Textual CBR im E-Commerce

Mirjam Minor, Mario Lenz

Unter Verwendung intelligenter Techniken können elektronische Einkäufe leichter und bequemer erledigt werden als mit herkömmlichen Systemen. Ein alter Traum der KI, nämlich das maschinelle Verstehen natürlicher geschriebener Sprache, wird durch Textual CBR zwar nicht realisiert, aber wenigstens simuliert: Kunden können ihre Wünsche in normalen Sätzen formulieren. Das elektronische Verkaufssystem findet dazu mit einem ähnlichkeitsbasierten Retrieval-Verfahren schnell eine passende Information oder ein passendes Produkt aus dem aktuellen Angebot. Das Angebot selbst ist ebenfalls weitestgehend in Textform beschrieben. Im Beitrag werden die grundlegenden Ideen von Textual CBR erläutert und mit anderen Ansätzen verglichen. Anschließend werden einige erfolgreiche Anwendungsbeispiele von Textual CBR im E-Commerce vorgestellt.

1 Einleitung

Zur Zeit herrscht Goldgräberstimmung im Online-Business; nahezu täglich erscheinen neue Zahlen und Prognosen von Wachstumsraten im Business-To-Business- und Business-To-Consumer-Geschäft, bisher (Sommer 2000) nur leicht unterbrochen von Börsenschwankungen. Tennenbaum hat drei Faktoren formuliert, deren Missachtung den tatsächlichen Erfolg von E-Commerce-Lösungen hemmen könnten (vgl. [Wilke98]): Ein System muss das Vertrauen (*confidence*) der Kunden bezüglich Sicherheit und Datenschutz gewinnen, es muss komfortabel zu bedienen sein (*convenience*) und es muss Waren anbieten, die den Kunden aufgrund des Preises, des Services oder der großen Auswahl interessieren (*content*). Für das Sortiment an Waren sind die Anbieter selbst verantwortlich, für Sicherheit und Datenschutz gibt es bereits sichere Übertragungsprotokolle und kryptographische Verfahren. Die bequeme Benutzung hingegen lässt bei vielen Systemen noch zu wünschen übrig und ist ein interessantes Feld für Forschungsarbeiten.

Noch sind die häufigsten Methoden, um das passende Produkt im Internet zu finden, Stichwortsuche in Produktdatenbanken und Navigieren in Produktkatalogen per Mausklick. Bei vielen Anwendungen, wie zum Beispiel dem Vorbestellen von Kinokarten in der Stadt Berlin, funktioniert das ganz gut. Bei großen Sortimenten werden diese Techniken aber leicht zum Ärgernis. Statt umständlichen Navigierens und riesiger Suchergebnisse sind einfachere Verfahren und präzisere Ergebnisse gefragt. In jüngster Zeit kommen vor allem für den *pre-sales*-Bereich vermehrt intelligente Techniken zum Einsatz, die diesen Mangel an Bedienerfreundlichkeit beheben sollen.

Das fallbasierte Schließen auf textuellen Daten (*Textual CBR, TCBR*) wurde ursprünglich für Aufgaben im Knowledge Management [LenzEtal98] [MinorHanft2000] entwickelt. TCBR eignet sich sehr gut für die intelligente Suche auf riesigen elektronischen Datenbeständen, die Texte enthalten. Im E-Commerce kann man damit zum Beispiel Produktkataloge oder Fehlerbeschreibungen durchsuchen, um den elektronischen Verkaufsprozess und die Aktivitäten zur Kundenpflege zielgerichtet und schnell zu gestalten. Eine wichtige Voraussetzung an die elektronischen Datenbestände ist, dass die Daten zu einem (evtl. auch mehreren) bestimmten Themenkreis gehören, da im TCBR spezifisches Wissen über die Domäne benutzt wird. Es wäre auch verwirrend für die Kunden, wenn etwa Urlaubsreisen, Jeanshosen und Grundstücke durcheinander angeboten würden.

