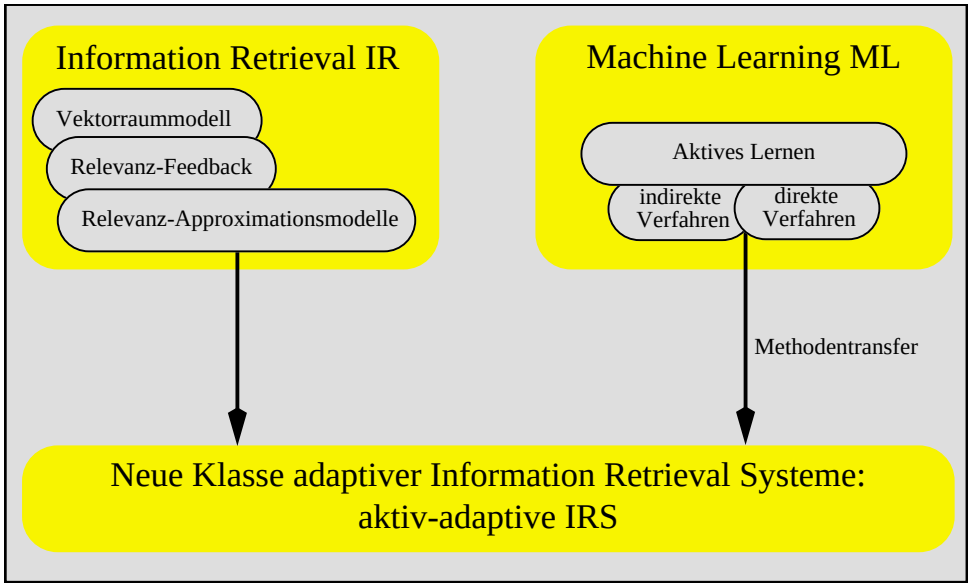
Polyrepräsentation, Relevanz-Approximation und aktives Lernen im Vektorraummodell des Information-Retrievals

von Günter Bachelier

Disputation 31.01.2002



1) Information Retrieval-Systeme (IRS)

Indexierung von Dokumenten

- Menge von m Dokumenten D_i; Menge von n Merkmalen F_j (z.B. Auftreten eines Termes oder eines Zeichen-n-Grams).
- Dokument-Indexierungsfunktion: Abbildung von D_i auf einen Punkt x_i in einem n-dimensionalen Dokumentvektorenraum (DVR).

Retrieval

- Anfrage (Query) Q_i durch User.
- Query-Indexierungsfunktion: Abbildung von Q_i auf einen Punkt q_i (Queryvektor) im DVR.
- Retrievalfunktion: Festlegung einer Dokumentvektorenteilmenge DVM_i; nachgewiesen wird Dokumentmenge.
- Rankingfunktion: Sortieren von Dokumentvektoren in DVM_i; nachgewiesen wird geordnete Dokumentliste.

Relevanz-Feedback (= interaktives Retrieval)

- Bewertung (binär, ..., reell) nachgewiesener Dokumente durch User.
- Dokumentbewertung wird für Dokumentvektoren übernommen => Relevanzwerte.

Approximation, Relevanz-Approximation

= Ersetzung Ursprungsfunktion f(x) durch Approximationsfunktion f(x)^, die entsprechend einem Qualitätsmaß nicht mehr als um einen Schwellenwert abweicht.

- Stützpunktbasiertes Approximationsmodell: Stützpunktmenge plus Approximationsverfahren AM(x) = (M, f(x | M)).
- Instanzbasiertes Approximationsmodell: Direkte Verwendung der Stimuli als Stützpunkte der Approximation.
- Prototypbasiertes Approximationsmodell: Erzeugt zunächst (kleinere) Menge von Prototypen als Stützpunkte.
- Anwendung im IR: Approximation von Relevanzwerten durch Feedback-Stützpunkte ergibt eine Form von User-Modell.

2) Aktives Lernen

= Lernsituationen, in denen der Lernende Einfluss auf die Komponenten des Lernprozesses besitzt.

