, geschweige denn geeignet verarbeitet werden.

drei Ebenen der Inhaltsverarbeitung

Bezüglich der Verwendung von Weltwissen unterscheidet man drei Ebenen, wie der Inhalt von Medienobjekten verwaltet und eine Ähnlichkeit berechnet werden könnte:

1. *syntaktische Ebene*: Hier erfolgt die Verarbeitung rein syntaktisch, ohne dass die Bedeutung der Medienobjekte berücksichtigt wird. Zum Beispiel können von Rasterbildern Farbverteilungen berechnet werden.
2. *semantische Ebene*: Auf dieser Ebene wird die Bedeutung von Medienobjekten verwaltet und für den Ähnlichkeitsvergleich verwendet. Zum Beispiel wird erkannt, dass auf einem Bild ein Baum abgebildet ist.
3. *pragmatische Ebene*: Auf der pragmatischen Ebene werden Medienobjekte interpretiert und thematischen Kategorien zugeordnet. Zum Beispiel kann ein Bild mit Bäumen zu einem Waldschadensbericht gehören.

Extraktion von Features der syntaktischen Ebene

Die meisten derzeit existierenden Feature-Extraktionsverfahren bewegen sich auf der syntaktischen Ebene. Ein hehres Ziel für die Zukunft ist die Realisierung des Übergangs von der syntaktischen zur semantischen Ebene. Für den allgemeinen Fall ist dies resultierend aus den eingangs erwähnten Problemen (noch) nicht erreichbar.

Da man in der Regel nicht in der Lage ist, Weltwissen geeignet im Computer abzubilden, stellt sich die Frage, welche Arten von Ähnlichkeit sich überhaupt berechnen lassen.

Um eine dem Menschen angepasste Ähnlichkeit nachzubilden, wurde von Psychologen untersucht, wie Menschen Reize wahrnehmen, bevor sie in der Lage sind, ihr Weltwissen zur Interpretation zu nutzen. Bei der menschlichen

Wahrnehmung von Reizen wird die so genannte pre-attentive von der attentiven Wahrnehmung unterschieden, wobei wir uns hier auf visuelle Reize beschränken. Die beiden Formen der Wahrnehmung unterscheiden sich in ihrer zeitlichen Dimension. Die pre-attentive Wahrnehmung erfolgt in den ersten 250 Millisekunden, in denen ein visueller Reiz auf das Auge einwirkt. Innerhalb dieser Zeitspanne ist der Mensch noch nicht in der Lage, sein Weltwissen zur Interpretation zu nutzen. Dies erfolgt in der darauf folgenden Phase, der attentiven Phase. Aufgrund des fehlenden Weltwissens ist für uns nur die pre-attentive Phase von Interesse.

pre-attentive versus attentive

Ein Ziel bei der Feature-Extraktion und beim Ähnlichkeitsvergleich ist es, die pre-attentive Ähnlichkeitswahrnehmung nachzubilden. Dazu wurden psychologische Experimente vorgenommen, die untersuchen, welche Feature in der pre-attentive-Phase wahrnehm- und unterscheidbar sind. Dies sind unter anderem Feature wie:

Nachbildung der pre-attentiven Wahrnehmung

- Linienorientierung,
- Länge, Breite, Größe von Objekten,
- Krümmung,
- Anzahl von Objekten und
- Farbe und Intensität von Objekten.

Leider sind Feature, die vom Menschen in der pre-attentiven Phase leicht wahrgenommen werden können, nicht immer leicht algorithmisch berechenbar. Diese Problematik ist aktueller Forschungsgegenstand der Bildverarbeitung und der Psychologie.

schwierige Berechnung pre-attentiver Feature-Werte

Die erwähnten Probleme beziehen sich auf den *allgemeinen* Fall. Wenn statt dessen von einer stark reglementierten Datenbasis ausgegangen wird, dann gibt es durchaus gute Lösungen, die dem menschlichen Ähnlichkeitsempfinden nahekommen. Das zur Interpretation notwendige Weltwissen ist stark eingeschränkt und kann daher durch Algorithmen nutzbar gemacht werden.

allgemeiner versus spezieller Fall

Beispiel 6.6

Ein Multimedia-System beinhaltet Passphotos. Ziel ist die Berechnung der Ähnlichkeit zwischen abgebildeten Personen. Für diesen speziellen Fall existieren relativ zuverlässige Verfahren der Ähnlichkeitsberechnung, die erfolgreich in der Kriminalistik eingesetzt werden.

spezielle Ähnlichkeit

Als Fazit gilt, je weniger die Medienobjekte reglementiert sind, und je ungenauer Ähnlichkeit festgelegt werden kann, desto schwieriger lässt sich eine computerberechnete Ähnlichkeit mittels eines Ähnlichkeitsmaßes auf der semantischen Ebene durchführen.

6.4 Konkrete Ähnlichkeitsmaße

*Vielfalt von
Funktionen und
Maßen*

In diesem Abschnitt werden konkrete Ähnlichkeitsmaße diskutiert. In der Wissenschaft werden viele verschiedene Funktionen und Maße vorgeschlagen, die durch ihre Kombination eine Vielfalt von Ähnlichkeitswertberechnungen erzeugen. Leider gibt es keine allgemein anerkannte Kombination von Funktionen und Maßen zur Berechnung von Ähnlichkeitswerten. Statt dessen stehen viele Alternativen zur Auswahl. Idealerweise sollten Funktionen und Maße gewählt werden, die möglichst gut dem subjektiven Ähnlichkeitsempfinden der potentiellen Anwender entsprechen, also im gemeinsamen Zusammenspiel gute Precision- und Recall-Werte erzielen.

*Beschreibung der
Eigenschaften*

Häufig können aus Kostengründen nicht alle möglichen Kombinationen durchgetestet und die beste Kombination ermittelt werden. Aus diesem Grund liegt der Schwerpunkt dieses Abschnittes nicht nur in der reinen Auflistung entsprechender Funktionen und Maße, sondern auch in der Beschreibung ihrer Eigenschaften. Diese sollen helfen, den Bezug zum beabsichtigten Ähnlichkeitsempfinden herzustellen.

