

Adaption von Information Retrieval-Verfahren zur automatisierten Produktsuche und -klassifikation

Stefan Naumann, Rolf Krieger, Norbert Kuhn,
Cordula Schürmann, Christian Sommer

Institut für Softwaresysteme in Wirtschaft, Umwelt und Verwaltung
Fachhochschule Trier, Standort Umwelt-Campus Birkenfeld
Postfach 1380, D-55761 Birkenfeld
naumann@umwelt-campus.de

Abstract: Die Suche nach Produkten über Merkmal/Wert-Paare in Datenbanken und im Internet ist ein wesentlicher Teil des elektronischen Beschaffungsprozesses. In diesem Beitrag wird ein Verfahren für die Suche in merkmalsbasierten Produktkatalogen vorgestellt. Das mehrstufige Verfahren adaptiert klassische Methoden aus dem Information Retrieval und der Fuzzy-Logik für diesen Suchvorgang. Ein Export der Suchergebnisse als XML-Produktkatalog unterstützt die Integration in bestehende Beschaffungsprozesse. Das Verfahren lässt sich auch zur automatischen Klassifikation vorhandener Produkte einsetzen.

1 Einführung und Motivation

Der Suchvorgang nach geeigneten Produkten auf Internet-Marktplätzen ist ein wesentlicher Teil des elektronischen Beschaffungsprozesses und findet im Regelfall merkmalsbasiert statt. Der Einkäufer legt in einer Anfrage die Eigenschaften eines Produkts fest, die durch Merkmale und ihre Werte vorgegeben werden. Als Ergebnis wird eine bewertete Liste relevanter Produkte erwartet. Besonders aufwändig ist dieser Suchvorgang im Fall komplexer Güter wie beispielsweise Industriemaschinen, die durch eine Vielzahl von Merkmalen beschrieben sind und daher in dieser Arbeit für die empirische Auswertung verwendet werden. Die Qualität der Suchergebnisse hängt dabei von der Genauigkeit der Anfrage, der Struktur des Produktkataloges, der Qualität des Datenbestandes und den Methoden des gewählten Suchverfahrens ab.

Grundlage für die Beschreibung der Produkte in einem Produktkatalog bildet ein *Klassifikationssystem*, das aus hierarchisch angeordneten *Produktgruppen* besteht, denen *Merkmale* und mögliche *Merkmalswerte* zugeordnet sind [LSD02; Kr02]. In einem *Produktkatalog* werden innerhalb der Klassifikation Produkte eindeutig den Produktgruppen zugeordnet und über Merkmal/Wert-Paare beschrieben.

Die hier beschriebenen Such- und Klassifizierungsstrategien werden in [Na03] eingeführt. Sie nutzen ein Klassifikationssystem und insbesondere die Produktmerkmale mit ihren Werten als indizierende Kennzeichen eines Produktes. Es wird davon ausgegangen, dass sich die Suche auf Produkte in der Datenbank eines Marktplatzes beschränkt. Insofern handelt es sich bei dem in dieser Arbeit beschriebenen Verfahren zum Produkt-Retrieval um eine Realisierung eines Data Retrieval / Faktenretrieval-Verfahrens [Ri79; Fu00], wobei wir jedoch auch Verfahren des Volltext-Retrievals anwenden. Neben einer kurzen Darstellung des in [Na03] vorgestellten Verfahrens werden in dieser Arbeit die Parametrisierung des erweiterten Fuzzy-Retrievals sowie eine XML-basierte Schnittstelle für den Export der Suchergebnisse (beispielsweise in Standard-Katalogformate) diskutiert. Dieser Schritt ist wichtig zur Integration der Recherche in den gesamten Beschaffungsprozess. Ergänzend werden allgemeine Probleme der Qualität von Produktdaten auf Internetmarktplätzen erläutert.

