Ein Hybridansatz zur Interaktion mit Retrievalsystemen

David Zellhöfer
Institut für Informatik
Brandenburgische Techn. Universität Cottbus
david.zellhoefer@tu-cottbus.de

Zusammenfassung

Gängige Retrievalsysteme nutzen qualitative oder quantitative Ansätze, um das subjektive Ähnlichkeitsempfinden der Nutzer bzw. ihre Präferenzen bezüglich einer Anfrage zu formulieren. Die beiden Ansätze treten dabei meist getrennt auf. Vertreter qualitativer Ansätze sehen quantitative Ansätze, die Ergebnismengen auf Basis numerischer Score-Werte und Gewichtungen erzeugen und damit eine Totalordnung generieren, meist als Spezialisierung ihrer Methoden.

Die vorgestellte Arbeit versucht, die Vorzüge beider Methoden zu verbinden, um dem Nutzer eine vereinfachte Interaktion während des Relevance Feedbacks mit dem Retrievalsystem zu bieten, ohne auf die höhere Ausdifferenzierung der Ergebnismenge quantitativer Ansätze verzichten zu müssen. Die Interaktion mit dem System geschieht dabei allein auf Basis der Angabe von intuitiv verständlichen Präferenzpaaren, die wiederum eine Halbordnung bilden. Hierbei muss angemerkt werden, dass die Angabe einer Präferenz $\geq$ immer auf einzelnen Paaren der Ergebnismenge geschieht und nicht mittels *abstrakter* Präferenzen, deren konkrete Angabe gerade im Feld des Multimedia Retrievals schwierig erscheint. Diese nutzerdefinierte Halbordnung stellt die Grundlage für ein maschinelles Lernverfahren auf Basis einer nicht-linearen Optimierung dar, welches aus den speziellen Präferenzen allgemeingültige Gewichtungen für die Anfrage generiert.

Erste Ergebnisse experimenteller Untersuchungen zeigen, dass der vorgestellte Ansatz gut einsetzbar ist und weiterverfolgt werden sollte.

1 Motivation

Retrievalsysteme sind gegenüber traditionellen Datenbanksystemen dadurch gekennzeichnet, dass sie eine Anfrage seitens des Nutzer mit relevanten Ergebnisobjekten beantworten. Diese Ergebnisobjekte müssen dabei eine Anfrage nicht komplett erfüllen, wie es im Fall eines DBS, das in der Regel auf der Booleschen Logik basiert, notwendig ist. Wichtiger ist bei diesen Systemen die Ähnlichkeit eines Objekts zur Anfrage. Das Konzept der Ähnlichkeit ist dabei subjektiv und basiert im wesentlichen auf den Erwartungen des Nutzers. Psychologische Studien stützen dies [2, 11]. Experimente im Information-Retrieval zeigen, dass einzelne Bedingungen einer Anfrage subjektiv wichtiger wahrgenommen werden und damit das Ähnlichkeitsempfinden steuern. Die Möglichkeit einzelne Bedingungen zu gewichten erhöht außerdem die Nutzerzufriedenheit [8].

Folglich nutzen gängige Retrievalsysteme Methoden, um das subjektive Ähnlichkeitsempfinden der Nutzer bzw. ihre Präferenzen bezüglich einer Anfrage zu formulieren. Präferenzen ermöglichen es dem Nutzer, einzelne Bedingungen der Anfrage stärker zu gewichten oder oder als besonders relevant anzugeben. Üblicherweise werden die Möglichkeiten zur Angabe von Präferenzen in zwei Klassen unterteilt: die quantitativen [10, 5] und die qualitativen Ansätze [4, 6]. Diese Klassen treten stets getrennt auf. Eine Kombination ist in der Regel nicht vorgesehen.

Die qualitativen Ansätze setzen voraus, dass der Nutzer bereits zum Zeitpunkt der Anfrageformulierung seine Präferenzen kennt. Präferenzen und Ergebnismengen werden hierbei mittels Halbordnungen abgebildet. Präferenzen erweitern die Boolesche Anfragelogik und lassen sich deshalb gut in relationalen DBS umsetzen. Beim Skyline-Operator [1] handelt es sich ebenfalls um einen qualitativen Ansatz.

In Abgrenzung dazu wird beim quantitativen Ansatz eine Totalordnung der Ergebnisobjekte auf der Grundlage einer numerischen Scoring-Funktion erzeugt, wie dies in Retrievalsystemen üblich ist. Präferenzen werden bei diesen Ansätzen in der Regel als numerischer Gewichtungswert über Anfragebedingungen ausgedrückt. Anhänger qualitativer Ansätze betrachten deshalb quantitative Verfahren als Spezialfall der qualitativen Verfahren, da jede Totalordnung auch eine Halbordnung ist. Dem kann entgegengebracht werden, dass ebenfalls jede Halbordnung mit der Dushnik-Miller-Dimension d durch den Schnitt von d Totalordnungen ausgedrückt werden kann.