In diesem Abschnitt werden folgende, konkrete Ähnlichkeitsmaße vorgestellt:

1. Feature-Kontrast-Modell nach Tversky,
2. Fuzzy-Feature-Kontrast-Modell von Santini und Jain,
3. Histogrammschnitt,
4. Kosinusmaß und
5. Ähnlichkeitsmaße aus der Taxonomie.

6.4.1 Feature-Kontrast-Modell nach Tversky

*binäre
Objekteigenschaften*

Tversky stellt in [208] ein spezielles Ähnlichkeitsmodell vor. Es geht von binären Eigenschaften aus. Jedem Medienobjekt kann, etwa im Kontext der Feature-Extraktion, eine Menge von Eigenschaften zugeordnet werden, welche das Objekt charakterisieren.

binäre Eigenschaften

Beispiel 6.7

Ein Bild b , auf dem ein roter Kreis und ein blaues Rechteck mittels einer Feature-Extraktion erkannt wurde, wird durch diese binären Eigenschaften $B = \{\text{roter-Kreis, blaues-Rechteck}\}$ charakterisiert.

Für ein Ähnlichkeitsmaß $s(a, b)$ zwischen zwei Objekten a und b auf der Grundlage der korrespondierenden Eigenschaftsmengen A und B stellt Tversky eine Reihe zu erfüllender Eigenschaften auf:

1. *Matching*: Das Ähnlichkeitsmaß ist eine Funktion über drei verschiedene, mengenwertige Komponenten:

$$s(o_1, o_2) = f(A \cap B, A \setminus B, B \setminus A)$$

2. *Monotonie*: Die Ähnlichkeit kann nur steigen, wenn die Schnittmenge nicht kleiner wird und die beiden Differenzen nicht größer werden:

$$s(a, b) \geq s(a, c) \text{ gdw.} \\ A \cap C \subseteq A \cap B, \quad A \setminus B \subseteq A \setminus C, \quad B \setminus A \subseteq C \setminus A$$

3. *Unabhängigkeit*: Bevor Unabhängigkeit gefordert wird, muss der Begriff der Übereinstimmung zweier Objektpaare definiert werden. $f(X, Y, Z)$ sei die Funktion für ein Ähnlichkeitsmaß mit $X = A \cap B$, $Y = A \setminus B$ und $Z = B \setminus A$. Wir schreiben weiterhin $V \approx W$, wenn X, Y und Z existieren, für die eine oder mehrere der folgenden Bedingungen gelten:

$$f(V, Y, Z) = f(W, Y, Z) \\ f(X, V, Z) = f(X, W, Z) \\ f(X, Y, V) = f(X, Y, W)$$

V und W sind dann sozusagen äquivalent. Zwei Objektpaare (a, b) und (c, d) stimmen in einer (zwei oder drei) Komponente(n) überein, wenn die entsprechenden Eigenschaften gelten:

$$(A \cap B) \approx (C \cap D) \\ (A \setminus B) \approx (C \setminus D) \\ (B \setminus A) \approx (D \setminus C)$$

Angenommen, die Paare (a, b) und (c, d) sowie die Paare (a', b') und (c', d') stimmen in denselben zwei Komponenten überein, während die Paare (a, b) und (a', b') sowie die Paare (c, d) und (c', d') in der übrigen Komponente übereinstimmen. Dann muss für die Eigenschaft der Unabhängigkeit folgende Bedingung gelten:

$$s(a, b) \geq s(a', b') \iff s(c, d) \geq s(c', d')$$

Abbildung 6.3 zeigt die Forderung nach Unabhängigkeit grafisch.

Auf der Grundlage dieser Eigenschaften formuliert Tversky den Repräsentationssatz des Feature-Kontrast-Modells.

$$\begin{array}{ccc}
 s(a, b) & \xrightarrow[2,3]{\approx} & s(c, d) \\
 \geq \Big| \begin{array}{c} 1 \\ \approx \end{array} & & \geq \Big| \begin{array}{c} 1 \\ \approx \end{array} \\
 s(a', b') & \xrightarrow[2,3]{\approx} & s(c', d')
 \end{array}$$

Abb. 6.3: Unabhängigkeit

Repräsentationssatz
des Feature-
Kontrast-Modells

Lemma 6.1

Angenommen, s sei ein Ähnlichkeitsmaß, für welches Matching, Monotonie und Unabhängigkeit erfüllt sind. Dann existiert eine Ähnlichkeitsfunktion S , eine nichtnegative Funktion f sowie zwei Konstanten $\alpha, \beta \geq 0$, so dass für alle Objekte a, b, c, d

$$S(a, b) \geq S(c, d) \iff s(a, b) \geq s(c, d)$$

und

$$S(a, b) = f(A \cap B) - \alpha f(A \setminus B) - \beta f(B \setminus A).$$

gelten. Dieser Satz besagt, dass jede Ähnlichkeitsordnung, welche Matching, Monotonie und Unabhängigkeit erfüllt, durch eine Linearkombination der Funktionswerte über der Menge der Gemeinsamkeiten ($A \cap B$) und den beiden Mengen der Unterschiede ($A \setminus B, B \setminus A$) nachgebildet werden kann.

Feature-Kontrast-
Modell und
Asymmetrie

Insbesondere lässt sich das Feature-Kontrast-Modell gut verwenden, um eine gewünschte Asymmetrie nachzubilden. Im vorigen Abschnitt wurde diskutiert, dass die Ähnlichkeit eines Objektes a mit relativ gering hervorstehenden Eigenschaften zu einem Objekt b mit relativ stark hervorstehenden Eigenschaften größer ist als umgekehrt. Im Feature-Kontrast-Modell geht man davon aus, dass die Stärke der Eigenschaften durch die Funktion f ausgedrückt werden kann¹:

$$f(B) > f(A)$$

Durch die Wahl der Konstanten $\alpha > \beta$ wird dann, wie gewünscht, eine Asymmetrie

$$S(a, b) > S(b, a)$$

erreicht.

