Dominanzproblem bei der Nutzung von Multi-Feature-Ansätzen

Thomas Böttcher
Technical University Cottbus-Senftenberg
Walther-Pauer-Str. 2, 03046 Cottbus
tboettcher@tu-cottbus.de

Ingo Schmitt
Technical University Cottbus-Senftenberg
Walther-Pauer-Str. 2, 03046 Cottbus
schmitt@tu-cottbus.de

ABSTRACT

Ein Vergleich von Objekten anhand unterschiedlicher Eigenschaften liefert auch unterschiedliche Ergebnisse. Zahlreiche Arbeiten haben gezeigt, dass die Verwendung von mehreren Eigenschaften signifikante Verbesserungen im Bereich des Retrievals erzielen kann. Ein großes Problem bei der Verwendung mehrerer Eigenschaften ist jedoch die Vergleichbarkeit der Einzeleigenschaften in Bezug auf die Aggregation. Häufig wird eine Eigenschaft von einer anderen dominiert. Viele Normalisierungsansätze versuchen dieses Problem zu lösen, nutzen aber nur eingeschränkte Informationen. In dieser Arbeit werden wir einen Ansatz vorstellen, der die Messung des Grades der Dominanz erlaubt und somit auch eine Evaluierung verschiedener Normalisierungsansätze.

Keywords

Dominanz, Score-Normalisierung, Aggregation, Feature

1. EINLEITUNG

Im Bereich des Information-Retrievals (IR), Multimedia-Retrievals (MMR), Data-Mining (DM) und vielen anderen Gebieten ist ein Vergleich von Objekten essentiell, z.B. zur Erkennung ähnlicher Objekte bzw. Duplikate oder zur Klassifizierung der untersuchten Objekte. Der Vergleich von Objekten einer Objektmenge O basiert dabei in der Regel auf deren Eigenschaftswerten. Im Bereich des MMR sind Eigenschaften (Features) wie Farben, Kanten oder Texturen häufig genutzte Merkmale. In vielen Fällen genügt es für einen erschöpfenden Vergleich von Objekten nicht, nur eine Eigenschaft zu verwenden. Abbildung 1 zeigt anhand des Beispiels eines Farbhistogramms die Schwächen einer einzelnen Eigenschaft. Obwohl beide Objekte sich deutlich unterscheiden so weisen sie ein sehr ähnliches Farbhistogramm auf.
Statt einer Eigenschaft sollte vielmehr eine geeignete Kombination verschiedener Merkmale genutzt werden, um mittels einer verbesserten Ausdruckskraft [16] genauere Ergebnissen zu erzielen. Der (paarweise) Vergleich von Objekten anhand

Figure 1: Unterschiedliche Objekte mit sehr hoher Farbähnlichkeit

von Eigenschaften erfolgt mittels eines Distanz- bzw. Ähnlichkeitsmaßes[1]. Bei der Verwendung mehrerer Eigenschaften lassen sich Distanzen mittels einer Aggregationsfunktion verknüpfen und zu einer Gesamtdistanz zusammenfassen. Der Einsatz von unterschiedlichen Distanzmaßen und Aggregationsfunktionen bringt jedoch verschiedene Probleme mit sich:
Verschiedene Distanzmaße erfüllen unterschiedliche algebraische Eigenschaften und nicht alle Distanzmaße sind für spezielle Probleme gleich geeignet. So erfordern Ansätze zu metrischen Indexverfahren oder Algorithmen im Data-Mining die Erfüllung der Dreiecksungleichung. Weitere Probleme können durch die Eigenschaften der Aggregationsfunktion auftreten. So kann diese z.B. die Monotonie oder andere algebraische Eigenschaften der Einzeldistanzmaße zerstören. Diese Probleme sollen jedoch nicht im Fokus dieser Arbeit stehen.
Für einen Ähnlichkeitsvergleich von Objekten anhand mehrerer Merkmale wird erwartet, dass die Einzelmerkmale gleichermaßen das Aggregationsergebnis beeinflussen. Häufig gibt es jedoch ein Ungleichgewicht, welches die Ergebnisse so stark beeinflusst, dass einzelne Merkmale keinen oder nur einen geringen Einfluss besitzen. Fehlen algebraische Eigenschaften oder gibt es eine zu starke Dominanz, so können die Merkmale und dazugehörigen Distanzmaße nicht mehr sinnvoll innerhalb einer geeigneten Merkmalskombination eingesetzt werden. Im Bereich der Bildanalyse werden zudem immer komplexere Eigenschaften aus den Bilddaten extrahiert. Damit wird auch die Berechnung der Distanzen basierend auf diesen Eigenschaften immer spezieller und es kann nicht sichergestellt werden welche algebraische Eigenschaften erfüllt werden. Durch die vermehrte Verwendung von vielen Einzelmerkmalen steigt auch das Risiko der Dominanz eines oder weniger Merkmale.
Kernfokus dieser Arbeit ist dabei die Analyse von Multi-Feature-Aggregationen in Bezug auf die Dominanz einzelner Merkmale. Wir werden zunächst die Dominanz einer Eigen-

[1]Beide lassen sich ineinander überführen [Sch06], im Folgenden gehen wir daher von Distanzmaßen aus.

In: G. Specht, H. Gamper, F. Klan (eds.): Proceedings of the 26th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 21.10.2014 - 24.10.2014, Bozen, Italy, published at http://ceur-ws.org.

schaft definieren und zeigen wann sich eine solche Dominanz manifestiert. Anschließend führen wir ein Maß zur Messung des Dominanzgrades ein. Wir werden darüber hinaus zeigen, dass die Ansätze bestehender Normalisierungsverfahren nicht immer ausreichen um das Problem der Dominanz zu lösen. Zusätzlich ermöglicht dieses Maß die Evaluation verschiedener Normalisierungsansätze.
Die Arbeit ist dabei wie folgt aufgebaut. In Kapitel 2 werden noch einmal einige Grundlagen zur Distanzfunktion und zur Aggregation dargelegt. Kapitel 3 beschäftigt sich mit der Definition der Dominanz und zeigt anhand eines Beispiels die Auswirkungen. Weiterhin wird ein neues Maß zur Messung des Dominanzgrades vorgestellt. Kapitel 4 liefert einen Überblick über bestehende Ansätze. Kapitel 5 gibt eine Zusammenfassung und einen Ausblick für zukünftige Arbeiten.

2. GRUNDLAGEN

Das folgende Kapitel definiert die grundlegenden Begriffe und die Notationen, die in dieser Arbeit verwendet werden. Distanzberechnungen auf unterschiedlichen Merkmalen erfordern in der Regel auch den Einsatz unterschiedlicher Distanzmaße. Diese sind in vielen Fällen speziell auf die Eigenschaft selbst optimiert bzw. angepasst. Für eine Distanzberechnung auf mehreren Merkmalen werden dementsprechend auch unterschiedliche Distanzmaße benötigt.
Ein Distanzmaß zwischen zwei Objekten basierend auf einer Eigenschaft p sei als eine Funktion $d : O \times O \mapsto \mathbb{R}_{\geq 0}$ definiert. Ein Distanzwert basierend auf einem Objektvergleich zwischen o_r und o_s über einer einzelnen Eigenschaft p_j wird mit $d^j(o_r, o_s) \in \mathbb{R}_{\geq 0}$ beschrieben. Unterschiedliche Distanzmaße besitzen damit auch unterschiedliche Eigenschaften. Zur Klassifikation der unterschiedlichen Distanzmaße werden folgende vier Eigenschaften genutzt:
Selbstidentität: $\forall o \in O : d(o, o) = 0$, Positivität: $\forall o_r \neq o_s \in O : d(o_r, o_s) > 0$, Symmetrie: $\forall o_r, o_s \in O : d(o_r, o_s) = d(o_s, o_r)$ und Dreiecksungleichung: $\forall o_r, o_s, o_t \in O : d(o_r, o_t) \leq d(o_r, o_s) + d(o_s, o_t)$.
Erfüllt eine Distanzfunktion alle vier Eigenschaften so wird sie als Metrik bezeichnet [11].
Ist der Vergleich zweier Objekte anhand einer einzelnen Eigenschaft nicht mehr ausreichend, um die gewünschte (Un-) Ähnlichkeit für zwei Objekte o_r,$o_s \in O$ zu bestimmen , so ist die Verwendung mehrerer Eigenschaften nötig. Für eine Distanzberechnung mit m Eigenschaften $p = (p_1 \dots p_m)$ werden zunächst die partiellen Distanzen $\delta^j_{rs} = d^j(o_r, o_s)$ bestimmt. Anschließend werden die partiellen Distanzwerte δ^j_{rs} mittels einer Aggregationsfunktion $agg : \mathbb{R}^m_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ zu einer Gesamtdistanz aggregiert. Die Menge aller aggregierten Distanzen (Dreiecksmatrix) für Objektpaar aus O, sei durch $\delta^j = (\delta^j_1, \delta^j_2 \dots, \delta^j_l)$ mit $l = \frac{n^2-n}{2}$ bestimmt. Dieser Ansatz erlaubt eine Bestimmung der Aggregation auf den jeweiligen Einzeldistanzwerten. Die Einzeldistanzfunktionen d^j sind in sich geschlossen und damit optimiert auf die Eigenschaft selbst.

3. DOMINANZPROBLEM

Bisher haben wir das Problem der Dominanz nur kurz eingeführt. Eine detaillierte Motivation und Heranführung an das Problem soll in diesem Kapitel erfolgen. Hierzu werden wir zunächst die Begriffe *Überbewertung* und *Dominanzproblem* einführen. Die Auswirkungen des Dominanzproblem auf das Aggregationsergebnis sollen anschließend durch ein Beispiel erläutert werden. Abschließend werden wir ein Maß definieren, um den Grad der Dominanz messen zu können.

3.1 Problemdefinition

Wie bereits erwähnt ist der Einsatz vieler, unterschiedlicher Eigenschaften (Features) und ihrer teilweise speziellen Distanzmaße nicht trivial und bringt einige Herausforderungen mit sich. Das Problem der Dominanz soll in diesem Unterabschnitt noch einmal genauer definiert werden.

Zunächst definieren wir das Kernproblem bei der Aggregation mehrerer Distanzwerte.

PROBLEM: *Für einen Ähnlichkeitsvergleich von Objekten anhand mehrerer Merkmale sollen die Einzelmerkmale gleichermaßen das Aggregationsergebnis beeinflussen. Dominieren die partiellen Distanzen δ^j_{rs} eines Distanzmaßes d^j das Aggregationsergebnis, so soll diese Dominanz reduziert bzw. beseitigt werden.*
Offen ist an dieser Stelle die Frage, wann eine Dominanz einer Eigenschaft auftritt, wie sich diese auf das Aggregationsergebnis auswirkt und wie der Grad der Dominanz gemessen werden kann.
Das Ergebnis einer Aggregation von Einzeldistanzwerten ist erneut ein Distanzwert. Dieser soll jedoch von allen Einzeldistanzwerten gleichermaßen abhängen. Ist der Wertebereich, der zur Aggregation verwendeten Distanzfunktionen nicht identisch, so kann eine Verfälschung des Aggregationsergebnisses auftreten. Als einfaches Beispiel seien hier zwei Distanzfunktionen d_1 und d_2 genannt, wobei d_1 alle Distanzen auf das Intervall $[0, 1]$ und d_2 alle Distanzen auf $[0, 128]$ abbildet. Betrachtet man nun eine Aggregationsfunktion d_{agg}, die Einzeldistanzen aufsummiert, so zeigt sich, dass d_2 das Aggregationsergebnis erheblich mehr beeinflusst als d_1.
Allgemein werden dann die aggregierten Distanzwerte stärker oder schwächer durch Einzeldistanzwerte einer (zur Aggregation verwendeten) Distanzfunktion beeinflusst als gewünscht. Wir bezeichnen diesen Effekt als eine Überwertung. Der Grad der Überbewertung lässt sich mittels Korrelationsanalyse (z.B. nach Pearson [10] oder Spearman [13]) bestimmen.

DEFINITION 1 (*Überbewertung einer Distanzfunktion*). *Für zwei Distanzfunktionen d^j und d^k, bei der die Distanzwerte δ^j in Abhängigkeit einer Aggregationsfunktion agg das Aggregationsergebnis stärker beeinflussen als δ^k, also die Differenz der Korrelationswerte*
$\rho(\delta^j, \delta^{agg}) - \rho(\delta^k, \delta^{agg}) > \epsilon$ ist, bezeichnen wir d^j als überbewertet gegenüber d^k.

Eine empirische Untersuchung hat gezeigt, dass sich ab einem Wert $\epsilon \geq 0.2$ eine Beeinträchtigung des Aggregationsergebnisses zu Gunsten einer Distanzfunktion zeigt.
Ausgehend von einer Überbewertung definieren wir das Problem der Dominanz.

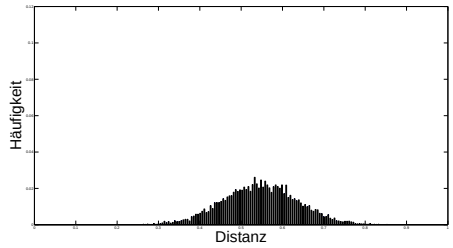
DEFINITION 2 (*Dominanzproblem*). *Ein Dominanzproblem liegt vor, wenn es eine Überbewertung einer Distanzfunktion d^j gegenüber d^k gibt.*

Das Problem einer Überbewertung bei unterschiedlichen Wertebereichen in denen die Distanzen abgebildet werden ist jedoch bereits weitreichend bekannt. In vielen Fällen kommen Normalisierungsverfahren (z.B. im Data-Mining [12] oder in der Biometrie [5]) zum Einsatz. Diese bereiten Distanzen aus verschiedenen Quellen für eine Aggregation vor.

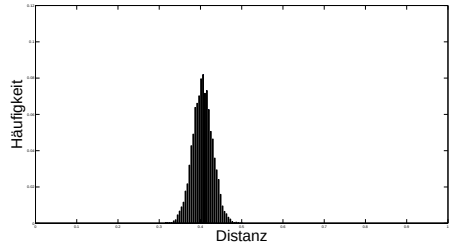
Zur Vermeidung einer Überbewertung werden Distanzen häufig auf ein festes Intervall normalisiert (i.d.R. auf [0,1]). Damit ist zumindest das Problem in unserem vorherigen Beispiel gelöst.
Das Problem der Dominanz tritt jedoch nicht nur bei unterschiedlichen Wertebereichen auf. Auch bei Distanzfunktionen, die alle auf den gleichen Wertebereich normalisiert sind, kann das Dominanzproblem auftreten. Im folgenden Abschnitt soll anhand eines Beispiels dieses Dominanzproblem demonstriert werden.

3.2 Beispiel eines Dominanzproblems

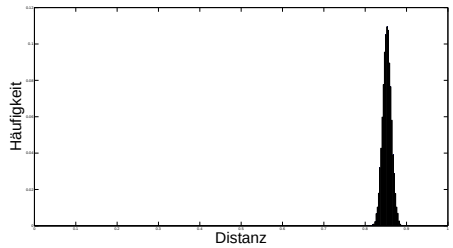
In Abbildung 2 sind drei Distanzverteilungen ν_1, ν_2 und ν_3 aus einer Stichprobe zu den zugehörigen Distanzfunktionen d_1, d_2 sowie d_3 dargestellt. Der Wertebereich der Funktionen sei auf das Intervall [0,1] definiert. Die Werte aus der Stichprobe treten ungeachtet der Normalisierung auf $[0, 1]$ jedoch in unterschiedlichen Intervallen auf. Die Distanzwerte der Stichprobe von ν_1 liegen im Intervall [0.2, 0.9], von ν_2 im Intervall [0.3, 0.5] und in ν_3 im Intervall [0.8, 0.9]. Auch wenn es sich hierbei um simulierte Daten handelt so sind solche Verteilungen im Bereich des MMR häufig anzutreffen.



(a) ν_1



(b) ν_2



(c) ν_3

Figure 2: Distanzverteilung verschiedener Distanzfunktionen (simulierte Daten)

Wir betrachten nun die Distanzfunktionen d_1 und d_2. Bezüglich einer beispielhaften Aggregationsfunktion[2] $agg_{\prod_{d_1,d_2}}(o_r, o_s) = d_1(o_r, o_s) * d_2(o_r, o_s)$ kann nun gezeigt werden, dass d_1 stärker den aggregierten Distanzwert beeinflusst als d_2.
In Abbildung 3 sind zwei verschiedene Rangfolgen aller 10 Distanzwerte zwischen fünf zufälligen Objekten der Verteilungen ν_1 und ν_2 dargestellt, sowie die Aggregation mittels $agg_{\prod}$. Die Distanz-ID definiert hierbei einen Identifikator für ein Objektpaar. Betrachtet man die ersten fünf Ränge der aggregierten Distanzen, so sieht man, dass die top-5-Objekte von Distanzfunktion d_1 komplett mit denen der Aggregation übereinstimmen, während bei Distanzfunktion d_2 lediglich zwei Werte in der Rangfolge der aggregierten Distanzen auftreten. Gleiches gilt für die Ränge 6–10. Damit zeigt die Distanzfunktion d_1 eine Dominanz gegenüber der Distanzfunktion d_2. Schaut man sich noch einmal die Intervalle der Verteilung ν_1 und ν_2 an, so zeigt sich, dass die Dominanz dem großen Unterschied der Verteilungsintervalle (0.7 vs. 0.2) obliegt. Eine Dominanz manifestiert sich also vor allem wenn eine große Differenz zwischen den jeweiligen Intervallen der Distanzverteilungen liegt.

3.3 Messung der Dominanz

Um die Überwertung aus unserem Beispiel und somit die Dominanz zu quantifizieren, wird die Korrelation zwischen den Distanzen von d_1 (d_2) und der aggregierten Distanzen aus d_{agg} bestimmt. Zur Berechnung der Korrelation können mehrere Verfahren genutzt werden. Verwendet man wie im obigen Beispiel nur die Ränge, so bietet sich Spearmans Rangkorrelationskoeffizient an [13].

$$\rho(A,B) = \frac{Cov(\mathrm{Rang}(A), \mathrm{Rang}(B))}{\sigma_{\mathrm{Rang}(A)} * \sigma_{\mathrm{Rang}(B)}} \text{ mit}$$
$$Cov(X,Y) = E\left[(X - \mu_x) * (Y - \mu_y)\right] \quad (1)$$

Hierbei sei $Cov(X, Y)$ die über den Erwartungswert von X und Y definierte Kovarianz. Bezogen auf das vorherige Beispiel erhalten wir eine Korrelation nach Spearman für d_1 von $\rho_1 = 0.94$ und für d_2 $\rho_2 = 0.45$. Die Differenz der Korrelationswerte liegt dabei bei $\rho_1 - \rho_2 = 0.49$. Ab $\epsilon = 0.2$ lässt sich eine Überbewertung einer Distanzfunktion feststellen. Somit haben wir mit $\rho_1 - \rho_2 = 0.49 > 0.2$ eine starke Überbewertung von d_1 gegenüber d_2 in Bezug auf das Aggregationsergebnis gezeigt.
Durch die Verwendung der Rangwerte gibt es allerdings einen Informationsverlust. Eine alternative Berechnung ohne Informationsverlust wäre durch Pearsons Korrelationskoeffizienten möglich [10]. Genügen die Ranginformationen, dann bietet Spearmans Rangkorrelationskoeffizient durch eine geringere Anfälligkeit gegenüber Ausreißern an [14].
Bisher haben wir die Korrelation zwischen den aggregierten Werten und denen aus je einer Distanzverteilung verglichen. Um direkt eine Beziehung zwischen zwei verschiedenen Distanzverteilungen bzgl. einer aggregierten Verteilung zu bestimmen, werden zunächst die zwei Korrelationswerte ρ_1 und ρ_2 der Distanzfunktionen d_1 und d_2 bzgl. ihres Einflusses auf das Aggregationsergebnis graphisch dargestellt [6]. Hierzu werden die jeweiligen Werte der Korrelation als Punkte in $[-1, 1]^2$ definiert. Für eine gleichmäßige Beeinflussung des Aggregationsergebnisses sollten sich die Punkte auf der Diagonalen durch den Koordinatenursprung mit

[2]Das Problem der Dominanz tritt auch bei anderen Aggregationsfunktionen wie Summe, Mittelwert etc. auf und kann zusätzlich eine Dominanz hervorrufen, z.B. bei der Minimum/Maximumfunktion.

Rang	d_1	Distanz-ID	d_2	Distanz-ID	$agg_{\prod}$	Distanz-ID
1	0.729	1	0.487	8	0.347	8
2	0.712	8	0.481	5	0.285	4
3	0.694	4	0.426	10	0.266	1
4	0.547	9	0.425	7	0.235	5
5	0.488	5	0.421	3	0.205	9
6	0.473	7	0.411	4	0.201	7
7	0.394	10	0.375	9	0.168	10
8	0.351	3	0.367	6	0.148	3
9	0.337	2	0.365	1	0.112	6
10	0.306	6	0.316	2	0.106	2

Figure 3: Dominanzproblem bei unterschiedlichen Verteilungen

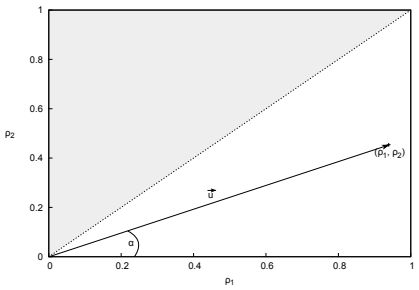


Figure 4: Graphische Darstellung der Korrelation ρ_1 und ρ_2 auf das Aggregationsergebnis

dem Anstieg $m = 1$ befinden. Wir bezeichnen diese Gerade als Kalibrierungslinie. Für unser Beispiel genügt es, nur positive Korrelationswerte zu betrachten. Damit kennzeichnen alle Punkte unterhalb dieser Linie einen größeren Einfluss durch d_1. Analog gilt bei allen Punkten oberhalb dieser Linie (grau schraffierter Bereich) eine größere Beeinflussung durch d_2. Abbildung 4 zeigt graphisch die Korrelation für unser Beispiel von ρ_1 und ρ_2 auf das Aggregationsergebnis. Um die Abweichung vom gewünschten Zustand zu bestimmen, ermitteln wir den Winkel zwischen dem Ortsvektor $\vec{u} = (\rho_1, \rho_2)^T$ durch den Punkt (ρ_1, ρ_2) und der horizontalen Koordinatenachse [6]. Der Winkel α ergibt sich dann durch $\alpha = \arctan\left(\frac{\rho_2}{\rho_1}\right)$ Dieser Winkel liegt zwischen $[0, \frac{\Pi}{2}]$, während die Kalibrierungslinie mit der horizontalen Achse einen Winkel von $\frac{\Pi}{4}$ einschließt. Für eine vorzeichenbehaftete Kennzeichnung der Überbewertung sollen nun alle Korrelationspunkte unterhalb der Kalibrierungslinie einen positiven Wert und alle Korrelationspunkte oberhalb einen negativen Wert erhalten. Für ein Maß der Dominanz definieren wir nun folgende Berechnung [6]:

$$Cal_{err}(\delta^i, \delta^j, \delta^{agg}) = 1 - \frac{4}{\pi}\arctan\left(\frac{Corr(\delta^j, \delta^{agg})}{Corr(\delta^i, \delta^{agg})}\right) \quad (2)$$

Hierbei definiert $Corr(X, Y)$ ein geeignetes Korrelationsmaß, in unserem Fall der Rangkorrelationskoeffizient von Spearman. Wir bezeichnen dieses Maß als Kalibrierungsfehler, wobei ein Fehler von 0 bedeutet, dass es keine Dominanz gibt und somit beide Distanzfunktionen gleichermaßen in das Aggregationsergebnis einfließen. Der Wertebereich des Kalibrierungsfehlers Cal_{err} liegt in $[-1, 1]$. Für unser Beispiel erhalten wir unter Verwendung von Spearmans Rangkorrelationskoeffizienten $Cal_{err}(d_1, d_2, d_{agg}) = 0.43$, womit erkennbar ist, dass d_1 das Aggregationsergebnis stärker beeinflusst als d_2.

Definition 3 (*Kalibrierungsfehler*). *Ein Kalibrierungsfehler liegt vor, wenn es eine Dominanz einer Distanzfunktion d_1 gegenüber d_2 gibt, d.h. die Korrelationswerte nicht auf der Kalibrierungslinie liegen. Entsprechend sind zwei Verteilungen von Distanzwerten kalibriert, wenn kein Kalibrierungsfehler auftritt.*

Analog zur Definition eines ϵ-Wertes zeigte eine empirische Untersuchung für einen Wert $\tau \geq 0.1$ eine ungleichmäßige Auswirkung auf das Aggregationsergebnis.

3.4 Zusammenfassung

Wir haben in diesem Kapitel gezeigt wann ein Dominanzproblem auftritt und wie groß der Einfluss auf das Aggregationsergebnis sein kann. Mit der Verwendung von Gleichung (2) ist es nun möglich den Grad des Dominanzproblems bzw. den Kalibrierungsfehler messen zu können. Ein Hauptgrund für das Auftreten des Dominanzproblem liegt in der Verteilung der Distanzen. Sind die Intervalle, in denen die Distanzen liegen unterschiedlich groß, so ist die Dominanz einer Eigenschaft unvermeidbar. Können diese Intervalle der Distanzverteilungen aneinander angeglichen werden ohne dabei die Rangfolge zu verletzen, so könnte dies das Dominanzproblem lösen. Weiterhin ermöglicht das Maß des Kalibrierungsfehlers die Evaluation von Normalisierungsansätzen.

4. STAND DER TECHNIK

Die Aggregation auf Basis mehrerer Eigenschaften ist ein weit verbreitetes Feld. Es gibt bereits eine Vielzahl von Arbeiten die sich mit dem Thema der *Score-Normalization* beschäftigten. Die Evaluierung solcher Ansätze erfolgt in vielen Fällen, vor allem im Bereich des IR, direkt über die Auswertung der Qualität der Suchergebnisse anhand verschiedener Dokumentenkollektionen, z.B. TREC-Kollektionen[3]. Dieses Vorgehen liefert aber kaum Anhaltspunkte, warum sich einige Normalisierungsansätze besser für bestimmte Anwendungen eignen als andere [6].
Betrachten wir zunächst verschiedene lineare Normalisierungen der Form normalize$(\delta) = y_{min} + \frac{\delta - x_{min}}{x_{max} - x_{min}}(y_{max} - y_{min})$ [15], wobei die Bezeichnungen x_{min}, x_{max}, y_{min} und y_{max} verschiedene Normalisierungsparameter darstellen. Tabelle 1 stellt einige solcher linearer Ansätze dar [15, 5, 9, 6].

Name	y_{min}	y_{max}	x_{min}	x_{max}
Min-Max	0	1	$min(\delta)$	$max(\delta)$
Fitting	$0 < a$	$a < b < 1$	$min(\delta)$	$max(\delta)$
ZUMV	0	1	μ_δ	σ_δ
ZUMV2	2	3	μ_δ	σ_δ
MAD	0	1	Median(δ)	$MAD(\delta)$

Table 1: Parameter lin. Normalisierungsverfahren

Neben den linearen Normalisierungsfunktionen gibt es auch zahlreiche nicht-lineare. Darunter fallen z.B. der tanh-Estimator [4] und die Double-Sigmoid [2] Normalisierung.

[3]Text Retrieval Conference (`http://trec.nist.gov/`)

Beide sind den linearen Verfahren jedoch sehr ähnlich. Avampatzis und Kamps stellen in [1] drei verschieden Normalisierungsverfahren vor, die alle auf der Annahme basieren, dass sich ein Score-Wert eine Summe aus einer Signal und einer Noise-Komponente zusammensetzen, wobei das Verhältnis der Summanden nur von dem Gesamt-Score abhängt [6, 1].
Für das Problem der Dominanz lässt sich einfach zeigen, dass diese Ansätze keinen direkten Einfluss auf die Distanzverteilung haben. Es werden maximal zwei statistische Merkmale (Minimum, Maximum, Median etc.) genutzt, um eine Normalisierung durchzuführen [9, 7]. Auch wenn diese Ansätze auf einigen Testkollektionen Verbesserungen in der Retrieval-Qualität erreichten, so kann nicht sichergestellt werden, dass diese Ansätze allgemein zu einer Verbesserung des Dominanzproblems beitragen. Besonders problematisch sind Ausreißer in den Verteilungen, die das Dominanzproblem bei einer Aggregation sogar noch verstärken können. Ebenfalls problematisch sind Aggregationen auf unterschiedlichen Distanzverteilungen, z.B. Normal- und Gleichverteilungen.
Es gibt allerdings auch Ansätze, die die Distanzverteilung als Grundlage zur Normalisierung heranziehen. Hierbei wird versucht die Distanzen aus unterschiedlichen Quellen so abzubilden, dass sie möglichst exakt gleiche Verteilungen besitzen. Die Ansätze von Manmatha [8] und Fernandez [3] analysieren dabei das probabilistische Verhalten von Suchmaschinen unter der Annahme, dass relevante Dokumente eine Normalverteilung und irrelevante eine exponentielle Verteilung besitzen. Diese Ansätze bieten zwar eine optimierte Normierung, erfordern aber gleichzeitig Zusatzinformation (z.B. über die Relevanz von Textdokumenten), die in vielen Anwendungsfällen gar nicht vorhanden sind.

5. ZUSAMMENFASSUNG UND AUSBLICK

In dieser Arbeit wurde ein Verfahren vorgestellt um die Dominanz unterschiedlicher Eigenschaften messen zu können. Hierzu wurde zunächst der Begriff der Dominanz und dessen Auswirkung untersucht. Anschließend wurde auf Basis eines von Distanzverteilungen ein Maß vorgestellt, mit dessen Hilfe der Grad der Dominanz bestimmt werden kann. Dies ermöglicht uns eine Überbewertung zu erkennen und die Qualität eines Normalisierungsverfahrens zu evaluieren. Die in dieser Arbeit vorgestellten Normalisierungsverfahren wiesen jedoch einige Schwächen auf. Hinzu kommt, dass die algebraischen Eigenschaften der zugrunde liegenden Distanzfunktionen gänzlich ungeachtet blieben (Problem fehlender Metrikeigenschaften). In zukünftigen Arbeiten soll daher ein Ansatz entwickelt werden, der beide Probleme gleichermaßen zu lösen versucht. Hierzu soll ein Verfahren der multivariaten Statistik, die multidimensionale Skalierung, verwendet werden. Zusätzlich sollen die Auswirkungen unterschiedlicher Normalisierungsansätze auf Dominanz und (Retrieval-) Qualität untersucht werden.

6. REFERENCES

[1] A. Arampatzis and J. Kamps. A signal-to-noise approach to score normalization. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 797–806, New York, NY, USA, 2009. ACM.

[2] R. Cappelli, D. Maio, and D. Maltoni. Combining fingerprint classifiers. In *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings*, pages 351–361, 2000.

[3] M. Fernández, D. Vallet, and P. Castells. Probabilistic score normalization for rank aggregation. In *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London*, volume 3936 of *Lecture Notes in Computer Science*, pages 553–556. Springer, 2006.

[4] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics - The Approach Based on Influence Functions*. Wiley, 1986. missing.

[5] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recogn.*, 38(12):2270–2285, Dec. 2005.

[6] R. Kuban. Analyse von Kalibrierungsansätzen für die CQQL-Auswertung. Bachelor's thesis, University of Cottbus-Senftenberg, Germany, Oct. 2012.

[7] J. H. Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, New York, NY, USA, 1997. ACM Press.

[8] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–275, New York, NY, USA, 2001. ACM Press.

[9] M. Montague and J. A. Aslam. Relevance score normalization for metasearch. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 427–433, New York, NY, USA, 2001. ACM.

[10] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42:59–66, 1988.

[11] H. Samet. *Foundations of Multidimensional And Metric Data Structures*. Morgan Kaufmann, 2006/08/08/ 2006.

[12] L. A. Shalabi and Z. Shaaban. Normalization as a preprocessing engine for data mining and the approach of preference matrix. In *DepCoS-RELCOMEX*, pages 207–214. IEEE Computer Society, 2006.

[13] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103, 1904.

[14] K. Varmuza and P. Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics*. Taylor & Francis, 2008.

[15] S. Wu, F. Crestani, and Y. Bi. Evaluating score normalization methods in data fusion. In *Proceedings of the Third Asia Conference on Information Retrieval Technology*, AIRS'06, pages 642–648, Berlin, Heidelberg, 2006. Springer-Verlag.

[16] D. Zellhöfer. Eliciting Inductive User Preferences for Multimedia Information Retrieval. In W.-T. Balke and C. Lofi, editors, *Proceedings of the 22nd Workshop "Grundlagen von Datenbanken 2010"*, volume 581, 2010.