2 Erweitertes Fuzzy-Retrieval

Unser Ansatz kann im Ablauf wie folgt beschrieben werden (vgl. auch [Na03]): Nach Erstellung einer Suchanfrage auf Basis von Produktmerkmalen werden für jedes Produkt und jedes Merkmal Ähnlichkeiten zwischen Anfragewert und Produktwert anhand hinterlegter Zugehörigkeitsfunktionen ermittelt. Diese Ähnlichkeiten werden unter Berücksichtigung anfragespezifischer Gewichte durch das erweiterte boolesche Modell zu einer Gesamtrelevanz für jedes Produkt sortiert zusammengefasst. Das Ergebnis wird dem Anfragenden schließlich in einem elektronischen Katalogformat zur Verfügung gestellt. Diese Suchstrategie lässt sich auch zur automatisierten Klassifikation nutzen: Statt Produkten werden dann die zugehörigen Produktgruppen zur Klassifikation angeboten.

2.1 Produkt-Retrieval-Modell

Ein *Produkt-Retrieval-Modell* besteht aus einem Klassifikationssystem, einem Produktkatalog, einer Menge von Anfragen nach Produkten (wobei eine Anfrage aus Produktgruppen und Merkmal/Wert-Paaren, die über logische Operatoren verknüpft sind, besteht) und einer Menge von Rankingfunktionen, die jedem Produkt p in Abhängigkeit von einer Anfrage q eine Relevanz zwischen 0 und 1 zuweisen.

2.2 Erweitertes Fuzzy-Retrieval

2.2.1 Ähnlichkeitsermittlung durch Zugehörigkeitsfunktionen

Im Produkt-Retrieval-Modell steht $\mathbf{z}: W_m \Delta W_m \downarrow [0,1]$ für die *Zugehörigkeitsfunktion*. Jedem Merkmal $\mathbf{m}$ ist (produktgruppenabhängig) mindestens eine solche Funktion zugeordnet, wobei mit W_m der Wertebereich des Merkmals bezeichnet wird. Die Funktion bestimmt die Ähnlichkeit zwischen Anfrage- und Produktwert. Um einen domänenabhängigen geeigneten Funktionsabfall zu erreichen wird die Grundfunktion in einen *ähnlichkeitsrelevanten* Bereich gestreckt oder gestaucht. Für eine konkrete Anfrage $\mathbf{q}$ wird dann der Anfragewert fixiert. Es ergibt sich eine eindimensionale Funktion, die als Variable Produkt-Merkmalwerte aufnimmt und deren Ähnlichkeit zur Anfrage bestimmt.

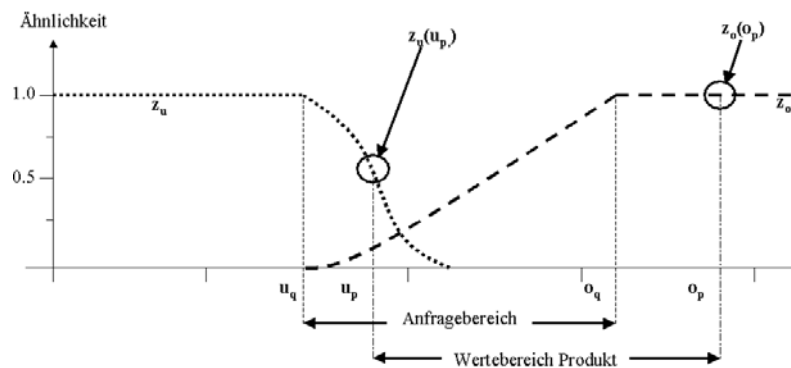


Abbildung 1: Zugehörigkeitsfunktionen eines Bereichsvergleiches

Wie in [Na03] beschrieben wird der Typ der Zugehörigkeitsfunktion neben der domänenabhängigen Semantik auch durch die Skalierung des Merkmals beeinflusst. So erfordern beispielsweise Bereiche (z.B. Drehzahlbereiche), mehrdimensionale Werte (z.B. Raummaße) oder Mehrfachwerte (z.B. gestufte Einstellungen) mehrere angepasste Zugehörigkeitsfunktionen. Abbildung 1 zeigt eine typische Zugehörigkeitsfunktion für einen Vergleich von Bereichen. Hier wird davon ausgegangen, dass der Anfragende ein Produkt mit Merkmalswerten sucht, die den Anfragebereich $[u_q, o_q]$ mindestens abdecken. Dem Merkmal wurden 2 Funktionen zugeordnet: z_u für die untere und z_o für die obere Schranke des Bereichs. Die Ähnlichkeit für einen Produkt-Wertebereich $[u_p, o_p]$ ergibt sich dann durch $z_u(u_p) \cdot z_o(o_p)$, wobei auch andere zusammenfassende Verfahren möglich sind.