Asymmetrie

Beispiel 6.8

In Abbildung 6.1 auf Seite 217 wurden zwei Bilder mit unterschiedlich hervorstehenden Eigenschaften vorgestellt. Die Bilder a, b können in diesem Beispiel folgendermaßen durch binäre Eigenschaften charakterisiert werden:

$$A = \{Haus\} \quad B = \{Haus, Baum\}.$$

¹Die Funktion f wird daher auch als Salient-Funktion bezeichnet.

Wenn die Funktion f die Kardinalität einer Menge berechnet und $\alpha = 2$ sowie $\beta = 1$ gilt, dann erhält man folgenden Ähnlichkeitswert für $S(a, b)$:

$$\begin{aligned} S(a, b) &= f(\{Haus\}) - 2f(\emptyset) - f(\{Baum\}) \\ &= 1 - 0 - 1 = 0 \end{aligned}$$

Dieser Wert unterscheidet sich vom Wert für $S(b, a)$:

$$\begin{aligned} S(b, a) &= f(\{Haus\}) - 2f(\{Baum\}) - f(\emptyset) \\ &= 1 - 2 - 0 = -1 \end{aligned}$$

Es gilt also:

$$S(a, b) \geq S(b, a)$$

Eine bis jetzt offen gelassene Frage betrifft die Eigenschaften des Feature-Kontrast-Modells: Erfüllt dieses Modell alle Eigenschaften eines Abstandsmaßes? Bei der Überprüfung der Eigenschaften Dominanz, Konsistenz und Transitivität müssen zwei Aspekte beachtet werden:

*Eigenschaften des
Feature-Kontrast-
Modells*

1. Das Feature-Kontrast-Modell liefert ein Ähnlichkeitsmaß, wohingegen die Eigenschaften Dominanz, Konsistenz und Transitivität einen Ähnlichkeitsabstand (also ein Unähnlichkeitsmaß) charakterisieren. Wir berücksichtigen dies durch eine entsprechende Umkehrung von Ungleichheitsbeziehungen.
2. Das Feature-Kontrast-Modell basiert auf binären Eigenschaften. Jede Dimension wird daher als eine binäre Eigenschaft aufgefasst.

*Ähnlichkeitsmaß
versus
Ähnlichkeitsabstand
binäre Eigenschaften*

Nachweis Dominanz: Die folgende Bedingung ist zu überprüfen:

$$d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right) \geq \max\left\{d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_1 \\ y_2 \end{pmatrix}\right), d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_1 \end{pmatrix}\right)\right\}$$

Gegeben seien die 4 Objekte a_1, b_1, b_2, b_3 mit den Eigenschaftsmengen

$$A_1 = \{x_1, y_1\} \quad B_1 = \{x_2, y_2\} \quad B_2 = \{x_1, y_2\} \quad B_3 = \{x_2, y_1\}.$$

Setzt man diese Mengen in die Bedingung

$$d(a_1, b_1) \geq \max\{d(a_1, b_2), d(a_1, b_3)\}$$

ein und kehrt das Ungleichheitszeichen um, dann erhält man:

$$\min\left\{\begin{array}{l} f(\emptyset) \quad - \quad \alpha f(\{x_1, y_1\}) \quad - \quad \beta f(\{x_2, y_2\}) \leq \\ f(\{x_1\}) \quad - \quad \alpha f(\{y_1\}) \quad - \quad \beta f(\{y_2\}), \\ f(\{y_1\}) \quad - \quad \alpha f(\{x_1\}) \quad - \quad \beta f(\{x_2\}) \end{array}\right\}$$

Aufgrund der Monotonieeigenschaft und den Untermengenbeziehungen zwischen den korrespondierenden Eigenschaftsmengen ist diese Ungleichung stets erfüllt.

Nachweis Konsistenz: Die folgende Konsistenzeigenschaft ist zu überprüfen:

$$\begin{aligned} d\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_1 \end{pmatrix}\right) &> d\left(\begin{pmatrix} x_3 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_4 \\ y_1 \end{pmatrix}\right) \\ &\iff \\ d\left(\begin{pmatrix} x_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right) &> d\left(\begin{pmatrix} x_3 \\ y_2 \end{pmatrix}, \begin{pmatrix} x_4 \\ y_2 \end{pmatrix}\right) \end{aligned}$$

Gegeben seien die 8 Objekte $a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4$ mit den Eigenschaftsmengen

$$\begin{aligned} A_1 &= \{x_1, y_1\} & A_2 &= \{x_3, y_1\} & A_3 &= \{x_1, y_2\} & A_4 &= \{x_3, y_2\} \\ B_1 &= \{x_2, y_1\} & B_2 &= \{x_4, y_1\} & B_3 &= \{x_2, y_2\} & B_4 &= \{x_4, y_2\} \end{aligned}$$

Setzt man diese Mengen in die Bedingung

$$\begin{aligned} d(a_1, b_1) &> d(a_2, b_2) \\ &\iff \\ d(a_3, b_3) &> d(a_4, b_4) \end{aligned}$$

ein und kehrt das Ungleichheitszeichen um, dann erhält man:

$$\begin{aligned} f(\{y_1\}) - \alpha f(\{x_1\}) - \beta f(\{x_2\}) &< \\ f(\{y_1\}) - \alpha f(\{x_3\}) - \beta f(\{x_4\}) & \\ &\iff \\ f(\{y_2\}) - \alpha f(\{x_1\}) - \beta f(\{x_2\}) &< \\ f(\{y_2\}) - \alpha f(\{x_3\}) - \beta f(\{x_4\}) & \end{aligned}$$

Wie man leicht überprüfen kann, entspricht diese Bedingung der Unabhängigkeit und ist damit immer erfüllt.

Nachweis Transitivität: Die Erfüllung der Transitivität

$$x_1|x_2|x_3 \text{ und } x_2|x_3|x_4 \implies x_1|x_2|x_4 \text{ und } x_1|x_3|x_4$$

Restriktion auf eine Dimension

ist etwas schwieriger nachzuweisen. Der entscheidende Punkt dieser Implikation liegt darin, dass sie sich nur auf eine Dimension bezieht und dass durch die Konsistenz die Dimensionen unabhängig voneinander wirken. Würde sich die Transitivität auf mehrere Dimensionen beziehen, könnte trotz Konsistenz leicht ein Gegenbeispiel gefunden werden.