2 TCBR in einer E-Commerce-Lösung

Ein TCBR-System im E-Commerce funktioniert nach folgendem Grundprinzip: Die Beschreibung eines Produkts oder Fehlers wird als Fall betrachtet und in einem Retrieval-

Server verwaltet. Kunden können in natürlicher Sprache Anfragen an den Retrieval-Server stellen. Dieser ermittelt in einem Retrieval-Prozess diejenigen Fälle aus seiner Fallbasis, die am besten zu der Anfrage (Query) passen, und präsentiert sie in einer Übersicht.
TCBR hat einige Vorteile, die es für den Einsatz im E-Commerce prädestinieren:

- Das Retrieval geht sehr schnell, d.h. durch Nutzung geeigneter Indizierungs- und Speicherstrukturen ist ein Zugriff auf die geeignetsten Fälle deutlich unter einer Sekunde möglich.
- Man kann die Fälle automatisch aus elektronischen Dokumenten erzeugen lassen, d.h. ein *case authoring* entfällt.
- Domänenwissen kann relativ einfach modelliert und im Retrieval ausgenutzt werden (siehe Kapitel 4).
- Die Fallbasis kann jederzeit ausgetauscht werden, um den aktuellen Zustand der Dokumentensammlung abzubilden.
- Die Bedienung ist intuitiv verständlich, da Anfragen in natürlicher Sprache eingegeben werden können.
- Die Suche ist anonym, d.h. es müssen keine benutzerbezogenen Daten gespeichert werden, was zum Beispiel in agentenorientierten Ansätzen zu Datenschutz- und Akzeptanzproblemen führen kann.
- Die CBR-typische Trennung von Problem und Lösung innerhalb eines Falls kann unterbleiben.

3 Textual Case Based Reasoning

Bei konventionellen CBR-Systemen müssen Fälle in einem sogenannten *case authoring* Prozess explizit eingegeben und repräsentiert werden. Dies entfällt beim TCBR, jedoch müssen Texte erst vorverarbeitet und in interne Strukturen überführt werden, bevor sie maschinell weiter verarbeitet und miteinander verglichen werden können.
Es genügt nicht, einfach die gemeinsam vorkommenden Wörter zu zählen, um herauszufinden, ob zwei Texte (zum Beispiel eine Anfrage und ein Text aus einem Fall) auch inhaltlich etwas miteinander zu tun haben. Zum einen drücken sich verschiedene Autoren sehr unterschiedlich aus, um dasselbe zu sagen. Persönliche Wortschätze, spezielle Abkürzungen und Schreibweisen von Fachbegriffen spielen dabei eine Rolle. Zum anderen sind vor allem im Deutschen sehr viele Satzbaumuster und grammatikalische Formen möglich, um einen Sachverhalt zu beschreiben. Ein reiner Vergleich von Buchstabenketten (*string matching*) liefert daher nur sehr unzureichende Ergebnisse, wie man dies etwa von den üblichen Suchmaschinen im WWW kennt.

Meist werden im TCBR die Fälle aus bereits bestehenden elektronischen Quellen automatisch generiert. Das können unstrukturierte elektronische Dokumente sein, die in Fälle mit genau einem Textabschnitt transformiert werden. Haben die Datenquellen bereits eine Struktur, wie zum Beispiel einen Frage- und Antwortteil in FAQ-Dokumenten, kann diese Struktur auch in das Fallformat übernommen werden, so dass ein Fall mehrere Abschnitte hat. Ein Fall kann Textpassagen und Attribut-Werte-Paare sogar aus mehreren Datenquellen, z.B. Datenbanken und Dokumentensammlungen, beziehen. Typische Attribut-Werte-Paare sind zum Beispiel die Automarke in Kleinanzeigen oder die Farbe eines Produkts. Für das Retrieval gibt es dann verschiedene Strategien:

1. Alle Abschnitte werden mit der ganzen Anfrage verglichen.
2. Es werden Abschnitte des Fallformats definiert, die im Retrieval berücksichtigt werden sollen. Die übrigen Abschnitte sind nur zur Information der Benutzer gedacht.
3. Die Query ist ebenfalls in Abschnitte unterteilt, und jeder Abschnitt der Anfrage wird separat mit seinem Partnerabschnitt im jeweiligen Fall verglichen.

Die zweite Strategie ist beispielsweise bei FAQ-Dokumenten sinnvoll, um die Anfrage nur mit dem Frageteil der Dokumente zu vergleichen. Die dritte Strategie lohnt sich nur, wenn die Benutzer unterschiedliche Abschnitte spezifizieren möchten, zum Beispiel die Beschreibung eines Hauses und des dazugehörigen Grundstücks. Die Quadratmeterzahlen im Abschnitt „Grundstück“ sollen natürlich nicht verglichen werden mit Größenangaben des Abschnitts „Wohnfläche“.