Beide Verfahren haben ihre Daseinsberechtigung in unterschiedlichen Einsatzszenarien. Qualitative Verfahren bieten sich vor allem für den Einsatz in relationalen DBS an, da sie sich dort ohne größere Hindernisse implementieren lassen [4, 6]. Sie setzen allerdings voraus, dass die Präferenzen des Nutzers bereits zum Zeitpunkt der Anfrageformulierung bekannt sind.

Beispiel 1.1 *Denkbar ist dies z.B. bei der Wohnungssuche. Hier kann der Benutzer klar angeben, dass er Wohnungen mit Balkon gegenüber solchen ohne bevorzugt und und wenn eine Wohnung die gleiche Lage hat, die günstigste gewählt wird.*

Tabelle 1: Beispieldatenbank Wohnungen

ID	Lage	Balkon	Etage	Miete
A	Kreuzberg	ja	1	510
B	Mitte	ja	2	490
C	Kreuzberg	ja	5	450
D	Kreuzberg	ja	4	500
E	Neukölln	ja	4	350

Die entstehende Halbordnung würde die Wohnungen B,C und E als relevant einstufen. Eine Aussage über eine Ordnung innerhalb der Kreuzberger Wohnungen kann nicht getroffen werden, da diese nicht spezifiziert wurde[1]. Die Interpretation der Präferenz folgt somit dem Ceteris-paribus-Prinzip. D.h. dass die die Präferenz erfüllenden Objekte vor allen anderen stehen und die übrigen als gleichwertig betrachtet werden.

Diese Form des Ergebnisses lässt sich nicht mit quantitativen Ansätzen erreichen. Für diesen Fall müsste eine Scoring-Funktion gefunden werden, welche die Präferenz aus Beispiel 1.1 ausdrückt. Da keine Präferenz zwischen den Kreuzberger Wohnungen und z.B. Wohnung B angegeben wurde, müsste der Score von B gleich dem der Kreuzberger Wohnungen sein. Daraus folgt, dass der Score aller Kreuzberger Wohnung ebenfalls gleich sein müsste, was einen Widerspruch darstellt, da gleichzeitig die günstigste Wohnung gewählt werden sollte [3]. Trotz dieser Einschränkung bieten quantitative Ansätze ein feiner ausdifferenziertes Ergebnis, was durch folgendes Beispiel verdeutlicht werden soll.

Beispiel 1.2 *Der Nutzer bevorzugt Wohnungen mit Balkon, wobei ein möglichst günstiger Preis wichtig ist. Tabelle 2 zeigt ein denkbares Ergebnis als Totalordnung, wobei die Teilbedingungen unterschiedlich mittels einer Scoring-Funktion gewichtet wurden, so dass der Preis den größten Einfluss hat.*

Die Wohnungen A und D erhalten hier die niedrigsten Score-Werte, da sie am teuersten sind. Über die Score-Werte ist es möglich anzugeben, inwiefern sich die einzelnen Objekte bzgl. ihrer Relevanz zur Anfrage unterscheiden. Qualitative Ansätze ermöglichen keine Unterscheidung zwischen „wesentlich“ und „geringfügig besser“. Die beiden Wohnungen wären als gleichwertig betrachtet worden (s.o.).
Da konkrete Präferenzen bei quantitativen Verfahren noch nicht zu Beginn der Anfrage feststehen müssen, ist es möglich, dass Präferenzen während der Interaktion mit dem System angepasst werden können. Dies

[1] Dies gilt natürlich auch für eine Ordnung zwischen B,C und E.

Tabelle 2: Ranking der Wohnungen in einem quantitativen Verfahren

Score	ID	Lage	Balkon	Etage	Miete
0.8	E	Neukölln	ja	4	350
0.7	C	Kreuzberg	ja	5	450
0.6	B	Mitte	ja	2	490
0.5	D	Kreuzberg	ja	4	500
0.4	A	Kreuzberg	ja	1	510

wird mit einer höheren Komplexität der Präferenzangabe erkauft, da Präferenzen als Gewichtswerte für Anfrageteile numerisch angegeben werden müssen, was gerade bei komplexen Anfragen oder schwer verständlichen Anfragebedingungen eine Hürde für den Nutzer darstellt. Ein ähnliches Problem während der Nutzerinteraktion tritt auch bei qualitativen Verfahren auf. So ist es in einem Retrievalsystem kaum denkbar, dass Anwender konkret angeben können, welche Feature-Werte, wie z.B. Textureigenschaften, ihnen gegenüber anderen wichtiger sind.

Durch geeignete Verfahren des maschinellen Lernens und die Nutzung eines speziellen Relevance-Feedback-Prozesses kann die Angabe konkreter Gewichte jedoch verborgen werden, wie im folgenden Abschnitt gezeigt wird.