Beispiel 6.9

Transitivität

Abbildung 6.4 zeigt ein positives und ein negatives Beispiel im zweidimensionalen Raum. In beiden Beispielen liegt der Punkt p_2 zwischen den Punkten p_1 und p_3 und der Punkt p_3 zwischen den Punkten p_2 und p_4 . Im linken Bild führt dies zu einem Abstand zwischen p_1 und p_4 , der bezüglich der anderen Abstände maximal ist. Damit wird die Transitivität im linken Bild erfüllt. Im rechten Bild wird sie jedoch verletzt, da der Abstand zwischen p_1 und p_4 zu klein ist.



Abb. 6.4: positives und negatives Beispiel der Transitivität

Da die Transitivität auf einer Dimension definiert ist, kann eine derartige Verletzung der Transitivität nicht auftreten. Es ist nicht möglich, eine Verletzung der Transitivität für Punkte auf einer Geraden zu konstruieren.

Damit erfüllt das Feature-Kontrast-Modell die Eigenschaften eines Ähnlichkeitsabstandes. Von den Eigenschaften einer Distanzfunktion wird die Symmetrie und die Dreiecksungleichung nicht erfüllt.

Leider weist das Feature-Kontrast-Modell einige Defizite auf, welche dessen Einsatz behindern:

Defizite des Feature-Kontrast-Modells

- **Abhängigkeit von Eigenschaftszahl:** Die Werte der Formel können abhängig von der Anzahl der beteiligten Eigenschaften sein. Dies demonstriert das folgende Beispiel.

Beispiel 6.10

Eigenschaftszahl

Angenommen, als Salient-Funktion f wird die Kardinalität der Eigenschaftsmenge verwendet. Seien weiterhin vier Objekte a, b, c, d mit ihren Eigenschaftsmengen $A = \{e_1, e_2\}$, $B = \{e_2, e_3\}$, $C = \{e_1, e_2, e_3, e_4\}$, $D = \{e_3, e_4, e_5, e_6\}$ gegeben. Sowohl das Paar (a, b) als auch das Paar (c, d) stimmen mit jeweils der Hälfte der Eigenschaften überein, während die andere Hälfte im jeweils anderen Objekt fehlt. Aus diesem Grund sollte man keinen großen Unterschied in den Ähnlichkeitswerten erwarten. Wenn $\alpha = \beta = 1$ gilt, dann beträgt jedoch der Ähnlichkeitswert des ersten Paares -1 und der des zweiten Paares -2 .

Beheben lässt sich dieses Problem bei der Verwendung der Kardinalität durch eine Normierung. Damit ergibt sich zur Berechnung der Ähnlichkeit zwischen zwei Objekten a und b folgende Formel:

$$S^{Norm}(a, b) = \frac{|A \cap B| - \alpha|A \setminus B| - \beta|B \setminus A|}{|A \cup B|}$$

- *Skalierung*: Ein Ähnlichkeitsmaß fordert eine Abbildung auf das Intervall $[0, 1]$. Der Maximalwert nach der normierten Formel beträgt 1 und der Minimalwert beträgt $-\max(\alpha, \beta)$. Eine Skalierung auf das gewünschte Intervall wird erreicht, wenn jeder Ähnlichkeitswert S^{Norm} zwischen zwei Objekten nach der normierten Formel folgendermaßen skaliert wird:

$$S^{[0,1]}(a, b) = \frac{S^{Norm}(a, b) + \max(\alpha, \beta)}{1 + \max(\alpha, \beta)}$$

- Das Feature-Kontrast-Modell basiert auf binären Eigenschaftswerten. In vielen Anwendungen liegen Eigenschaftswerte jedoch als reelle Werte vor. Diese können prinzipiell durch Festlegung bestimmter Intervallgrenzen in eine binäre Form überführt werden. Allerdings ist dies mit einem hohen Aufwand, willkürlichen Intervallgrenzen und einem Informationsverlust verbunden.

Ein Ansatz zur Behebung der Restriktion auf binäre Eigenschaftswerte wird von Santini und Jain vorgestellt.

6.4.2 Fuzzy-Feature-Kontrast-Modell von Santini und Jain

Fuzzy-Prädikat

In der Arbeit [170] von Santini und Jain wird mit Hilfe der Fuzzy-Logik die Einschränkung des Feature-Kontrast-Modells auf binäre Eigenschaften überwunden. Dabei wird jede der n Dimensionen als ein Fuzzy-Prädikat μ_i über den Objekten aufgefasst. Die Feature-Werte müssen Werte aus dem Intervall $[0, 1]$ sein.

Für die Berechnung des Ähnlichkeitsmaßes sind Mengenoperationen notwendig, die folgendermaßen realisiert werden:

- *Mengendurchschnitt*:

$$\mu_{\cap}(a, b) = \{\min(\mu_1(a), \mu_1(b)), \dots, \min(\mu_n(a), \mu_n(b))\}$$

- *Mengendifferenz*:

$$\mu_{\setminus}(a, b) = \{\max(\mu_1(a) - \mu_1(b), 0), \dots, \max(\mu_n(a) - \mu_n(b), 0)\}$$

- *Salient-Funktion*: Hier wird die Fuzzy-Kardinalität verwendet:

$$f(\{\mu_1, \dots, \mu_n\}) = \sum_{i=1}^n \mu_i.$$

Die Ähnlichkeit nach diesem Modell berechnet sich damit nach folgender Formel:

$$\begin{aligned} S(a, b) = & \sum_{i=1}^n \min(\mu_i(a), \mu_i(b)) \\ & - \alpha \sum_{i=1}^n \max(\mu_i(a) - \mu_i(b), 0) \\ & - \beta \sum_{i=1}^n \max(\mu_i(b) - \mu_i(a), 0). \end{aligned}$$

Die Berechnungen von

$$S^{Norm}(a, b) = \frac{S(a, b)}{\sum_{i=1}^n \max(\mu_i(a), \mu_i(b))}$$

und

$$S^{[0,1]}(a, b) = \frac{S^{Norm}(a, b) + \max(\alpha, \beta)}{1 + \max(\alpha, \beta)}$$

führen, analog zum ursprünglichen Feature-Kontrast-Modell, eine Normierung und Abbildung auf das Intervall $[0, 1]$ durch.