3.1 Indizierung von Texten

Die Ähnlichkeit zweier Texte wird mit Hilfe von Konzepten bestimmt, die die Texte repräsentieren. Die Textabschnitte innerhalb der Fälle werden dazu auf Mengen von Informationseinheiten abgebildet. Im Textual CBR besteht eine Informationseinheit aus einem natürlichsprachlichen Begriff mit seinen verschiedenen Ausprägungen. Das können grammatikalische Formen, synonyme Formulierungen, Abkürzungen oder sogar Übersetzungen in andere Sprachen sein (siehe Abb. 1).

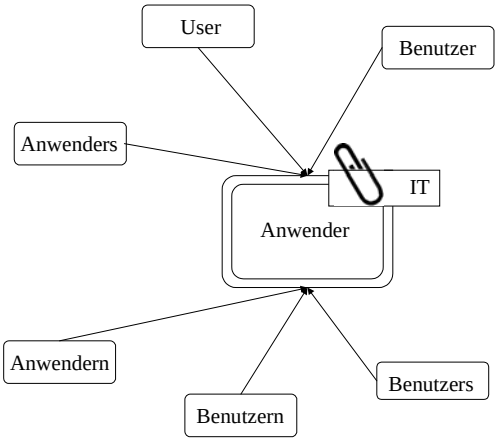

Abb.1: Beispiel einer Informationseinheit

Jede Informationseinheit ist außerdem einer Wortschatzkategorie zugeordnet. Neben der Kategorie „allgemeinsprachlicher Grundwortschatz“ gibt es Fachbegriffe verschiedener Domänen, zum Beispiel der allgemeinen IT-Welt oder einer bestimmten Produktlinie.

Ein Lexikon mit Informationseinheiten wird sowohl bei der Indizierung der Fälle als auch bei der Vorverarbeitung der Anfragen für das Retrieval verwendet. Die Abbildung von Texten auf Mengen von Informationseinheiten geschieht automatisch.

In orenge[1], tec:inno’s[2] komponentenbasierter Suchtechnologie, geht man sogar noch einen Schritt weiter: Anstelle der oben angesprochenen Kategorien stehen Klassen, zwischen denen Vererbungs- und Referenzbeziehungen definiert werden können, wie man sie aus objektorientierten Systemen kennt. Eine Informationseinheit ist dann ein Konzept, d.h. letztlich eine Instanz einer solchen Klasse. Vorteilhaft dabei ist, dass es möglich wird, in den zugrundeliegenden Modellen klassenspezifische Dinge zu berücksichtigen, wie etwa spezielle Ähnlichkeitsmaße oder aber auch spezielle Methoden, wie Klassenwerte aus Texten extrahiert werden können. In letzterem Sinne stellen die oben erläuterten Informationseinheiten nur

[1] Open Retrieval ENGinE
[2] www.tecinno.com

einen Spezialfall dar, genauso ist es auch möglich mittels *feature extraction* [RiloffLehnert94] automatisch Werte von weiteren Attributen zu extrahieren, etwa den Preis eines Produktes oder das Datum der Erstellung eines Dokumentes. Dadurch wird es möglich, auch strukturierte Informationen wie Attribut-Wert-Paare direkt aus Texten heraus zu instanziieren.

Auch jedes Attribut-Wert-Paar wird auf eine Informationseinheit abgebildet. Ein Fall wird nach der Indizierungsphase also durch eine Menge von Informationseinheiten repräsentiert.

3.2 Retrieval

Um Retrieval-Prozesse durchführen zu können, muss eine Ähnlichkeitsfunktion definiert sein, mit deren Hilfe die Fälle der Fallbasis in eine partielle Ordnung bezüglich der Anfrage (Query) gebracht werden können. Diese partielle Ordnung bestimmt die Auswahl und Reihenfolge der *best matching cases*, die als Suchergebnis angezeigt werden. Im Textual CBR wird eine kompositorische Ähnlichkeitsfunktion verwendet: Lokale Ähnlichkeitsbeziehungen zwischen Informationseinheiten werden ausgewertet, um einen globalen Ähnlichkeitswert der Query zu einem Fall auszurechnen. Oft genügt eine gewichtete Summe über die lokalen Ähnlichkeitswerte als globale Ähnlichkeitsfunktion:

Beispiel einer einfachen globalen Ähnlichkeitsfunktion
Sei E die Menge bekannter Informationseinheiten und $sim : E \times E \rightarrow [0,1]$ eine lokale Ähnlichkeitsfunktion. $Query \subseteq E$ sei eine Menge von Informationseinheiten, die die Anfrage repräsentiert, und $Case \subseteq E$ eine, die einen Fall repräsentiert.

$$SIM(Query, Case) = \sum_{e_i \in Query} \sum_{e_j \in Case} sim(e_i, e_j).$$

Durch eine Normierung des Werts bezüglich der Kardinalität von $Query$ kann man den Wertebereich von SIM zwischen 0 und 1 halten.