2 Ein Hybridansatz für das Relevance Feedback

Die vorgestellte Arbeit verbindet die Vorzüge beider Ansätze, um dem Nutzer eine vereinfachte Interaktion während des Relevance Feedbacks mit dem Retrievalsystem zu bieten, ohne auf die höhere Ausdifferenzierung der Ergebnismenge quantitativer Ansätze verzichten zu müssen. Die Interaktion mit dem System geschieht dabei allein auf Basis der Angabe von intuitiv verständlichen Präferenzpaaren, die wiederum eine Halbordnung bilden, da die Angabe von Gewichtswerten eine zu große Hürde für den Anwender darstellt.

Abbildung 1 zeigt den konzeptionellen Interaktionsablauf. Das Verfahren ist intern quantitativ und nutzt zur Anfrageformulierung und Gewichtung die Anfragesprache CQQL [10]. Die Nutzerschnittstelle hingegen verwendet eine qualitative Präferenz-Metapher. Hierbei muss angemerkt werden, dass die Angabe einer Präferenz $\geq$ immer auf einzelnen Paaren der Ergebnismenge geschieht und nicht mittels *abstrakter* Präferenzen, wie sie bei traditionellen qualitativen Verfahren üblich sind. Die Formulierung abstrakter Präferenzen ist gerade bei komplexen Retrievalsystem schwierig, da sie eine Kenntnis über zugrundeliegende Mechanismen und die Anfrage voraussetzt. Bei der Ergebnismenge handelt es sich um eine Totalordnung (Rank), die anhand der Scores der Objekte sortiert ist. Der erste Rank basiert dabei auf einer gewichteten CQQL-Anfrage und einem initialen Gewichtungsschema, z.B. zufälligen oder Nutzerprofil-basierten Gewichtswerten. Weitere mögliche Gewichtungsschemata finden sich in [13].

Sämtliche Präferenzpaare bilden die Menge P, welche mitsamt der zugrundeliegende Anfrage die Eingabe eines Downhill-Simplex-Lernalgorithmus [7] bilden. P ist dabei frei von widersprüchlichen Präferenzen, z.B. Zyklen in der abgeleiteten Halbordnung. Details zu den verwendeten Algorithmen finden sich in [9]. Der Lernalgorithmus dient zum Finden von konkreten Gewichtswerten, die zusammen mit der Anfrage einen neuen Rank erzeugen, welcher den angegebenen Präferenzen des Nutzers entspricht. Diese Modifikation hat zur Folge, dass sich die Ergebnisobjekte mit der Zeit an die Erwartung des Nutzers (Finaler Rank) anpassen, da die Gewichte direkt aus den nutzerdefinierten Präferenzen P' abgeleitet werden.

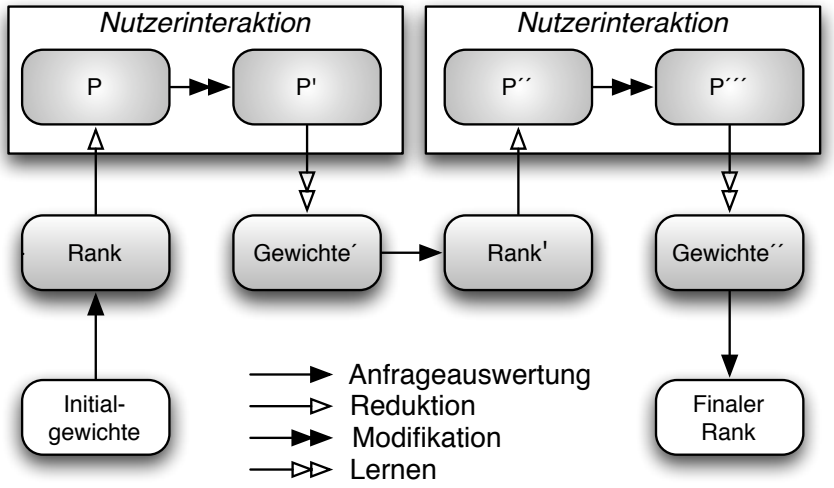

Abbildung 1: Verfeinerung von Gewichten durch Nutzerinteraktion mittels Präferenzen P

3 Fazit und Ausblick

Das vorgestellte Verfahren wurde experimentell anhand eines Szenarios zur Kamerasuche überprüft [9]. Dabei wurde gezeigt, dass das konzeptionelle Relevance-Feedback-Modell zu einer Annäherung an das vom Nutzer erwartete Ergebnis führt und ein Lernen von Gewichtswerten aus qualitativen Präferenzen möglich ist. Der zusätzliche Transformationsschritt und der Einsatz eines maschinellen Lernverfahrens zur Ermöglichung des Hybridverfahrens wirkt sich außerdem nicht kritisch auf die Laufzeit des Systems aus. Abb. 2 zeigt die Laufzeit der Lernalgorithmus in Abhängigkeit der Parameter Anzahl der Gewichtsvariablen und der nutzerdefinierten Präferenzen.