Normierung und Skalierung

6.4.3 Histogrammschnitt

Der Histogrammschnitt berechnet einen Ähnlichkeitswert zwischen zwei Histogrammen. Wir gehen zunächst von einem normalisierten Histogramm h_a mit jeweils n Werten aus, die relative Häufigkeiten ausdrücken:

Histogramm mit relativen Häufigkeiten

$$\sum_{i=1}^n h_a[i] = 1$$

Der Histogrammschnitt zwischen zwei Histogrammen h_a und h_b berechnet sich nach folgender Formel:

$$S_{nH}(h_a, h_b) = \sum_{i=1}^n \min(h_a[i], h_b[i]).$$

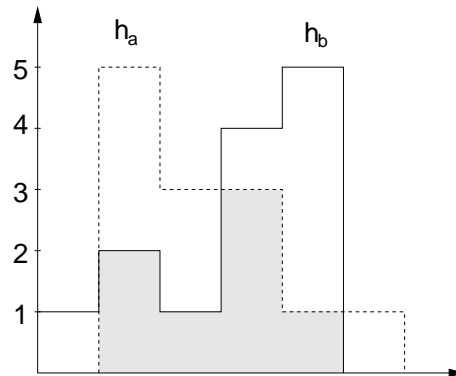


Abb. 6.5: Histogrammschnitt

Histogrammschnitt

Beispiel 6.11

In Abbildung 6.5 sind zwei Histogramme durch zwei Linien unterschiedlicher Linienarten dargestellt. Der Histogrammschnitt entspricht der markierten Fläche, welche den Flächenschnitt beider Histogramme beschreibt.

Ähnlichkeitsabstand

Aufgrund der Normierung der Histogramme ist garantiert, dass die Ähnlichkeitswerte auf das Intervall $[0, 1]$ abgebildet werden. Eine Umwandlung zu einem Ähnlichkeitsabstand kann mittels der Formel

$$d_{S_{nH}}(h_a, h_b) = 1 - S_{nH}(h_a, h_b)$$

Ähnlichkeitsabstand ist Distanzfunktion

erfolgen. Der Histogrammschnitt als Ähnlichkeitsabstand ist eine Distanzfunktion, da alle Distanzeigenschaften erfüllt sind. Auf einen Nachweis wird hier verzichtet.

Histogramm mit absoluten Häufigkeiten

Von den von Tversky geforderten Eigenschaften kann keine Eigenschaft überprüft werden. Dies ergibt sich aus der Forderung nach Normierung der Werte, die bewirkt, dass die Häufigkeiten nicht frei manipuliert werden können. Die Summe der Werte muss immer 1 betragen.

Verzichtet man bei den Histogrammen auf die Normierung, beschreiben Histogramme also absolute Häufigkeiten, erhält man ein Ähnlichkeitsmaß für beliebige Histogramme. Damit die Ähnlichkeitswerte im Intervall $[0, 1]$ liegen, ist eine Skalierung erforderlich:

$$S_H(h_a, h_b) = \frac{\sum_{i=1}^n \min(h_a[i], h_b[i])}{\sum_{i=1}^n h_a[i]}$$

Die Umwandlung zu einem Ähnlichkeitsabstand erfolgt analog zum normierten Fall. Von den Distanzeigenschaften ist nur die Selbstidentität erfüllt. Die

Positivität kann leicht verletzt werden, da das Minimum nur einen der zu vergleichenden Werte liefert.

*keine
Distanzfunktion*

Aus der Formel ist ersichtlich, dass die Symmetrie nicht erfüllt werden kann. Hier ergibt sich jedoch die Frage, wann $S(h_a, h_b) < S(h_b, h_a)$ gilt? Offensichtlich ist dies genau dann erfüllt, wenn $\sum_{i=1}^n h_a[i] > \sum_{i=1}^n h_b[i]$ gilt. Diese Summenberechnung kann also als eine Salient-Funktion² angesehen werden.

fehlende Symmetrie

Der nicht normierte Histogrammschnitt als Ähnlichkeitsabstand erfüllt die Eigenschaften Dominanz und Konsistenz. Die Transitivität ist nicht überprüfbar, da keine Histogramme h_1, h_2, h_3 in die Reihenfolge $h_1|h_2|h_3$ gebracht werden können.

*Dominanz und
Konsistenz*

Zwischen dem Histogrammschnitt und der L_1 -Distanzfunktion gibt es einen Zusammenhang. Dieser wird in Abbildung 6.5 ersichtlich. Der markierte Bereich entspricht dem Histogrammschnitt. Die L_1 -Distanzfunktion berechnet die Fläche zwischen den Histogrammkurven und der Schnittfläche. Da die Schnittfläche zu beiden Histogrammen gehört, ergibt die Summe aus dem doppelten Histogrammschnitt und der L_1 -Distanz die Gesamtsumme beider Histogramme:

*Histogrammschnitt
versus
 L_1 -Distanzfunktion*

$$d_{L_1}(h_a, h_b) + 2S_{nH}(h_a, h_b) = \sum_{i=1}^n h_a[i] + \sum_{i=1}^n h_b[i].$$

Falls die Histogramme normiert sind, kann der Histogrammschnitt direkt aus der L_1 -Distanzfunktion berechnet werden:

$$S_{nH}(h_a, h_b) = 1 - \frac{d_{L_1}(h_a, h_b)}{2}.$$

Der Ähnlichkeitsabstand bezüglich des Histogrammschnitts auf normierten Histogrammen ergibt dann:

$$d_{S_{nH}}(h_a, h_b) = 1 - S_{nH}(h_a, h_b) = \frac{d_{L_1}(h_a, h_b)}{2}.$$

Aus diesem Zusammenhang wird ersichtlich, dass dieser Ähnlichkeitsabstand alle Distanzeigenschaften erfüllt.