Beispiel einer globalen Ähnlichkeitsfunktion, die leere Abschnitte ignoriert
Seien $S^1_{Case}, S^2_{Case}, ..., S^k_{Case}$ die Abschnitte eines Falls ($S^1_{Query}, S^2_{Query}, ..., S^k_{Query}$ die einer Query), wobei die Ähnlichkeitsfunktion zwischen Partnerabschnitten definiert ist als

$$SIM(S^l_{Case}, S^l_{Query}) = \begin{cases} \sum_{e_i \in S^l_{Case}} \sum_{e_j \in S^l_{Query}} sim(e_i, e_j), \text{ wenn } S^l_{Case} \neq \emptyset \text{ und } S^l_{Query} \neq \emptyset \\ 0, \text{ sonst.} \end{cases}$$

Die globale Ähnlichkeitsfunktion ist dann definiert als

$$SIM(Query, Case) = \frac{1}{\alpha} \cdot SIM(S^1_{Case}, S^1_{Query}) + \frac{1}{\alpha} \cdot SIM(S^2_{Case}, S^2_{Query}) + \ldots + \frac{1}{\alpha} \cdot SIM(S^k_{Case}, S^k_{Query}).$$

Die globalen Ähnlichkeitswerte setzen sich aus den Werten der partiellen Ähnlichkeitsfunktion zwischen den Partnerabschnitten zusammen. $SIM(S^l_{Case}, S^l_{Query})$ *wird auf 0 gesetzt, wenn der Abschnitt l in der Query, im Fall oder in beiden leer ist.* $1/\alpha$ *dient der Normierung des globalen Ähnlichkeitswertes, wobei* $\alpha \in N$ *die Anzahl beiderseits gefüllter Abschnitte ist.*

Beide Beispielfunktionen liefern den Wert 0 für Fallrepräsentationen, die völlig disjunkt zur Anfrage sind, und 1 für zur Query identische Fallrepräsentationen.
Die Werte für die lokalen Ähnlichkeitsbeziehungen sind wie die Informationseinheiten selbst in einem Lexikon definiert. Semantisch spiegeln sie sprachliche Ähnlichkeiten im Sinne eines Thesaurus oder domänenspezifische Ähnlichkeiten zum Beispiel zwischen Produktnamen wieder.

Mit Hilfe eines *Case Retrieval Nets* kann ein effizientes Retrieval auf den Fällen durchgeführt werden. In einem *spreading-activation*-Prozess werden initiale Aktivierungen von der Anfrage durch das Netz bis zu den Fällen propagiert (siehe [LenzBurkhard96]).

4 Die Pflege der Wissensbasis

Das Wissen eines TCBR-Systems steckt in den Lexika (Informationseinheiten und lokale Ähnlichkeitsbeziehungen zwischen Informationseinheiten) und in den Fällen. Die Pflege der Fallbasis ist relativ einfach, da die Fälle entweder aus vorhandenen Texten extrahiert oder ohne besondere Vorkenntnisse direkt für das CBR-System geschrieben werden können. Auch die Pflege der Lexika ist leicht verständlich, aber leider etwas aufwendiger. Warum sich der Aufwand lohnt, wird im Vergleich mit Information-Retrieval-Techniken in Kapitel 5 deutlich. Ein Grundstock an allgemeinsprachlichen und IT-Vokabeln kann aus Online-Wörterbüchern automatisch extrahiert werden. Dazu kommen Fachbegriffe, die aus Glossaren und Fachtexten in das Lexikon eingepflegt werden müssen. Die lokalen Ähnlichkeitsbeziehungen stammen teilweise aus elektronischen Thesauri, teilweise müssen sie von Hand in das Ähnlichkeiten-Lexikon eingetragen werden. Vor allem die fachspezifischen Ähnlichkeitsbeziehungen können nur von Fachleuten der Domäne definiert werden, da beispielsweise Produktnamen, die ähnlich klingen, durchaus verschiedene Produkte bezeichnen können.
Je nach Spezialisierungsgrad der Domäne ist es einfach oder schwierig, geeignete Experten für die Modellierung zu finden. Die Domäne Gebrauchtwagenmarkt beispielsweise ist im Vergleich zum Verkauf elektronischer Bauteile einfach zu erfassen, da viele Autofahrer sich mit Gebrauchtwagen auskennen. Die Experten können in zeitlichen Abständen den Inhalt der Lexika aktualisieren und dies zum Beispiel in Stoßzeiten auf die lange Bank schieben. Das Retrieval-Verfahren kann auch mit Lücken in den Lexika brauchbare Ergebnisse liefern. Je sorgfältiger jedoch die Wissensakquisition für die Lexika durchgeführt wird, desto präzisere Ergebnisse liefert die Ähnlichkeitsberechnung.