2.2.2 Zusammenfassung der Einzelähnlichkeiten durch Rankingfunktionen

Um die merkmalsbezogenen Ähnlichkeiten zu einem Gesamtranking zusammenzufassen, adaptieren wir das erweiterte boolesche Modell, da es über Parametervariation des

Potenzfaktors p mehrere Rankingverfahren abbilden kann [SFW83]. Tabelle 1 zeigt Suchergebnisse nach einer Drehmaschine in unserer Testplattform, die bei einem Datenbestand von etwa 8.000 Industriemaschinen rund 500 Merkmale in über 500 Produktgruppen hat. Ausgewertet wurde jeweils mit dem Parameter $p=1$ und $p=3$. Als Zugehörigkeitsfunktion wurde generell die Trapezfunktion gewählt.

	Herstellungs-jahr	Drehdurch-messer über Bett	Drehdurch-messer	Drehzahl-bereich	Dreh-länge	Transport-kosten	Ranking	
Anfrage	1985..1985	800..800	400..400	100..1000	1000..1000	Nach Vereinb.	P = 1.0	P = 3.0
Gewichtsvektor	1.0	0.5	1.0	0.5	1.0	0.5		
Maschine 3580	1983	820	465	56..1120	1000	Nach Vereinb.		
Gewichtsvektor	0.9979849	0.95	0.675	0.0912	1.0	1.0	82.07	55.01
Maschine 3453	1983	820	465	56..1120	1000			
Gewichtsvektor	0.9979849	0.95	0.675	0.0912	1.0	0	70.96	41.31
Maschine 3457	1983	820	465	56..1120	1000			
Gewichtsvektor	0.9979849	0.95	0.675	0.0912	1.0	0	70.96	41.31
Maschine 3458	1983	820	465	56..1120	1000			
Gewichtsvektor	0.9979849	0.95	0.675	0.0912	1.0	0	70.96	41.31
Maschine 3681	1981		400		1000			
Gewichtsvektor	0.9959698	0	1.0	0	1.0	0	66.57	32.00

Tabelle 1: Ranking-Ergebnisse mit erweitertem Fuzzy-Retrieval ($p=1$ und $p=3$)

2.3 Bewertung der Ergebnisse

Betrachtet man die vorgeschlagene Suchstrategie, so lässt sich feststellen, dass genauere Suchergebnisse als bei einer rein booleschen Anfrage oder einer Suche entlang der Produktgruppenstruktur möglich sind. Das Verfahren lässt gewichtete Anfragen zu und ermittelt eine Gesamtbewertung aufgrund von Einzelgewichten der Merkmal/Wert-Paare. Die Vorteile des erweiterten Fuzzy-Retrievals sind die unscharfe Verarbeitung von Anfragen, die spezifische Anpassung von Zugehörigkeits- und Rankingfunktionen an die Anwendungsdomäne sowie die Verwendbarkeit beliebiger logischer Verknüpfungen. Im Vergleich der Rankings für $p=1$ und $p=3$ lässt sich erkennen, dass der kleinere Wert für p ein höheres Ranking pro Produkt ergibt, die Rangfolge der Ergebnisse aber unverändert ist. So hat beispielsweise die Maschine 3580 bei $p=1$ eine Relevanz von 82,07 % und bei $p=3$ eine Relevanz von 55,01 %, ist aber dennoch der beste Treffer.