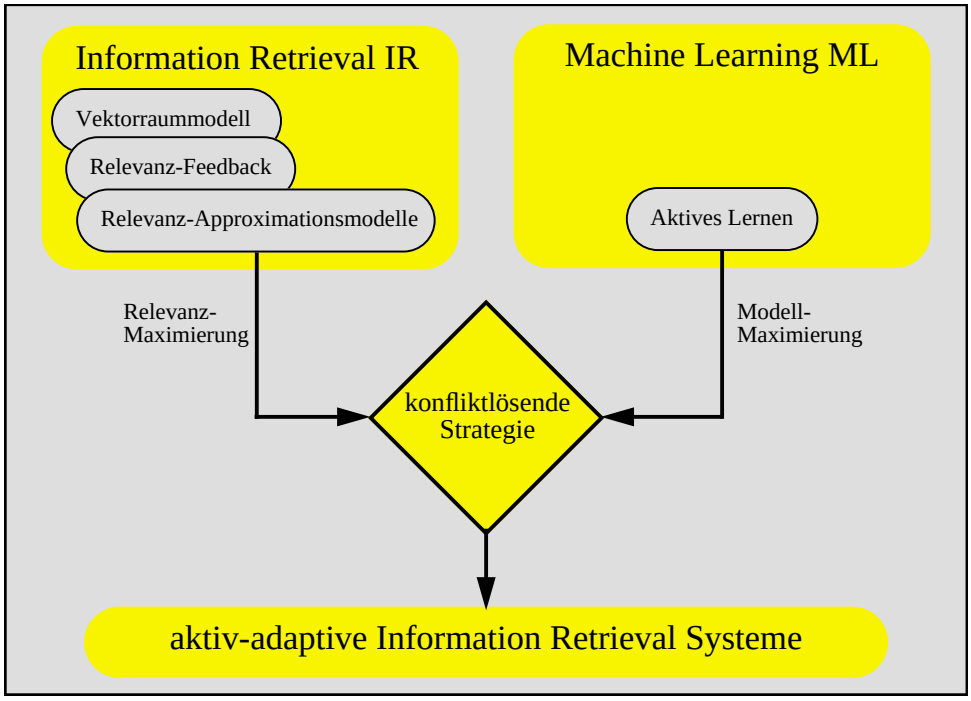
- Rationales aktives Lernen: Lernende verfolgt mit seiner Einflussnahme explizite Effektivitäts- und/oder Effizienz-Ziele.
- Aktives Lernen bei geschlossener Lernmenge: Gegeben ist nicht erweiterbare Menge unbewerteter Kandidatenvektoren.
- Indirekte Verfahren: Ermittlung von Eigenschaften eines Kandidatenstimulus (z.B. Relevanz-Varianz) und Rückschluss auf Eigenschaft des Modells, das sich ergibt, wenn der vervollständigte Stimulus in Lernmenge aufgenommen wird.
- Direkte Verfahren: Probeweise Erzeugung einer Menge alternativer Modelle durch Erweiterung der Lernmenge um eine Kandidatenstimuli-Teilmenge, Bewertung der Modelle (z.B. durch Bias-Quadrat-Integral, Varianz-Integral) und Auswahl des Modells mit den besten Eigenschaften (Modell-Selektion).

3) Aktiv-adaptive IRS

Aktiv-adaptive IRS erfordern Integration von Relevanz- und Modell-Maximierungskriterium:

- Relevanz-Maximierungskriterium: Auswahl der Dokumentvektoren, von denen das IRS annimmt, dass die zugehörigen Dokumente die höchsten Relevanzwerte besitzen => Verzerrung des Approximationsmodells; mangelnde Diskriminationsfähigkeit; systematisch zu gute Schätzungen.
- Modell-Maximierungskriterium: Auswahl der Dokumentvektoren, von denen das IRS annimmt, dass sie zu der größten Modellverbesserung führen werden, wenn sie als vervollständigte Stimuli m = (x, rel(x)) in die Lernmenge M aufgenommen werden => Konflikt mit beschränkter Bewertungsbereitschaft des Users, da gute wie schlechte Relevanzwerte benötigt werden.

=> Konflikt zwischen Relevanz- und Modell-Maximierungskriterium



Konfliktlösende Strategie

Interpretation der Konfliktsituation als Zwei-Ziel-Optimierungsproblem, das mit Hilfe der Pareto-Optimalität und darauf aufbauender Hierarchisierungsstrategien gelöst werden kann (z.B. Zwei-Ziel-Maximierung: Maximierung von Relevanzschätzung und Relevanz-Varianz bei indirektem Verfahren).