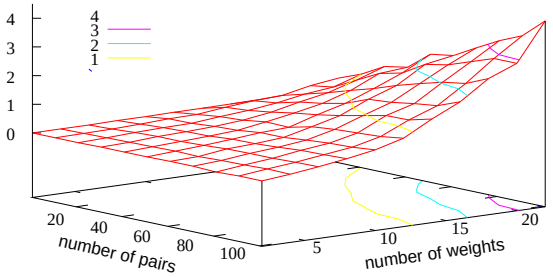

Abbildung 2: Laufzeit [s] in Abhängigkeit der Anzahl von Gewichten ($weights$) und Präferenzen ($pairs$)

Die Laufzeit für weniger als 10 Gewichtsvariablen innerhalb einer Anfrage und 40 Präferenzen liegt dabei deutlich unter einer Sekunde. Mit einer höheren Anzahl von Präferenzen ist nicht zu rechnen, da Nutzer in der Regel vor umfangreichen Interaktionen zurückschrecken [12]. Trotzdem erhöht eine große Anzahl von Gewichten und Präferenzen die Laufzeit nur langsam.

Abschließend kann festgestellt werden, dass sich qualitative und quantitative Verfahren kombinieren lassen. Im Bereich der Nutzerinteraktion sind qualitative Verfahren klar überlegen, da es Nutzern leichter fällt, Vergleiche mittels „besser als“ anzugeben als numerische Werte zu verwenden. Durch maschinelle Lernverfahren kann intern jedoch weiter ein quantitatives Verfahren genutzt werden, um eine möglichst große Ausdifferenzierung innerhalb der Ergebnismenge zu gewährleisten. Dies ist gerade im Retrievalbereich wünschenswert – bietet aber auch in anderen Szenarien bessere Möglichkeiten um beispielsweise die Ergebnismenge schrittweise zu vergrößern.

Literatur

[1] BÖRZSÖNYI, STEPHAN, DONALD KOSSMANN und KONRAD STOCKER: *The Skyline Operator*. In: *Proceedings of the 17th International Conference on Data Engineering*, Seiten 421–430, Washington, DC, USA, 2001. IEEE Computer Society.

[2] BRUCE, VICKI und PATRICK R. GREEN: *Visual Perception -physiology, psychology and ecology (2nd ed., reprinted)*. Lawrence Erlbaum Associates, Publishers, Hove and London, UK, 1993.

[3] CHOMICKI, JAN: *Querying with Intrinsic Preferences*. In: *EDBT '02: Proceedings of the 8th International Conference on Extending Database Technology*, Seiten 34–51, London, UK, 2002. Springer-Verlag.

[4] CHOMICKI, JAN: *Preference formulas in relational queries*. ACM Trans. Database Syst., 28(4):427–466, 2003.

[5] FAGIN, R. und E. L. WIMMERS: *A Formula for Incorporating Weights into Scoring Rules*. Theoretical Computer Science, 239(2):309–338, 2000.

[6] KIESSLING, W.: *Foundations of Preferences in Database Systems*. In: *Proc. of the 28th Int. Conf. on Very Large Data Bases, VLDB'02, Hong Kong, China, August, 2002*, Seiten 311–322. Morgan Kaufmann Publishers, 2002.

[7] NELDER, J. A. und R. MEAD: *A Simplex Method for Function Minimization*. Computer Journal, 7:308–313, 1965.

[8] SALTON, GERARD, EDWARD A. FOX und HARRY WU: *Extended Boolean Information Retrieval*. Commun. ACM, 26(11):1022–1036, 1983.

[9] SCHMITT, I. und D. ZELLHÖFER: *Lernen nutzerspezifischer Gewichte innerhalb einer logikbasierten Anfragesprache*. In: *Datenbanksysteme in Business, Technologie und Web, BTW'09, Lecture Notes in Informatics P-144*, Seiten 137–156. GI-Edition, 2009.

[10] SCHMITT, INGO: *QQL: A DB&IR Query Language*. The VLDB Journal, 17(1):39–56, 2008.

[11] SELFRIDGE, O. G.: *Pandemonium. A paradigm for learning*. The mechanics of thought processes, 1959.

[12] SHNEIDERMAN, BEN und CATHERINE PLAISANT: *Designing the user interface: Strategies for effective human–computer interaction*. Pearson, Boston, 4. ed. Auflage, 2005.

[13] ZELLHÖFER, D. und I. SCHMITT: *A Poset Based Approach for Condition Weighting*. In: *6th International Workshop on Adaptive Multimedia Retrieval, to appear*, 2009.