5 Ähnliche Ansätze

Traditionell beschäftigen sich die Forschungsrichtungen *Information Retrieval* (IR) und *Natural Language Processing* (NLP) mit der Informationssuche in elektronischen Texten. Beide Techniken werden auch im E-Commerce eingesetzt, wobei NLP im E-Commerce bisher wohl eher von akademischem oder spielerischem Interesse ist.
Es gibt beispielsweise ein *Pattern-Matching*-System namens Lydia[3], das sich in der Gestalt einer Schlange mit potentiellen Buchkäufern „unterhält“. Lydia ist relativ „dumm“ und macht den Verkaufsprozess vielleicht lustiger, aber keineswegs schneller und bequemer. Das in

[3] http://www.liveclub.de

NLP-Systemen enthaltene Wissen (z.B. in pattern oder parse trees) ist relativ aufwendig zu modellieren und deshalb inflexibel gegenüber Änderungen. Experten, die Domänenwissen in das System eingeben wollen, müssen entweder selbst KI-Experten sein oder mit solchen zusammenarbeiten. Letzteres ist sehr zeitintensiv und kann leicht zu Missverständnissen führen.

Information Retrieval (neuerdings auch *Knowledge Retrieval)* wird im E-Commerce erfolgreich für eine verbesserte Katalogsuche eingesetzt. Die Firma Verity[4] bietet zum Beispiel E-Commerce-Systeme mit IR-Techniken für Gesetzes-Informationsdienste oder Urlaubsreisen an. Texte werden mit Hilfe von *Stemming*-Algorithmen indiziert und zum Beispiel auf Vektoren von Indizes abgebildet. Die Ähnlichkeitsberechnung erfolgt durch Vergleichen der Indizes. Anstelle von sprachlichem und ontologischem Wissen werden statistisch ermittelte Werte von Indexeinträgen verwendet. Ein bekanntes Verfahren ist das *Vector-Space*-Modell [SaltonMcGill83], das ein algebraisches Maß als Ähnlichkeitsfunktion einsetzt und auch gerne in attributbasierten CBR-Systemen verwendet wird. Im Gegensatz zu TCBR-Systemen sind IR-Systeme nicht auf domänenspezifische Daten fixiert. Einerseits bringt dies Vorteile, da kein Aufwand für Wissensakquisition anfällt, andererseits sind die Ähnlichkeitsfunktionen weniger mächtig als wissensbasierte: Das Problem unterschiedlicher Formulierungen (*paraphrase problem*) kann dazu führen, dass ein Kunde nicht findet, was er sucht. Werden außerdem Texte aus verschiedenen Kontexten indiziert, kann zusätzlich das Problem mehrdeutiger Begriffe auftreten (*ambiguity problem*). Eine „Blume“ in einer Weinhandlung hat eine andere Semantik als eine „Blume“ in einem Gartencenter. Im TCBR dagegen wird dieses Problem durch die Fokussierung auf eine bestimmte Domäne weitestgehend umgangen.

Fallbasiertes Schließen selbst wird natürlich auch in seiner konventionellen strukturierten Variante für E-Commerce-Anwendungen eingesetzt. Bei Analog Devices werden elektronische Bauteile im Business-To-Business-Geschäft mit einem von IMS[5] entwickelten System[6] auf Basis von tec:inno‘s[7] CBR-Works sehr erfolgreich verkauft (vgl. [VollrathEtal98]). Die Benutzer spezifizieren ein gewünschtes Produkt mit einigen Attributen, die elektrische Eigenschaften der Operationsverstärker beschreiben und deren Werte in physikalischen Einheiten oder als Informationen wie z.B. „das beste“, „in der Art“ oder „weniger als“ ausgedrückt sind. Die Ähnlichkeitsfunktionen sind hochgradig komplex, da sie physikalische Eigenschaften z.B. in logarithmischen Arbeitskurven von Bauteilen modellieren müssen. Fehlende Werte (*missing values*) und unscharfe Begriffe verkraftet das System ebenso wie im TCBR. Durch die Strukturierung der Anfrage in feste Attributnamen sind natürlich nicht so viele Freiheiten bezüglich der Produktdaten und Anfragen wie im CBR mit freien Texten gegeben.