6.4.4 Kosinusmaß

Ein sehr weit verbreitetes Ähnlichkeitsmaß ist das Kosinusmaß. Dieses Maß ist auf Vektoren eines Vektorraums definiert und beschreibt den Kosinus des eingeschlossenen Winkels zwischen zwei Vektoren bezüglich des Nullvektors. Der Name dieses Ähnlichkeitsmaßes ergibt sich aus der Tatsache, dass über das Skalarprodukt der Kosinus des Winkels berechnet wird.

*eingeschlossener
Winkel*

²Zur Erläuterung einer Salient-Funktion siehe Seite 224.

Wenn a und b zwei Vektoren darstellen, dann berechnet sich das Kosinusmaß folgendermaßen:

$$S_{\cos}(a, b) = \frac{\langle a, b \rangle}{\|a\| * \|b\|}$$

$\langle a, b \rangle = a^T * b$ bezeichnet dabei das Skalarprodukt und $\|a\| = \sqrt{a^T * a}$ den Betrag. Das Kosinusmaß ermittelt zwischen zwei beliebigen Vektoren einen Wert aus dem Intervall $[-1, 1]$. Eine Abbildung auf das Ähnlichkeitsintervall $[0, 1]$ erhält man durch eine Halbierung gefolgt von der Addition mit $1/2$.

Skalierung

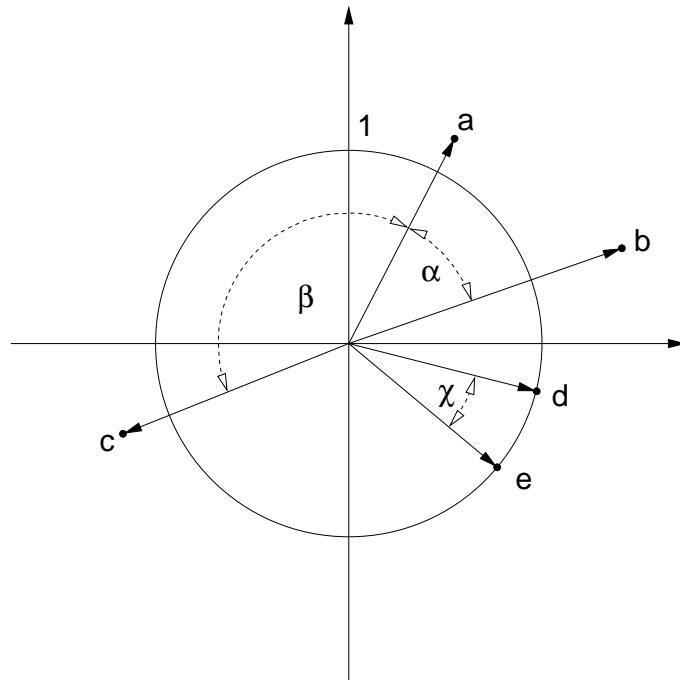


Abb. 6.6: Kosinusmaß

Kosinusmaß

Beispiel 6.12

In Abbildung 6.6 sind fünf Vektoren im zweidimensionalen Raum abgebildet. β beispielsweise bezeichnet den eingeschlossenen Winkel zwischen den Vektoren a und c . Die Kosinusfunktion berechnet dafür einen negativen Wert.

Bei der Berechnung des Kosinusmaßes können verschiedene Sonderfälle beachtet werden:

- *nichtnegative Werte*: Sind für alle Dimensionen der Vektoren die Werte nichtnegativ, kann der maximale Winkel nicht größer als 90 Grad sein. Damit erzeugt das Ähnlichkeitsmaß Werte aus dem Intervall $[0, 1]$, so dass eine anschließende Intervallanpassung nicht nötig ist.

In Abbildung 6.6 haben die Vektoren a und b nichtnegative Werte, liegen also im ersten Quadranten. Der Winkel ist kleiner als 90 Grad und ergibt damit einen positiven Wert aus dem Intervall $[0, 1]$.

- *längennormierte Vektoren*: Unter längennormierten Vektoren werden hier Vektoren der Länge 1 verstanden. Wie man sich leicht klar machen kann, ist der eingeschlossene Winkel unabhängig von einer Änderung der Länge eines Vektors.

In Abbildung 6.6 sind die Vektoren d und e normiert, liegen also auf dem abgebildeten Einheitskreis. Damit können sich normierte Vektoren nur in ihrer Richtung ändern, aber nicht mehr in ihrer Länge. Diese Invarianz geht also verloren.

Längeninvarianz

Um aus einem Ähnlichkeitswert aus dem Intervall $[0, 1]$ einen Ähnlichkeitsabstand zu erzeugen, wird dieser Wert von 1 abgezogen.

$$d_{cos}(a, b) = 1 - S_{cos}(a, b) = 1 - \frac{\langle a, b \rangle}{\|a\| * \|b\|}$$

Diese Abstandsfunktion ist ein Semi-Pseudo-Distanzfunktion beziehungsweise eine Semi-Distanzfunktion, wenn von längennormierten Vektoren ausgegangen wird:

Ähnlichkeitsabstand ist Semi-Pseudo-Distanzfunktion

- *Selbstidentität und Symmetrie*: Die Selbstidentität und Symmetrie kann direkt aus der Formel abgelesen werden.
- *Positivität*: Aufgrund der Längeninvarianz können unterschiedliche Vektoren einen Abstand von 0 ergeben, wenn sie sich nur in ihrer Länge unterscheiden. Dies ist nicht der Fall, wenn man von längennormierten Vektoren ausgeht. Damit ist die Erfüllung der Positivität abhängig von der Längennormierung.
- *Dreiecksungleichung*: Die Dreiecksungleichung kann nicht erfüllt werden. Beispiele, die dies demonstrieren, können leicht gefunden werden, werden hier aber nicht angegeben.