2.4 Qualität der Produktdaten

Das erweiterte Fuzzy-Retrieval ist allerdings nur dann erfolgreich einzusetzen, wenn die Datenqualität eine funktionale Auswertung bzw. den Vergleich über Korrelationsmatrizen [OMT91] überhaupt zulässt. Beim Import verschiedener Produktkataloge in unsere Testplattform sind wiederholt Probleme im zu verarbeitenden Datenbestand aufgetreten, vor allem in den Katalogteilen, die durch den Nutzer (z.B. Produkteinpfege durch Anbieter in einem elektronischen Marktplatz) eingetragen werden können. So finden sich beispielsweise Angaben von Maßeinheiten direkt hinter Merkmalswerten (z.B. 2.000 EUR), obwohl hierfür eigene Felder vorgesehen sind. Ein angebotenes Merkmal "Sonstiges" wird häufig für eine unscharfe, merkmalsbasierte Produktbeschreibung genutzt,

dafür vorgesehene Merkmale werden nicht gepflegt. Wir entwickeln daher zur Zeit Verfahren und Tools zur Analyse und Bereinigung von Produkt-Datenbeständen, um einem differenzierten Suchverfahren wie dem erweiterten Fuzzy-Retrieval eine geeignete Grundlage zu bieten.

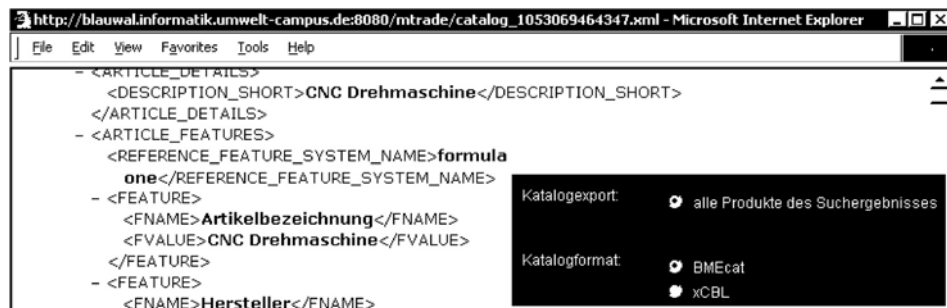


Abbildung 2: Schnittstelle zum Katalogexport mit Beispielkatalog

3 XML-Export der Suchergebnisse

Für die weitere Verwendung von Suchergebnissen oder die Integration der Recherche in den Beschaffungsprozess ist es notwendig, die gefundenen und gerankten Produkte in einem geeigneten Format zu exportieren. Im Bereich der elektronischen XML-basierten Katalogformate sind BMEcat [Bu02] und xCBL [Co02] bekannte Standardformate. Um ein Suchergebnis dem Anfragenden zur Weiterverarbeitung geeignet zur Verfügung stellen zu können, wurde für unsere Testplattform eine Schnittstelle geschaffen, die mittels eines konfigurierbaren XML-Mappingdokumentes aus einem Suchergebnis heraus eine XML-basierte Katalogstruktur erzeugt. Abbildung 2 zeigt die Verwendung der Schnittstelle und einen Auszug aus dem Katalogdokument. Um dem Anfragenden nicht nur die gefundenen Produkte zur Verfügung zu stellen, sondern auch das Ranking, ist es notwendig im Export ein weiteres Produktmerkmal zu ergänzen, welches die Relevanz des Produkts enthält. Diese Information kann in der weiteren Angebotsauswertung im Einkauf als Entscheidungskriterium genutzt werden. Ausgehend von einer strukturierten, XML-kodierten Anfrage – die verschiedenen Marktplätzen zur Verfügung gestellt werden kann – erfolgt über das klassifikationsbasierte Produkt-Retrieval ein Resultatsexport in strukturierter Form, der den Geschäftsprozess abrundet und eine direkte Integration in die ERP- / E-Procurement-Systeme des Anfragenden ermöglicht.

4 Ausblick

Das vorgestellte Produkt-Retrieval-Verfahren errechnet die Ähnlichkeit zwischen Produkt-Merkmalswerten über Zugehörigkeitsfunktionen der Fuzzy-Logik und ermittelt über die Berechnungsformeln des erweiterten booleschen Modells eine Ähnlichkeit

zwischen Suchanfrage und Produkt. Um die Ergebnisse des vorgestellten Verfahrens weiter zu verbessern, müssen die Zugehörigkeitsfunktionen sowohl von der Art der Funktion als auch in ihren Parametern variiert werden. Wie in dieser Arbeit gezeigt, hat eine Änderung des Parameterwertes p von 1 auf 3 beim erweiterten booleschen Modell das Ranking nicht beeinflusst, sondern nur zu einer Verschiebung des Relevanzbereiches geführt. In diesem Zusammenhang sind auch andere Verfahren zur Berechnung der Gesamtrelevanz zu betrachten. So hat eine Adaption des Vektorraum-Modells [RB99] mit Merkmal/Wert-Paaren als Indizes bei großen Datenbeständen ebenfalls zu relevanten Ergebnissen geführt [Na03]. Um den Betreiber eines elektronischen Marktplatzes oder eines Webshops zu entlasten, muss auch untersucht werden, inwieweit die Kalibrierung der vielfältigen Parameter automatisiert werden kann. Hier sind geeignete Lern- und Anpassungsalgorithmen zu konzipieren. Auch Algorithmen zur qualitativen Verbesserung des Datenbestandes sind zu entwickeln, da gerade auf öffentlichen Marktplätzen die Datenerfassung durch einen breiten, unterschiedlich qualifizierten Personenkreis erfolgt. Die vollständige Integration der Suche in den Beschaffungsprozess eröffnet ebenfalls erhebliches Optimierungspotenzial. So können Ähnlichkeitswerte einzelner Merkmale bei der Angebotsauswertung im Einkauf genutzt werden.

Literatur

- [Bu02] Bundesverband Materialwirtschaft, Einkauf und Logistik (2002): BMEcat 1.2, <http://www.bmecat.org/>, Abruf am 9. Juli 2003
- [Co02] Commerce One (2002): xCBL 4.0, <http://www.xcbl.org/>, Abruf am 9. Juli 2003
- [Fu00] Fuhr, Norbert (2000): Models in Information Retrieval, Veröffentlichung der Universität Dortmund
- [Kr02] Krieger, Rolf; Kuhn, Norbert; Mees, Michael; Naumann, Stefan; Schürmann, Cordula; Sommer, Christian (2002): Verwendbarkeit von Klassifikationssystemen und Katalogstandards zum Aufbau von elektronischen Handelsbörsen für gebrauchte Maschinen. In: Weinhardt, Christof; Holtmann, Carsten (Hrsg.): E-Commerce: Netze, Märkte, Technologien. Heidelberg
- [LSD02] Leukel, Jörg; Schmitz, Volker; Dorloff, Frank-Dieter (2002): Referenzmodell für den Austausch von Produktklassifikationssystemen im E-Business. In: Loos, Peter; Gronau, Norbert (Hrsg.): E-Business - Integration industrieller ERP-Architekturen. Göttingen
- [Na03] Naumann, Stefan; Krieger, Rolf; Kuhn, Norbert; Schürmann, Cordula; Sommer, Christian: Such- und Klassifizierungsstrategien in elektronischen Produktkatalogen. Veröffentlichung im Rahmen der 6. Internationalen Tagung Wirtschaftsinformatik 2003, Dresden
- [OMT91] Ogawa, Yasushi; Morita, Tetsuya; Kobayashi, Kiyohiko (1991): A fuzzy document retrieval system using the keyword connection matrix and a learning method. In: Fuzzy Sets and Systems 39:163-179.
- [RB99] Ribeiro-Neto, Berthier; Baeza-Yates, Ricardo (1999): Modern Information Retrieval. Addison-Wesley, Harlow
- [Ri79] van Rijsbergen, Cornelis J. (1979): Information Retrieval, Butterworths
- [SFW83] Salton, Gerard; Fox, Edward A.; Wu, Harry (1983): Extended Boolean Information Retrieval. In: Communications of the ACM 26(11), 1022-1036