Eine andere Technik aus der Nachbarschaft intelligenter Systeme ist *Active Collaborative Filtering* (ACF). Anstelle von Domänenwissen, wie im TCBR, werden Statistiken über das Verhalten der Benutzer gesammelt und ausgewertet. Data Mining-Techniken ermitteln zum Beispiel, welche Artikel oft im Zusammenhang mit anderen Artikeln gekauft wurden. Das Verkaufssystem macht den Kunden Vorschläge, welche Produkte gut zu den gerade gewählten passen könnten. Bei Amazon[8] bekommt man nach einer Produktsuche mit Hilfe von ACF Informationen zu weiteren Büchern, Videos oder CDs. Bei jeder Art von Benutzerprofilen stellen sich allerdings immer Fragen des Datenschutzes und der Akzeptanz vonseiten der Kunden. Wer nicht möchte, dass das eigene Kaufverhalten mitprotokolliert wird, schaltet die „Intelligenz“ des Systems einfach aus.

---

[4] http://www.verity.com
[5] http://www.imsgrp.com
[6] http://www.imsgrp.com/analog/query.htm
[7] http://www.tecinno.com
[8] www.amazon.com

Es gibt eine Reihe von Techniken, die so ähnlich wie TCBR funktionieren, aber (noch) nicht in E-Commerce-Systemen eingesetzt werden, obwohl sie durchaus dafür in Frage kämen. FAQFinder [BurkeEtal97] ist ein System, das fallbasiertes Schließen mit Techniken des Dokumentenmanagement verbindet und wie TCBR einen elektronischen Thesaurus verwendet, um seinen Retrieval-Mechanismus auf einer semantischen Wissensbasis aufzubauen. FAQFinder beantwortet Anfragen in natürlicher Sprache mit Einträgen aus FAQ-Dokumenten. Im Gegensatz zu TCBR ist FAQFinder nicht domänenspezifisch, sondern benutzt für alle Gebiete denselben Wortschatz. Die Ähnlichkeitsfunktion ist dadurch universeller aber schwächer als im TCBR, wo Fachbegriffe (z.B. PostScript Drucker) und deren Ähnlichkeitsbeziehungen (z.B. von PostScript Drucker zu Drucker) verwendet werden.
Im fallbasierten System CATO [BrueninghausAshley99] werden Gesetzestexte für Retrieval-Prozesse indiziert. Das Indizierungsverfahren wird durch maschinelles Lernen von Textzusammenfassungen, die bereits manuell auf abstrakte Konzepte eines Fakts abgebildet wurden, automatisiert. Die initiale Zuordnung von Konzepten zu Texten ist sehr aufwändig.

6 Anwendungsbeispiele für TCBR im E-Commerce

6.1 Gebrauchtwagen im Internet

Der Kübler-Verlag in Lampertheim ist europaweit führend auf dem Gebiet der Kleinanzeigen-Blätter. Quoka, sozusagen die Online-Division des Verlages, hat es sich auf die Fahnen geschrieben, diesen immensen *Content* nun auch im WWW zugreifbar zu machen. Eine erste Lösung wurde auf Basis von tec:inno Produkten für den Bereich Gebrauchtwagen implementiert und ging im März 2000 unter www.autoaktuell.de online. Obwohl man ein gebrauchtes Auto sicherlich relativ leicht durch eine Reihe von Attributen in strukturierter Form repräsentieren kann, waren für diese *pre-sales*-Applikation Komponenten des TCBR notwendig, da es eines der vorrangigen Ziele war, die vorliegenden Daten aus den Print-Medien automatisch mit in das System aufnehmen zu können. Es war daher notwendig, mittels Technologien aus orenge Komponenten bereitzustellen, die den typischen Kleinanzeigen-Jargon insofern verstehen, dass damit eine interne strukturierte Repräsentation möglich wird, auf Basis derer wiederum eine intelligente Suche realisiert werden kann. Eine weitere Herausforderung war die internationale Ausrichtung von Quoka, d.h. es musste gewährleistet sein, dass mit geringem Aufwand erstens Anzeigen in verschiedenen Sprachen verarbeitet werden können und zweitens dass zum Beispiel ein Besucher aus der Schweiz sowohl deutsche, französische als auch italienische Angebote erhält.
Das Ergebnis dieser Arbeiten ist unter der oben angegebenen URL verfügbar. Derzeit sind allein weit über 100.000 Kleinanzeigen aus dem deutschsprachigen Raum im System abrufbar, die mehrmals wöchentlich aktualisiert werden. Erweiterungen sowohl hinsichtlich neuer Regionen als auch für andere Arten von Kleinanzeigen sind geplant.