Als nächstes sollen die von Tversky geforderten Eigenschaften des Ähnlichkeitsabstandes überprüft werden:

Tversky-Eigenschaften

- *Dominanz*: Die Dominanz fordert, dass der Abstand zwischen zwei Punkten P_1 und P_2 kleiner wird, wenn für den Punkt P_2 der Wert für die Dimension x von dem Punkt P_1 übernommen wird. Für die Verletzung kann ein Gegenbeispiel angegeben werden. Die Dominanz gilt also nicht.

*Verletzung der
Dominanz*

Beispiel 6.13

In Abbildung 6.7 werden zwei Punkte P_1 und P_2 im zweidimensionalen Fall angezeigt. Wird der Punkt P_2 so modifiziert, dass er den x -Wert von Punkt P_1 übernimmt, wächst der Winkel anstatt, wie durch die Dominanz gefordert, zu schrumpfen. Der modifizierte Punkt ist der Punkt P'_2 .

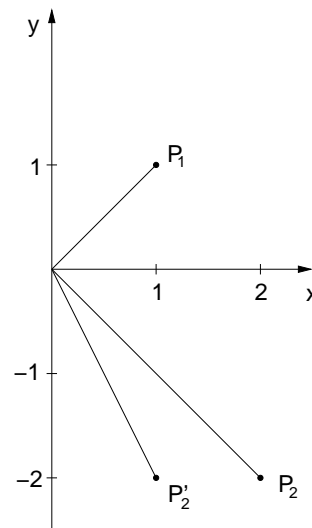


Abb. 6.7: keine Dominanz

- *Konsistenz:* Die Eigenschaft Konsistenz ist für den Ähnlichkeitsabstand des Kosinusmaß ebenfalls nicht erfüllt. Dies demonstriert die Abbildung 6.8. Während $\alpha < \beta$ für einen y -Wert gilt, kann ein anderer y -Wert, aber mit denselben x -Werten $\alpha' > \beta'$ erzeugen.
- *Transitivität:* Die Transitivität wird erfüllt. Die Reihenfolgen $x_1|x_2|x_3$ und $x_2|x_3|x_4$ können im zweidimensionalen Fall nur erfüllt werden, wenn die Vektoren, wie in Abbildung 6.8 dargestellt, auf einer Linie angeordnet liegen. Dann sind immer auch die Reihenfolgen $x_1|x_2|x_4$ und $x_1|x_3|x_4$ erfüllt.

*Variante des
Kosinusmaßes*

An dieser Stelle sei eine Variante des Kosinusmaßes aus der Psychologie, vorgeschlagen von Ekman in [52], erwähnt, ohne jedoch genauer auf deren Eigen-

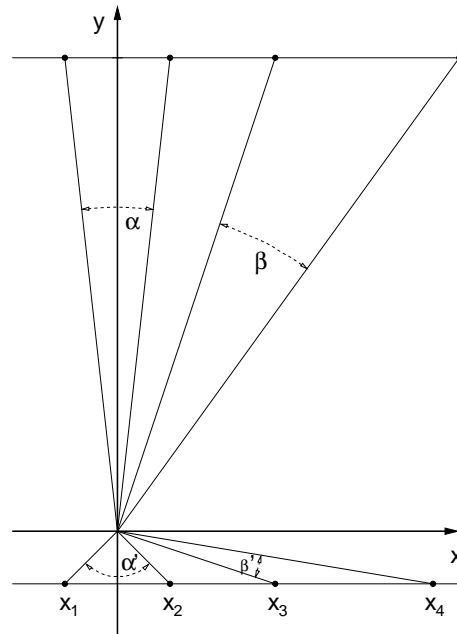


Abb. 6.8: keine Konsistenz

schaften einzugehen:

$$S_{Ekman}(x, y) = \frac{mx * \cos \theta + my * \cos \theta}{mx + my} \text{ mit}$$

$$mx = \begin{cases} \|x\| & \text{wenn } \|y\| \cos \theta \leq \|x\| \\ \|y\| * \cos \theta & \text{sonst} \end{cases}$$

$$my = \begin{cases} \|y\| & \text{wenn } \|x\| \cos \theta \leq \|y\| \\ \|x\| * \cos \theta & \text{sonst} \end{cases}$$

Dieses Maß berücksichtigt neben dem Kosinus des eingeschlossenen Winkels θ auch die Vektorlängen $\|x\|$ und $\|y\|$.

Zum Schluss soll gezeigt werden, dass es noch eine zweite Möglichkeit gibt, einen Ähnlichkeitsabstand auf der Grundlage des Kosinusmaß abzuleiten.

In Abbildung 6.9 entspricht der Punkt A dem längennormierten Vektor a und der Punkt B dem längennormierten Vektor b . Als Abstandswert zwischen a

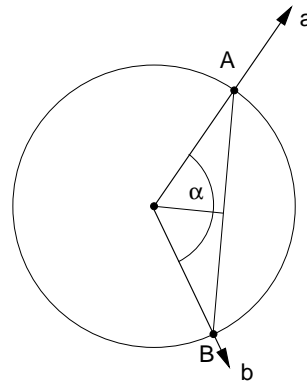


Abb. 6.9: Abstandsberechnung aus Kosinusmaß

und b kann die euklidische Distanz zwischen A und B verwendet werden:

$$\begin{aligned}
 d_{\cos 2}(a, b) &= \|A - B\| \\
 &= 2 * \sin \alpha / 2 \\
 &= \sqrt{2 * (1 - \cos \alpha)} \\
 &= \sqrt{2 * (1 - S_{\cos}(a, b))} \\
 &= \sqrt{2 * \left(1 - \frac{\langle a, b \rangle}{\|a\| * \|b\|}\right)}
 \end{aligned}$$

Aufgrund der Verwendung der euklidischen Gesetze erfüllt die Abstandsfunktion alle Eigenschaften einer Distanzfunktion und damit auch alle Eigenschaften eines Ähnlichkeitsabstandes, wenn von längennormierten Vektoren ausgegangen wird. Ansonsten kann die Positivität nicht garantiert werden.