6.2 Intelligente Kundenpflege am Beispiel des SIMATIC KnowledgeManagers

Der SIMATIC Knowledge Manager (SKM) [LenzEtal98a] [Lenz99] ist ein umfangreiches Self-Service-Angebot von Siemens Automation & Drives[9]. Mit weltweit ca. 52.000 Mitarbeitern, 66 Fertigungsstandorten und ca. 13 Mrd. DM Umsatz ist Siemens A&D der Marktführer im Bereich Industrieautomatisierung. Kunden und Mitarbeiter können durch den Zugriff auf das in mehr als zehntausend FAQs (Frequently Asked Questions) und in anderen Support-Dokumenten enthaltene Fachwissen der Servicetechniker Probleme selbst lösen, für die sonst ein Anruf bei der SIMATIC Hotline nötig wäre. Vorrangiges Ziel bei diesem Projekt

[9] http://www.ad.siemens.de:8080/cgi-bin2/skm_00/skm2/skmCGI?Func=Login&Access=ext&LANGUAGE=D

war daher die *call avoidance*, d.h. die Hilfe zur Selbsthilfe, damit Kunden nur noch bei wirklich schwierigen Problemen anrufen und sich die hochqualifizierten Hotline-Mitarbeiter somit auf diese konzentrieren können. Derzeit werden bei Siemens A&D hierdurch ca. 500 TDM pro Monat im *post-sales*-Bereich eingespart.

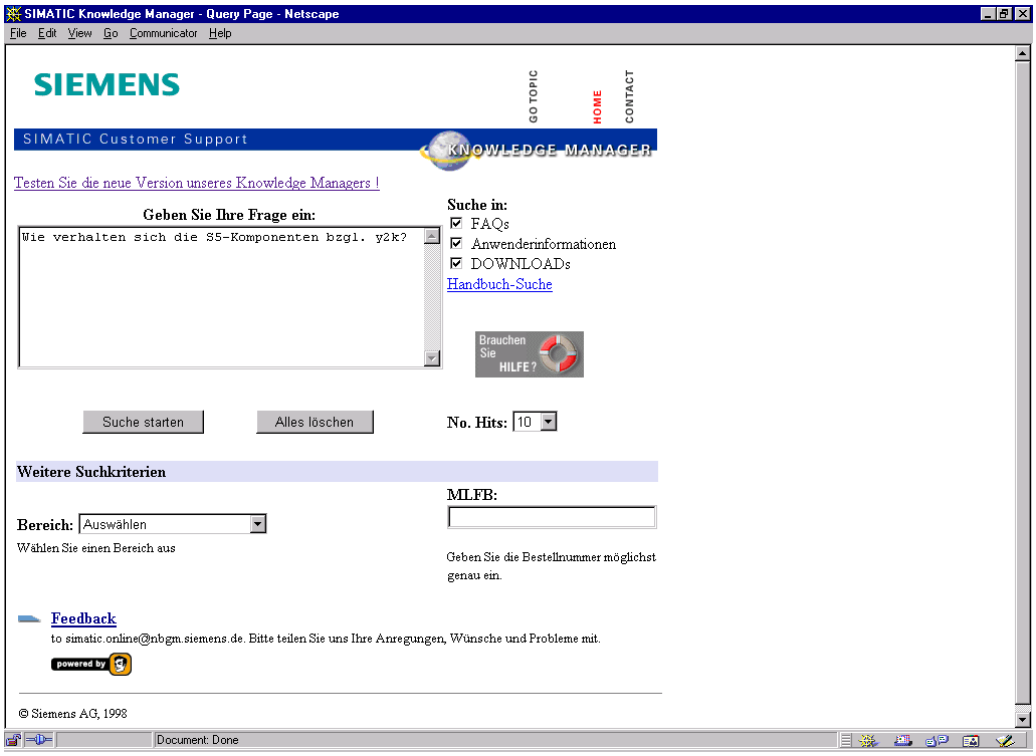

Abb. 1: Der SIMATIC Knowledge Manager im WWW

Eine technische Besonderheit des SKM ist, dass die Benutzer Filterkriterien wie zum Beispiel bestimmte Produktlinien angeben können. Neben der Bereitstellung des TCBR-Systems in Intra- und Internet werden auch regelmäßig offline-Varianten mit aktuellen Falldaten auf CD-Rom ausgeliefert.

7 Zusammenfassung und Ausblick

In diesem Fachbeitrag wurden die Grundlagen und jüngste Entwicklungen des Fallbasierten Schließens auf textuellen Daten erläutert, die Eigenschaften dieser Technik im Hinblick auf E-Commerce-Anwendungen diskutiert und schließlich einige dieser Anwendungen vorgestellt. Textual CBR eignet sich sowohl für das Business-To-Business- als auch für das Business-To-Consumer-Geschäft. Seine besonderen Stärken entfalten sich, wenn Daten in Textform vorliegen, die nicht einfach in einen Katalog einsortiert werden können und durch eine Stichwortsuche unzureichend durchforstet würden. Stammen die Daten aus demselben Themenkreis beispielsweise einer Produktfamilie, so bringt ein wissensbasiertes Retrieval wesentlich genauere Suchergebnisse als Information-Retrieval-Verfahren.
Obwohl es schon einige erfolgreiche E-Commerce-Anwendungen gibt, die Textual CBR für die Produktauswahl und für die Kundenpflege einsetzen, gibt es eine Vielzahl potentieller Anwendungsgebiete im Web. Denkbar wäre etwa der Verkauf von Ergebnissen einer Informationsrecherche in Webtexten oder eine Kombination von TCBR mit agentenorientierten Techniken zum Beispiel für die Verwaltung der Dienste-Wissensbasis eines Assistenzagenten.

Literatur

[BrueninghausAshley99] S. Brüninghaus, K. Ashley: Bootstrapping Case Base Development with Annotated Case Summaries. In: Proceedings of ICCBR-99, LNAI 1650, Springer Verlag, Berlin, 1999.

[BurkeEtal97] R. Burke, K. Hammond, V. Kulyukin, S. Lytinen, S. Schoenberg: Natural Language Proessing in the FAQ Finder System: Results and Prospects. In: Working Notes AAAI Spring Symposium NLP for the WWW, Stanford University, CA, 1997.

[KunzeHuebner98] M. Kunze, A. Hübner: CBR on Semi-structured Documents: The ExperienceBook and the FAllQ Project. In: Proceedings of the GWCBR-98, IMIB Report, University of Rostock, 1998.

[Lenz99] M.Lenz: Case Retrieval Nets as a Model for Building Flexible Information Systems, Dissertation, Humboldt-Universität zu Berlin, 1999.

[LenzBurkhard96] M. Lenz, H. D. Burkhard: Case Retrieal Nets: Basic Ideas and Extensions. In: G. Görz, S. Hölldobler (Hrsg.), KI-96: Advances in Artificial Intelligence, LNAI 1137, Springer Verlag, Berlin, 1996.

[LenzEtal98] M. Lenz, A. Hübner, M. Kunze: Textual CBR. In: Case-Based Reasoning Technology - From Foundations to Applications, Lenz M., Burkhard HD., Bartsch-Spörl B., Wess S. (Hrsg.), LNAI 1400,Springer Verlag, Berlin, 1998.

[LenzEtal98a] M. Lenz, A. Hübner, M. Kunze: Question Answering with Textual CBR. In: T. Andreasen, H. Christiansen, H. Larsen (Hrsg.), Flexible Query Answering Systems, S. 236 - 247, LNAI 1495, Springer Verlag, Berlin, 1998.

[MinorHanft2000] M. Minor, A. Hanft: Corporate Knowledge Editing with a Life Cycle Model. In: Proceedings of GWCBR 2000, DaimlerChrysler Research and Technology, Ulm, 2000.

[RiloffLehnert94] E. Riloff, W. Lehnert: Information Extraction as a Basis for High-Precision Text Classification. In: ACM Transactions on Information Systems, Vol. 12, No. 3, S. 296 – 333, 1994.

[SaltonMcGill83] G. Salton, M. McGill: Introduction to Modern Information Retrieval, Mc Graw-Hill, New York, 1983.

[VollrathEtal98] I. Vollrath, W. Wilke, R. Bergmann: Case-based reasoning support for online catalog sales. In: IEEE Internet Computing, July-August 1998.

[Wilke98] W. Wilke, Knowledge Management for Intelligent Sales Support in Electronic Commerce, Dissertation, Universität Kaiserslautern, 1998.