6.4.5 Ähnlichkeitsmaße aus der Taxonomie

binäre Eigenschaften

Dieses Ähnlichkeitsmaß basiert auf verschiedenen Mengen von binären Eigenschaften. Ein Objekt x kann also durch eine Menge X erfüllter Eigenschaften aus der Gesamtmenge von Eigenschaften U charakterisiert werden. Die Ähnlichkeit berechnet sich nach folgender Formel:

$$S_{tax_1}(x, y) = \frac{|X \cap Y| + |(U \setminus X) \cap (U \setminus Y)|}{|U|}$$

Der Wert ergibt sich aus dem Verhältnis aus gemeinsam erfüllten und gemeinsam nicht erfüllten Eigenschaften zur Gesamtanzahl von Eigenschaften. Der korrespondierende Ähnlichkeitsabstand ergibt sich durch die Subtraktion von

1:

$$d_{tax_1}(x, y) = 1 - \frac{|X \cap Y| + |(U \setminus X) \cap (U \setminus Y)|}{|U|}$$

Der Ähnlichkeitsabstand erfüllt alle Eigenschaften einer Distanzfunktion. Die Selbstidentität, Positivität und Symmetrie kann direkt aus der Formel abgelesen werden. Für den Nachweis der Dreiecksungleichung geben wir nur die Idee an: Wie in Abbildung 6.10 angedeutet, kann die Dreiecksungleichung durch Vereinigungen der disjunkten Mengen 1, ..., 8 dargestellt werden. Aus dieser Darstellung kann leicht die Erfüllung dieser Ungleichung nachgewiesen werden.

*Ähnlichkeitsabstand
ist Distanzfunktion*

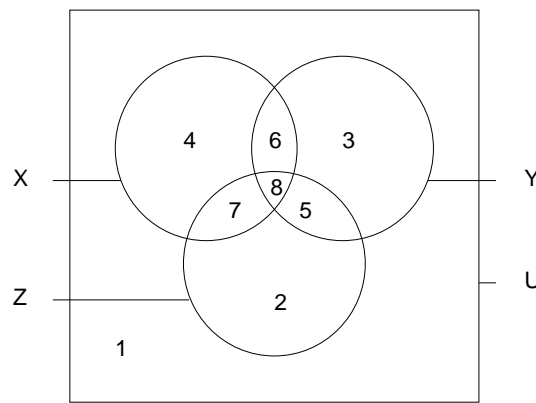


Abb. 6.10: Nachweis der Dreiecksungleichung

Ein anderes Ähnlichkeitsmaß aus der Taxonomie vernachlässigt die unerfüllten binären Eigenschaften:

$$s_{tax_2}(x, y) = \frac{|X \cap Y|}{|X \cup Y|}$$

*Berücksichtigung
nur erfüllter binärer
Eigenschaften*

Auch dieses Ähnlichkeitsmaß kann durch Subtraktion von 1 in eine Distanzfunktion umgewandelt werden. Diese Funktion erfüllt alle Eigenschaften einer Distanzfunktion und ist daher auch ein Ähnlichkeitsabstand. Die Eigenschaften der Selbstidentität, Positivität und Symmetrie können aus der Formel abgelesen werden. Für den Nachweis der Dreiecksungleichung eignet sich wieder die Mengenrepräsentation aus Abbildung 6.10.

Distanzfunktion

Eine weitere Variante eines Ähnlichkeitsmaßes, welche die Symmetrie aufgrund einer Gewichtung verletzt, sei hier erwähnt, ohne dass wir auf deren Eigenschaften genauer eingehen:

*gewichtetes
Ähnlichkeitsmaß*

$$s_{tax_3}(x, y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha * |X \setminus Y| + \beta * |Y \setminus X|}$$

Wenn $\alpha = \beta = 1$ gilt, dann entspricht dieses Maß dem Ähnlichkeitsmaß S_{tax_2} .

Neben diesen aufgeführten Ähnlichkeitsmaßen aus der Taxonomie gibt es eine große Vielzahl weiterer Ähnlichkeitsmaße, die etwa von Felix Brosius in [27] aufgelistet werden.

Tabelle 6.1 fasst die Eigenschaften der eingeführten Ähnlichkeitsabstände zusammen. Die Angabe „k.A.“ steht dabei für „keine Angabe möglich“.

Eigenschaft	$d_{S_{nH}}$	d_{S_H}	d_{cos}	d_{cos2}	d_{tax_1}	d_{tax_2}
Selbstidentität	✓	✓	✓	✓	✓	✓
Positivität	✓	–	–	✓	✓	✓
Symmetrie	✓	–	✓	✓	✓	✓
Dreiecksungleichung	✓	–	–	✓	✓	✓
Dominanz	k.A.	✓	–	✓	✓	✓
Konsistenz	k.A.	✓	–	✓	✓	✓
Transitivität	k.A.	k.A.	✓	✓	✓	✓

Tabelle 6.1: Eigenschaften der Ähnlichkeitsabstände

6.5 Aggregation von Ähnlichkeitswerten

In Abbildung 3.5 auf Seite 108 wurde die Konstruktion eines RSV-Wertes bezüglich einer komplexen Anfrage skizziert. Bei der Berechnung des Anfrageergebnisses mussten verschiedene Ähnlichkeitswerte zu einem endgültigen Ähnlichkeitswert aggregiert werden. In diesem Abschnitt sollen für die Aggregation von Ähnlichkeitswerten Verfahren vorgestellt werden. Das folgende Beispiel demonstriert die Notwendigkeit von Aggregationen in einem realen Szenario.

Suche nach Stoffen

Beispiel 6.14

In einer Bilddatenbank, die Abbildungen von Stoffen für die Produktion von Kleidungsstücken enthält, soll nach einem bestimmten Stoff gesucht werden. Vorgabe für die Suche sind dabei ein bestimmtes Muster und eine bestimmte Farbe. Diese beiden Eigenschaften führen pro Stoffabbildung der Datenbank zu zwei Ähnlichkeitswerten, die zu einem endgültigen Ähnlichkeitswert kombiniert werden müssen.

Anforderungen

An eine Aggregatfunktion *agg*, die Ähnlichkeitswerte für ein Objekt aggregiert, werden bestimmte Forderungen gestellt: