Information Systems 1

1. Business Intelligence

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Business Economics and Information Systems
& Institute for Computer Science
University of Hildesheim
http://www.ismll.uni-hildesheim.de

1. Business Intelligence, Data Warehousing und Data Mining

2. Web Usage Mining

3. Recommender-Systeme

Was ist Business Intelligence?

Der Begriff **Business Intelligence** wurde bereits 1989 von Howard Dresner (ab 1993 Gartner Group) geprägt:

> Ein interaktiver Prozeß des Untersuchens und Analysierens strukturierter, domänen-spezifischer Informationen (die oft in einem Data Warehouse gespeichert sind), um **Geschäftstrends oder -muster zu erkennen**, wobei Einsichten abgeleitet werden und Schlußfolgerungen gezogen werden.
>
> Der Business Intelligence-Prozeß umfaßt die Kommunikation der Ergebnisse sowie die Durchführung von Änderungen.
>
> Domänen sind u.a. Kunden, Zulieferer, Produkte, Dienstleistungen und Konkurrenten. [Gro04]

Data Warehousing

1991, zwei Jahre später, hat Inmon den Begriff Data Warehousing geprägt.

> Ein Data Warehouse ist eine
>
> – domänen-orientierte,
>
> – integrierte,
>
> – die Zeit berücksichtigende
>
> – und nicht-flüchtige
>
> Datensammlung zur **Unterstützung von Management-Entscheidungsprozessen**. [Inm92]

Data Warehousing

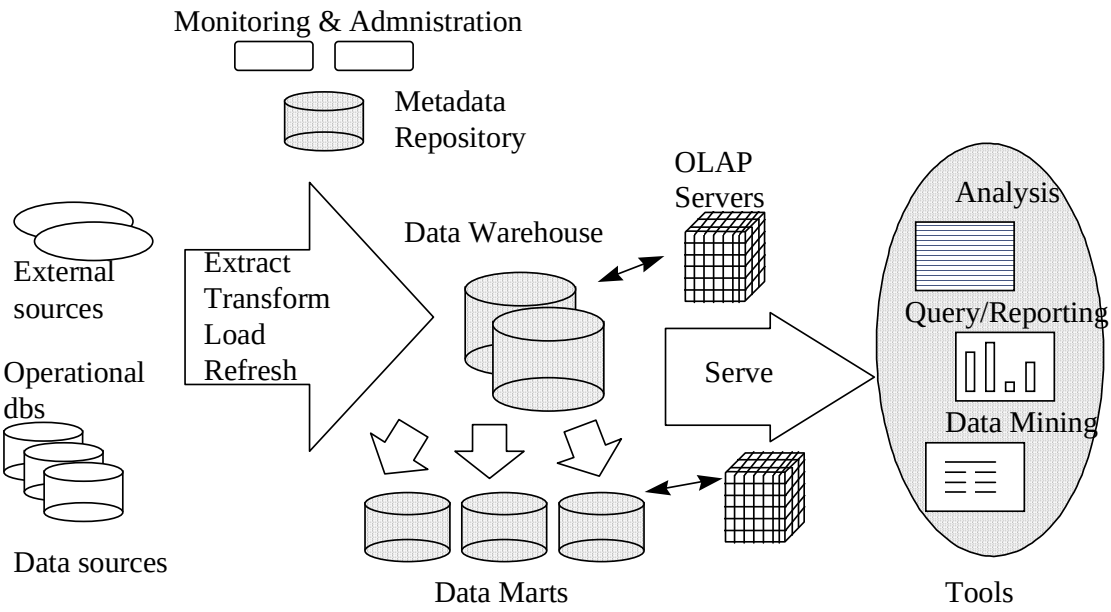
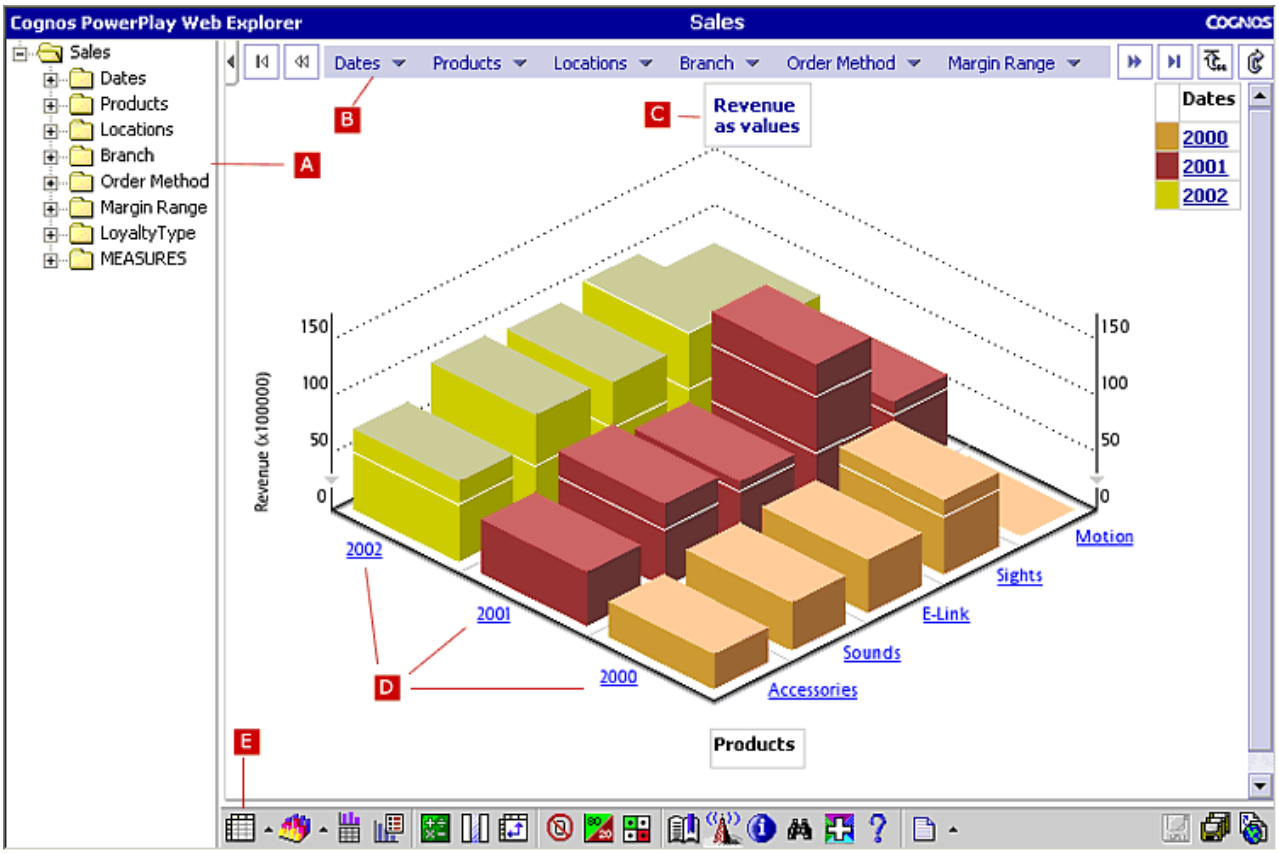
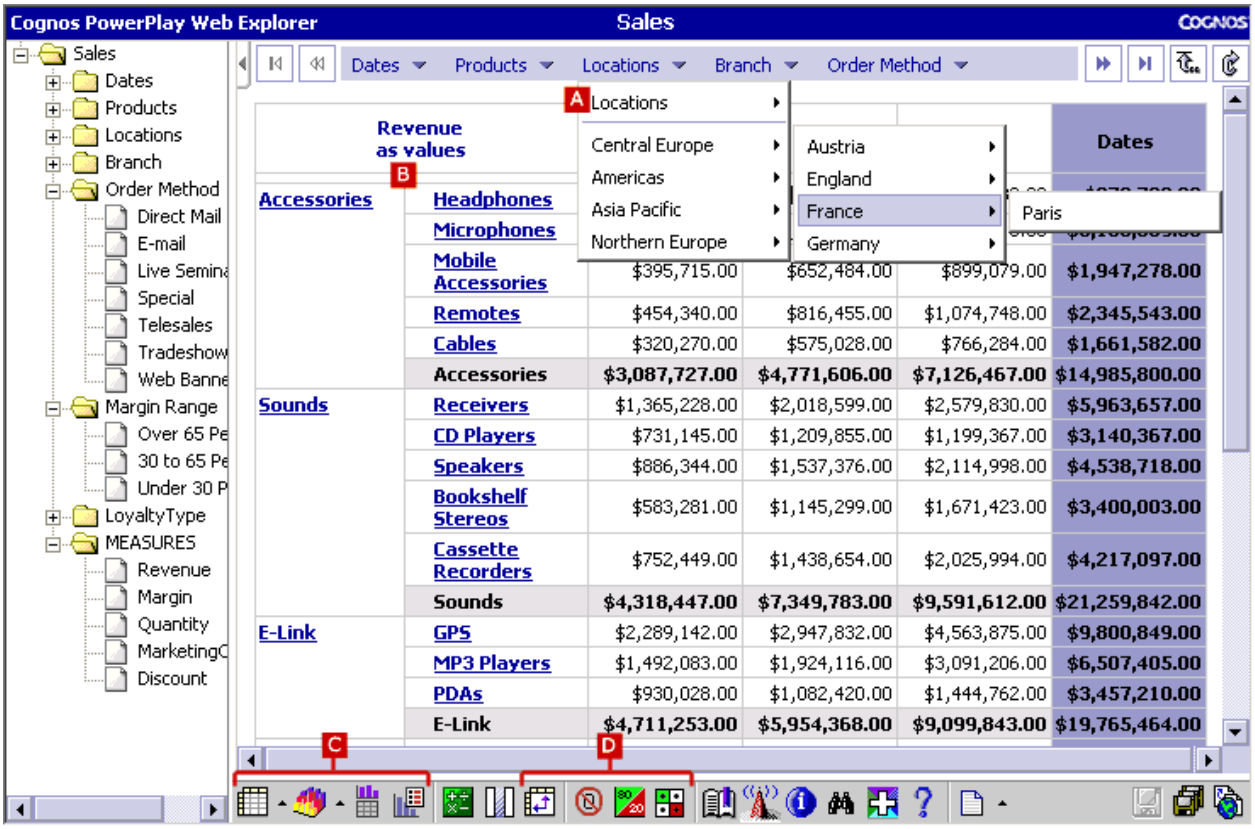


Abbildung 1: Data Warehouse-Architektur [CD97].

Online Analytical Processing (OLAP)



Online Analytical Processing (OLAP)



Data Mining

1995, nochmals vier Jahre später, wurde von Fayyad die Begriffe Knowledge Discovery in Databases (KDD) und Data Mining geprägt.

Knowledge Discovery in Databases (KDD) bezeichnet den nicht-trivialen Prozeß der Identifikation

– valider,

– neuartiger,

– potentiell nützlicher und

– klar verständlicher

Muster in Daten. [FPSS96]

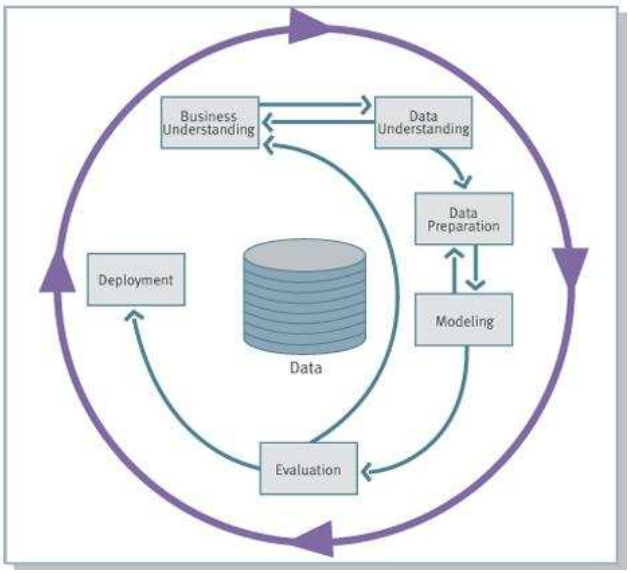


Abbildung 4: Data Mining-Prozeßmodell CRISP [CCK+00, S. 13].

Business Intelligence-Aspekte

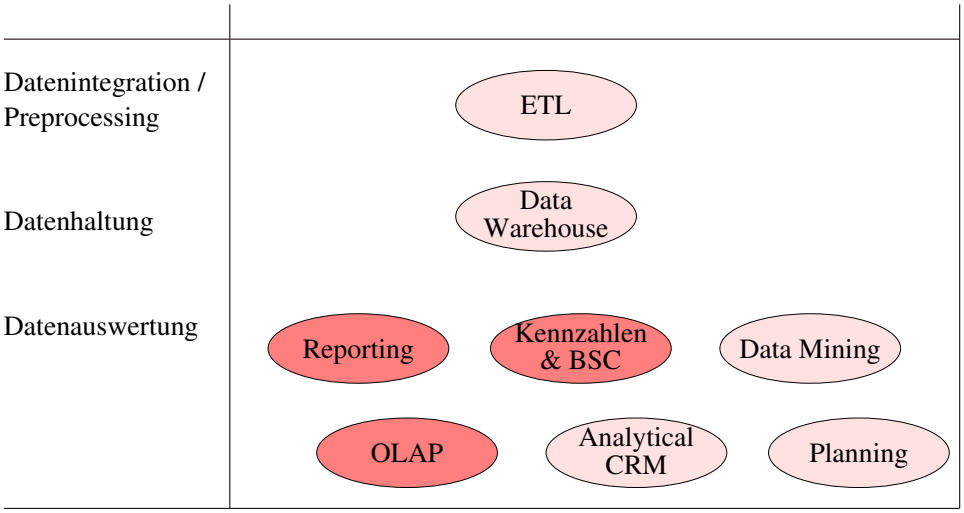


Abbildung 5: Business Intelligence-Aspekte (vgl. auch [DG02, S. 33]).

Business Intelligence-Aspekte

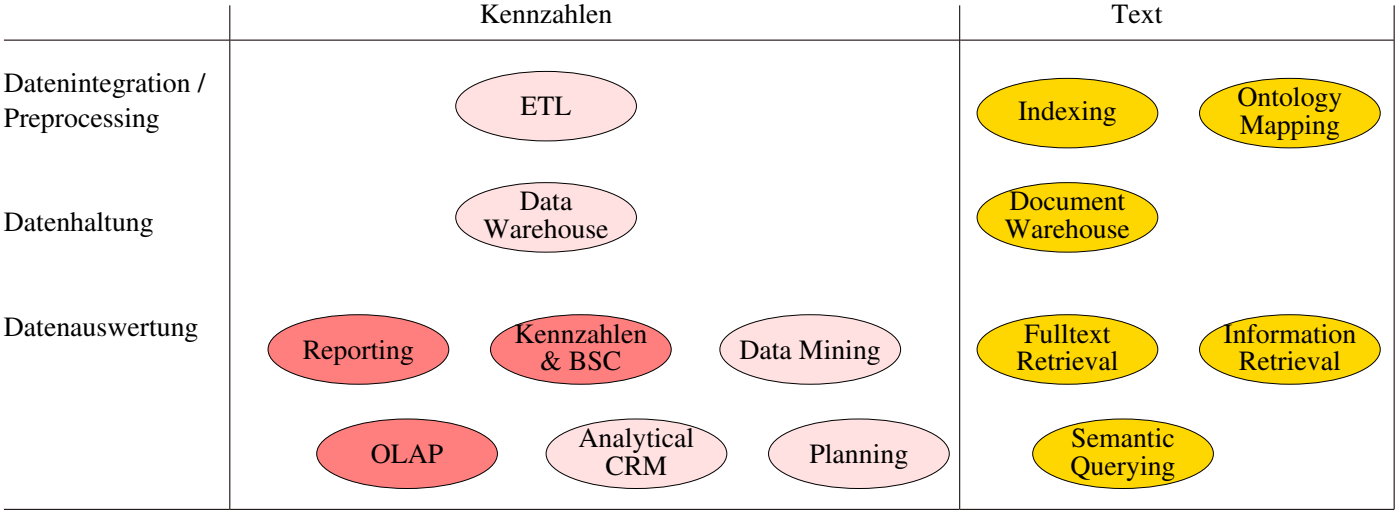


Abbildung 6: Business Intelligence-Aspekte in Abhängigkeit vom Datentyp.

Business Intelligence-Aufgaben

Reporting
Automatisierung regelmäßiger Reports,
auch über Grenzen von Organisationseinheit hinaus.

ad hoc-Reporting, Browsing
instantanes Erstellen von Reports nach Benutzervorgaben,
Verlinken verschiedener Reports zur leichteren Navigation.

Dashboard
tagesaktuelle, benutzerspezifische Übersicht über die wichtigsten Kennzahlen.

Analyse
diverse Auswertungen der erfassten historischen Daten
z.B. Entwicklung von Kundensegmenten, Warenkorbanalysen,
etc.

Vorhersage
Prognose der zeitlichen Entwicklung bestimmter Kennzahlen (z.B. Absatzzahlen) basierend auf historischen Daten.

Business Intelligence für KMU?

Traditionell wird Business Intelligence als größere Investition für größere Unternehmen gesehen.

Seit mehreren Jahren drängen aber auch Anbieter für KMU-Lösungen in den Markt:

Microsoft SQL-Server; Sharepoint; PerformancePoint

Pentaho (open source)

JasperSoft (open source)

BEE (open source)

Openi (open source)

SpagoBI (open source)

Business Intelligence für KMU?

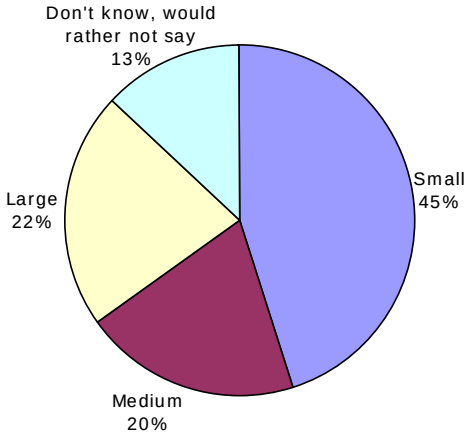


Abbildung 7: Teilnehmer der Verdana Reserach Open Source BI-Studie nach Unternehmensgröße [Res06].

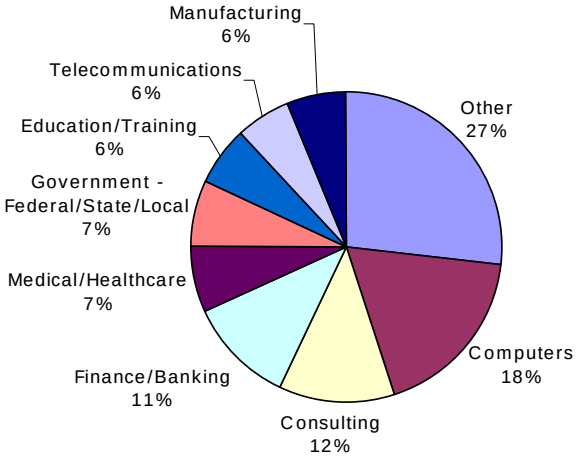


Abbildung 8: . . . und nach Branche [Res06].

Business Intelligence für KMU?

83% haben bereits eine OS BI-Lösung deployed oder denken darüber nach.

Bisher meistens kleine Installationen bis 200 Benutzer (79%).

Aber am Ende größere Installationen geplant:
- 37% mehr als 1000 Benutzer
- 24% 201-1000 Benutzer

Hauptgrund für den Einsatz von Open Source Business Intelligence-Lösungen:
- 20% aufgrund Interesses eines Meinungsführers
- 16% geringere Kosten als kommerzielle Lösung
- 12% geringere Kosten als eigene Lösung

Funktionsumfang: 54% zufrieden, 38% benötigen mehr.
- 16% bessere Unterstützung von Sicherheit
- 16% mehr Datenquellen-Adapter
- 13% verbesserte Verwaltung
- 11% Metadaten-Schicht zur Abfrageentwicklung

[Res06]

1. Business Intelligence, Data Warehousing und Data Mining

2. Web Usage Mining

3. Recommender-Systeme

Web Server-Logfiles

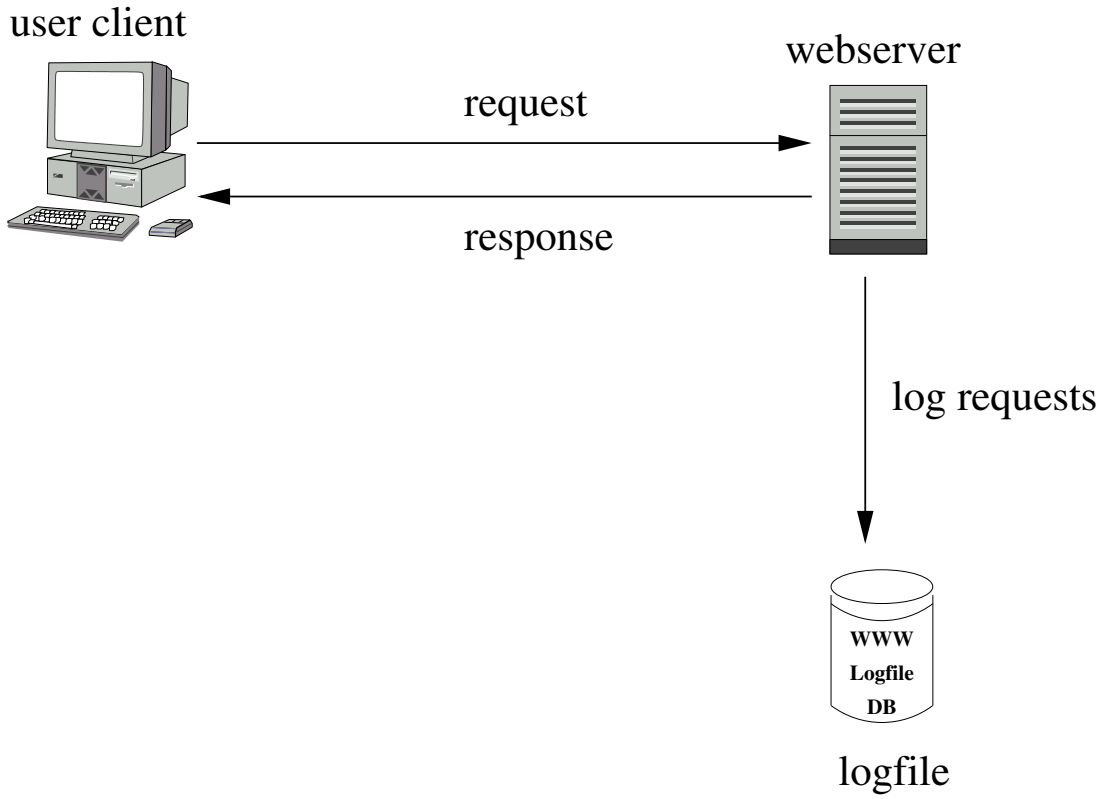


Abbildung 9: Protokollieren der Verwendung eines Webservers.

Business Events

Web Server-Logfile	Business Event-Logfile
GET /rec.jsp	<new-session id='10222' sid='15'/>
GET /rec.jsp?sid=15&q=1&type=2&price=3	<query id='10223' ref='10222' sid='15' nr='1'> <par name='type' value='2'/> <par name='price' value='3'/> </query> <show-products id='10224' ref='10223' from='0'> <p id='1014'/><p id='1143'/><p id='1216'/> <p id='1033'/><p id='1022'/> </show-products>
GET /view.jsp?sid=15&pid=1014	<detailview id='10225' ref='10224' pid='1014'/>
GET /view.jsp?sid=15&pid=1216	<detailview id='10226' ref='10224' pid='1216'/>
GET /rec.jsp?sid=15&q=1&type=2&price=3&fr=6	<show-products id='10227' ref='10223' from='5'> <p id='1221'/><p id='1045'/><p id='1176'/> </show-products>
GET /view.jsp?sid=15&pid=1045	<detailview id='10228' ref='10227' pid='1045'/>

Web Data Warehouse-Architektur

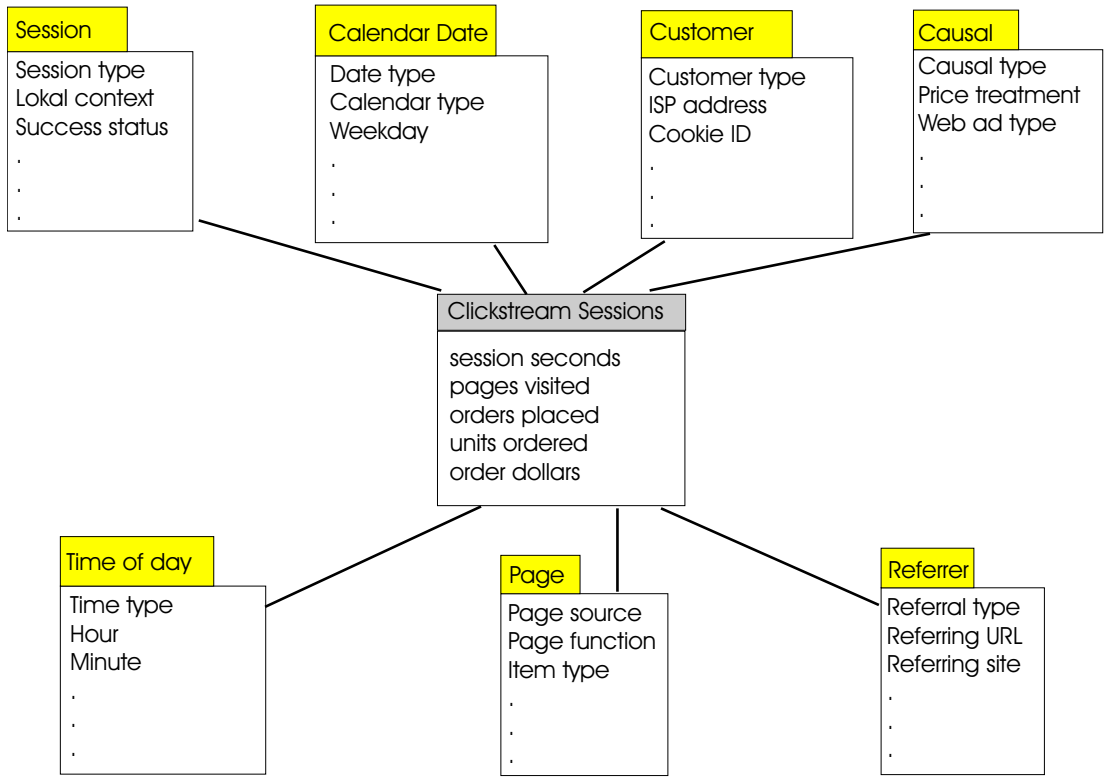
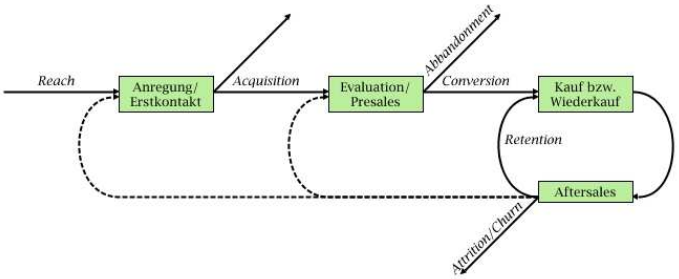


Abbildung 10: Web Data Warehouse-Architektur auf Seiten-Niveau [KM00].

e-Metrics

Kontaktaufnahme:

$$\text{total site reach} = \frac{\text{Zahl der Besucher}}{\text{Zahl der Internetnutzer}}$$



Sales:

$$\text{customer conversion rate} = \frac{\text{Zahl der Kunden}}{\text{Zahl der Besucher}}$$

$$\text{customer acquisition rate} = \frac{\text{Zahl der Kunden}}{\text{Zahl der click-through-Besucher}}$$

Presales:

$$\text{acquisition rate} = \frac{\text{Zahl der click-through-Besucher}}{\text{Zahl der Besucher}}$$

Aftersales:

$$\text{repeat customer conversion rate} = \frac{\text{Zahl der Wiederkäufer}}{\text{Zahl der Kunden}}$$

Analog Kostengrößen (Cost per Visitor/Click-through-Visitor/Customer/Repeat Customer).

Warenkorb-Analyse

Nr.	Warenkorb
1	A,C,D,K, M
2	A,C,D,L
3	A,B,C,D,H,K,L
4	A,B,C,D,I,K, L
5	A,B,C,E,F,G
6	A,B,C,D,G,K,L,M
7	B,C,E,I
8	A,B,C,D,F,I,K,M
9	A,B,C,D,K,L,M
10	A,B,C,D,F,L
⋮	⋮

Abbildung 11: Warenkörbe (flache Transaktionsdaten).

Muster (= häufige Teilwarenkörbe):

⋮

A,C,D : sup $= 0.8$

A,C,D,K : sup $= 0.6$

⋮

Assoziationen zwischen Mustern:

⋮

A,C,D $\rightarrow$ K : sup $= 0.6$, conf $= 0.75$

⋮

Navigationsmuster / Pfade

Muster und Assoziationsregeln lassen sich auch in reichhaltigeren Datenstrukturen effizient berechnen [ST03, GST01], z.B. für

- Navigationspfade,
- Warenkorb-Sequenzen,
- Warenkörbe mit Produkthierarchie,
- Warenkorb-Sequenzen mit Produkthierarchie,
- ...

„Pfadbruchstücke“ berücksichtigen z.B. die Reihenfolge von Seitenaufrufen.

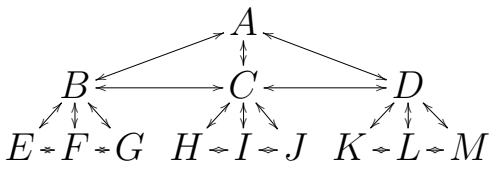


Abbildung 12: Sitegraph.

Nr.	Pfad	Kauf
1	ADK(D)C(D)M	nein
2	ACDKL	ja
3	ABCH(C)DLK	ja
4	AB(A)DLK(LD)CI	nein
5	ABGF(GB)C(B)E	nein
6	ADLK(L)M(LD)CBG	nein
7	CI(C)BE	ja
8	ABF(B)CIJ(IC)DK(D)M	ja
9	ABCDMLK	ja
10	ABF(BA)DKL(KD)CBF	nein

Abbildung 13: Pfade.

Beispiel

Kleines Beispiel:
Daten von Produkten, die während 3–4 aufeinanderfolgenden Besuchern von 12 Kunden aus einem Sortiment von 4 Produkten a,b,c und d gekauft wurden.

Erste Idee:
Wir suchen nach Produkten, die bei einem Besuch häufig zusammen gekauft wurden (häufige Warenkörbe).

$\Longrightarrow$ Alle Warenkörbe haben gleichen Support 3 !

	gekaufte Produkte beim			
Kunde	1. Besuch	2. Besuch	3. Besuch	4. Besuch
1	a,b	c	b,c	a,b,c
2	b	d	b,d	a,c,d
3	c	a,b	a,c	b,c,d
4	d	b	a,d	d
5	b,d	a,b,d	a,c,d	-
6	a,b,d	b,c	b,c,d	a,b,c,d
7	c	a,d	c,d	b,d
8	b	a,c	b,c	-
9	c,d	a,c	a,b	a
10	a,d	a,b,d	a,b,c,d	a
11	a,c,d	b,c,d	a	-
12	a,b,c	c,d	a,b,c	a,b,c,d

Zweite Idee:
Wir suchen nach häufigen Folgen gleicher Warenkörbe.
$\Longrightarrow$ Es gibt 3 Folgen der Länge 2 mit Support 2; alle anderen Folgen kommen höchstens einmal vor.

$\Longrightarrow$ Besitzen die Daten überhaupt irgendeine Struktur?

Kunde	gekaufte Produkte beim 1. Besuch	2. Besuch	3. Besuch	4. Besuch
1	5	3	8	11
2	2	4	9	13
3	3	5	6	14
4	4	2	7	4
5	9	12	13	-
6	12	8	14	15
7	3	7	10	9
8	2	6	8	-
9	10	6	5	1
10	7	12	15	1
11	13	14	1	-
12	11	10	11	15

Codierung	Warenkorb
1	a
2	b
3	c
4	d
5	a,b
6	a,c
7	a,d
8	b,c

Codierung	Warenkorb
9	b,d
10	c,d
11	a,b,c
12	a,b,d
13	a,c,d
14	b,c,d
15	a,b,c,d

Richtige Idee:
Wir suchen nach **Warenkorb-Sequenzen** (Substrukturen der Ordnung 2).
$\Longrightarrow$ Es gibt mehrere solcher Warenkorb-Sequenzen mit Support mindestens 3:

- {a,b}, {c}, {b,c}, {a,b,c}
- {d}, {b,d}, {a,c,d}
- {a,c}, {b,c}

$\Longrightarrow$ 75% der Daten werden durch diese Substrukturen beschrieben.

$\Longrightarrow$ 3 Kundensegmente konnten identifiziert werden.

Kunde	gekaufte Produkte beim 1. Besuch	2. Besuch	3. Besuch	4. Besuch
1	a,b	c	b,c	a,b,c
2	b	d	b,d	a,c,d
3	c	a,b	a,c	b,c,d
4	d	b	a,d	d
5	b,d	a,b,d	a,c,d	-
6	a,b,d	b,c	b,c,d	a,b,c,d
7	c	a,d	c,d	b,d
8	b	a,c	b,c	-
9	c,d	a,c	a,b	a
10	a,d	a,b,d	a,b,c,d	a
11	a,c,d	b,c,d	a	-
12	a,b,c	c,d	a,b,c	a,b,c,d

1. Business Intelligence, Data Warehousing und Data Mining
2. Web Usage Mining
3. Recommender-Systeme

Recommender-Systeme sind Online-Informationssysteme, die

- Kunden **Produkte empfehlen** (automatisches Verkaufen),
- im Gegensatz zu statischen Listen (Sonderangebote, editor's choice, etc.) gewöhnlich **personalisiert** und auf den individuellen Kunden ausgerichtet sind,
- **Kundenprofile** bestehend aus expliziten Produktbewertungen und impliziten Produktbewertungen verwenden, um Personalisierung zu erreichen,
- auch unter der Bezeichnung *collaborative filtering* bekannt sind.

Beispiel (1/3)

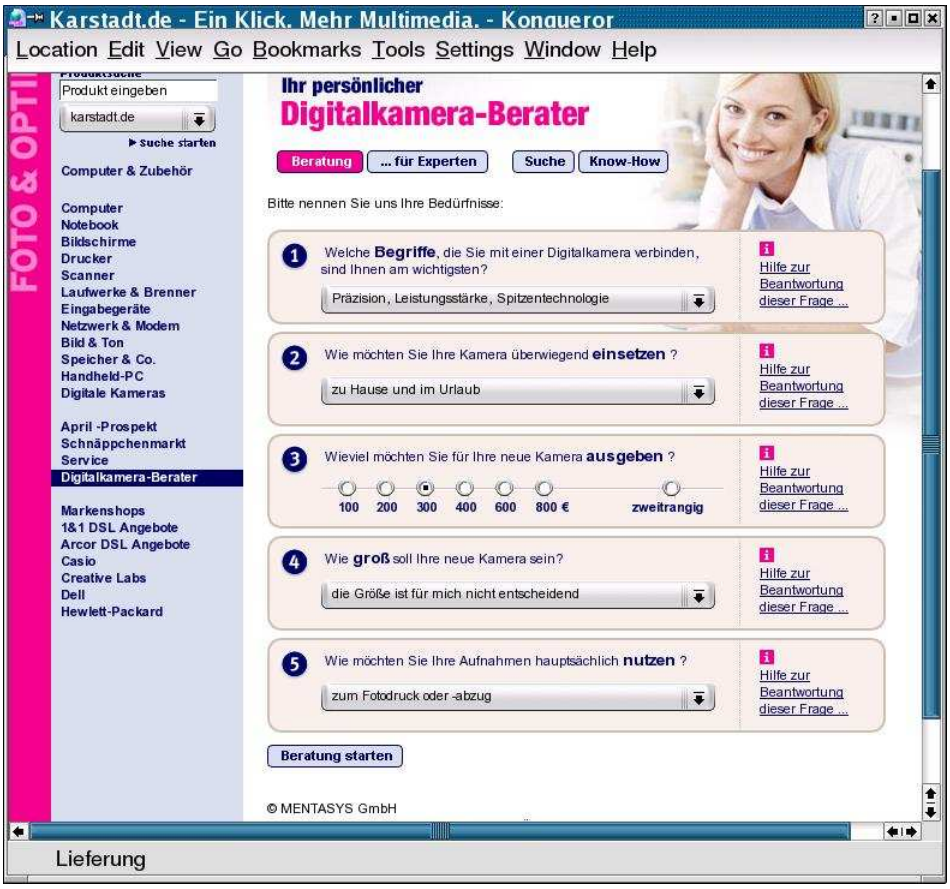


Abbildung 14: Anonymes Recommender-System: Aufgabenbeschreibung.

Beispiel (2/3)

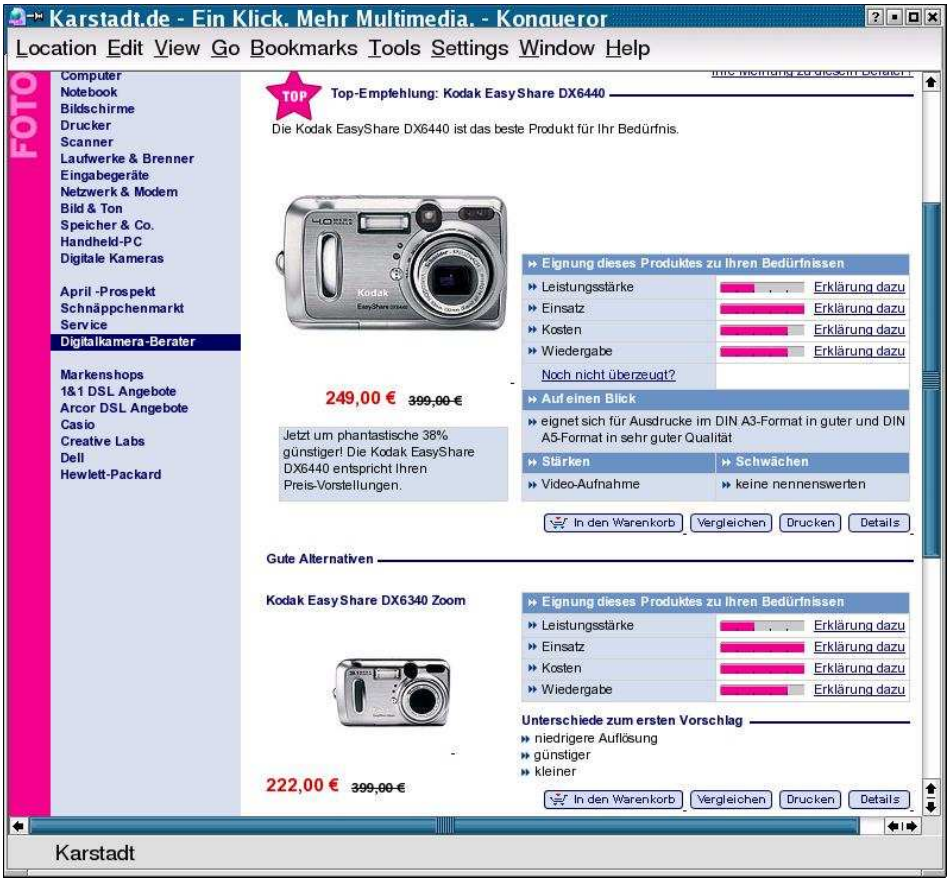


Abbildung 15: Anonymes Recommender-System: Vorschlagsliste.

Beispiel (3/3)

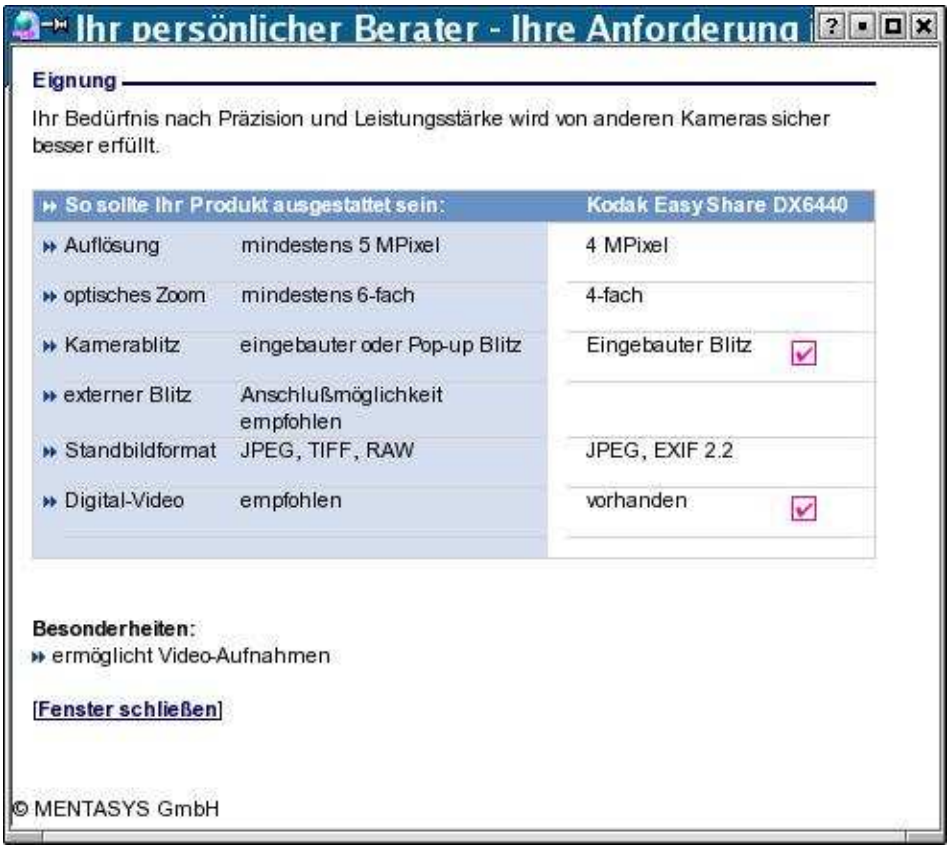
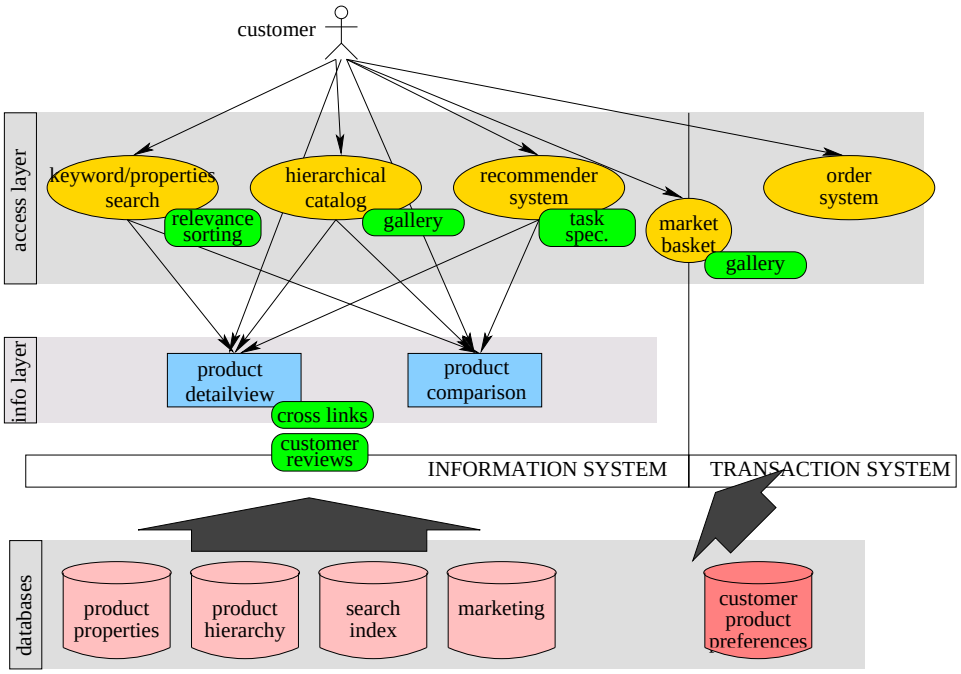
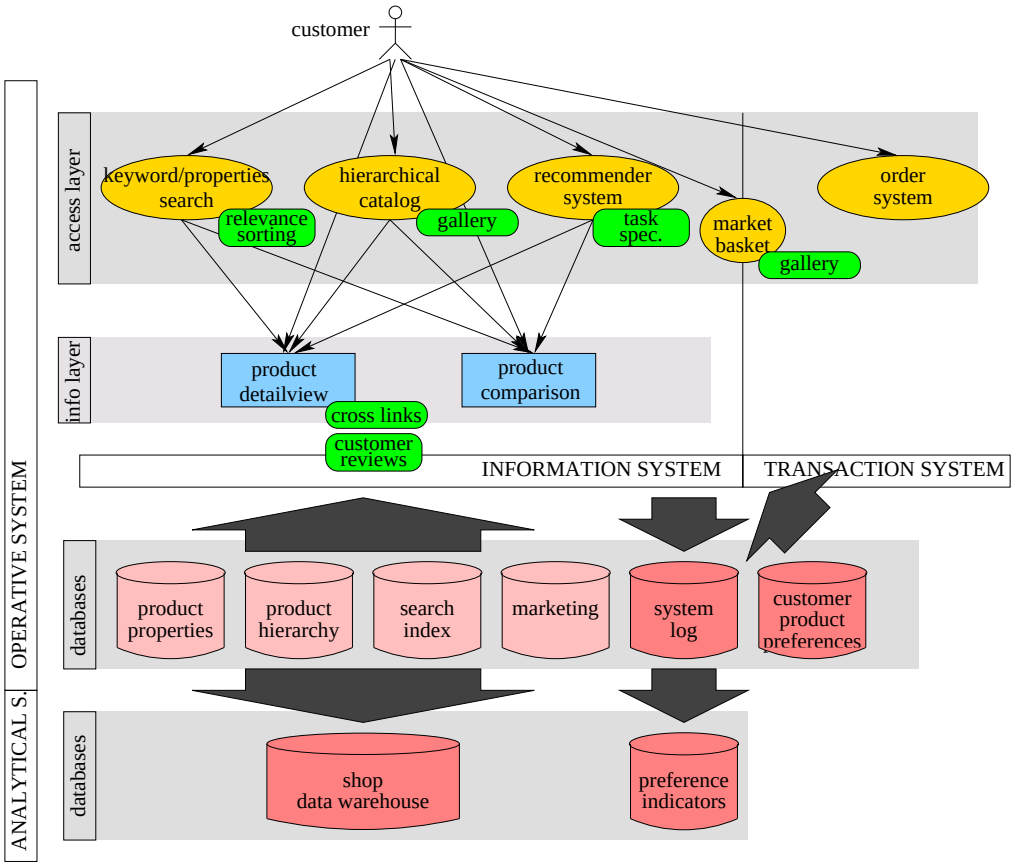


Abbildung 16: Anonymes Recommender-System: Vorschlags-Begründung.

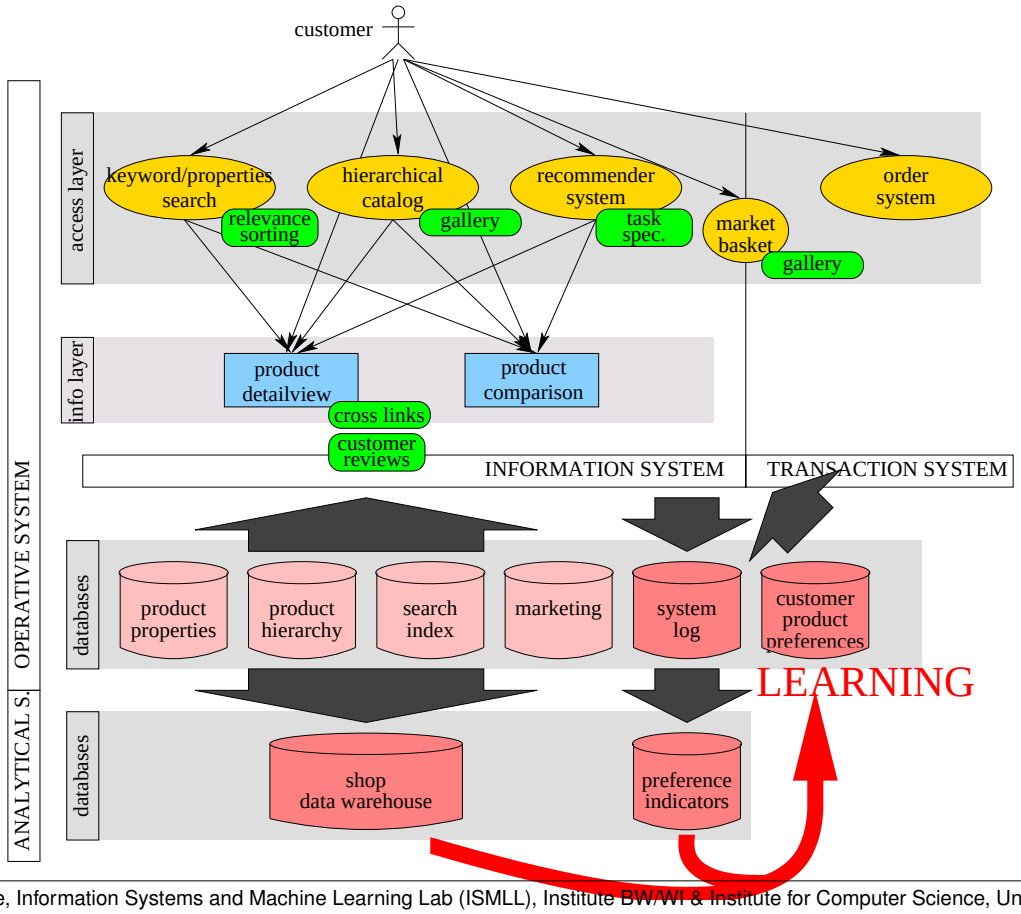
Operatives Shop-System



Analytisches System / Informations-Extraktion („was Kunden tun“)



Lernen von Kundenpräferenzen („was Kunden wollen“)



Predictive model specification

Models predicting **viewing** / **buying probabilities**:
a) model setup:

$$\mathcal{X} \longrightarrow [0,1]^P, \qquad \begin{array}{l}\mathcal{X} \text{ space of task specifications} \\ P \text{ set of products}\end{array}$$

b) training data (binary preference indicators):

x_1	x_2	$\ldots$	x_n		p_{1014}	p_{1015}	$\ldots$	p_{1243}
1	0	$\ldots$	1		1	0	$\ldots$	0
1	2	$\ldots$	0	$\longrightarrow$	0	1	$\ldots$	1
0	2	$\ldots$	2		0	0	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$

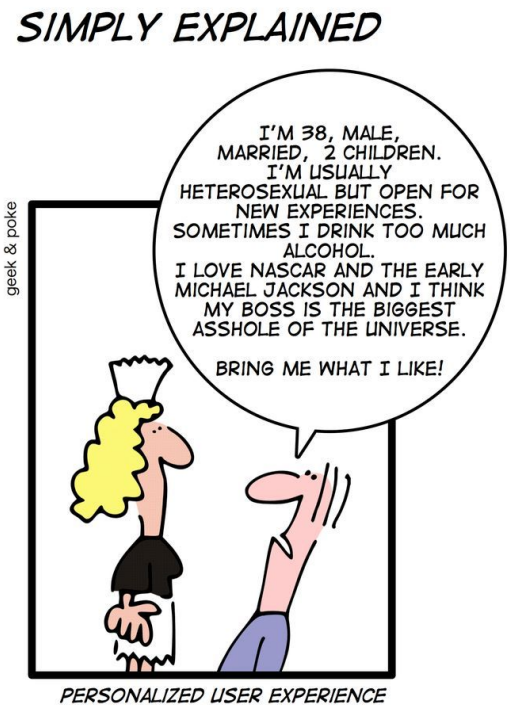
c) predictions:

x_1	x_2	$\ldots$	x_n		p_{1014}	p_{1015}	$\ldots$	p_{1243}
1	2	$\ldots$	0		0.010	0.009	$\ldots$	0.003
0	0	$\ldots$	0	$\longrightarrow$	0.007	0.003	$\ldots$	0.011
1	0	$\ldots$	1		0.002	0.005	$\ldots$	0.007
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$

Predictive model evaluation

model	mindev	nodes	$\text{wrec}_{\text{train}}$	$\text{wrec}_{\text{test}}$
random	–	–	0.092	0.097
static	–	1	0.460	0.437
set of trees	0.005	6.0	0.575	0.469
	0.003	12.1	0.605	0.480
	0.002	20.9	0.632	0.478
	0.001	47.6	0.693	0.460
single tree	0.01	5	0.506	0.447
	0.005	15	0.543	0.461
	0.002	47	0.586	0.486
	0.0015	71	0.601	0.495
	0.001	173	0.643	0.474
random forest	–	–	0.828	0.465

Alternative Explanation



Zusammenfassung

1. Business Intelligence-Systeme integrieren die Informationen eines Unternehmens und stellen damit eine wichtige Quelle für Entscheider dar, auch für detaillierte Protokolldaten (**Web Data Warehouse, e-Metrics**).

2. Diese Informationen können aber auch tiefer analysiert werden,
   - um verdichtere Reports zu bekommen (**Kundensegmente**, **Kaufverhalten**) oder
   - um damit operative Systeme zu betreiben (**Recommender-Systeme**).

3. Die Analyse großer und komplexer Datenbestände wird durch den Einsatz von Methoden des **maschinellen Lernens** / **Data Minings** ermöglicht.

4. Maschinelles Lernen kann für viele BI-Fragestellungen eingesetzt werden, aber auch in vielen anderen Bereichen (ingenieurwiss., medizininformatische Anwendungen etc.).

Zum Lesen

Ergänzend zum Lesen:

Kenneth C. Laudon, Jane P. Laudon, Detlef Schoder (^{2}2009): *Wirtschaftsinformatik — Eine Einführung,* Kapitel 11 „Entscheidungsunterstützung“.

Literatur

[CCK+00] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. Crisp-dm 1.0 step-by-step data mining guide, 2000.

[CD97] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and olap technology. *SIGMOD Record*, 26(1):65–74, 1997.

[DG02] Carsten Dittmar and Peter Gluchowski. Synergiepotenziale und herausforderungen von knowledge management und business intelligence. In Uwe Hanning, editor, *Knowledge Managment und Business Intelligence*, pages 27–41. Springer, 2002.

[FPSS96] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–29. AAAI Press / MIT Press, 1996.

[Gro04] Garnter Group. Glossary, term business intelligence, 2004.

[GST01] Wolfgang Gaul and Lars Schmidt-Thieme. Mining generalized association rules for sequential and path data. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM), San Jose*, pages 593–596, 2001.

[Inm92] William H. Inmon. *Data architecture: the information paradigm*. QED, Boston, 2nd edition, 1992.

[KM00] Ralph Kimball and Richard Merz. *The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse*. John Wiley & Sons, 2000.

[Res06] Ventana Research. Open source bi survey results, 2006.

[RST06] Steffen Rendle and Lars Schmidt-Thieme. Object identification with constraints. In *Proceedings of 6th IEEE International Conference on Data Mining (ICDM) 2006, Hong Kong*, 2006.

[ST03] Lars Schmidt-Thieme. *Assoziationsregel-Algorithmen für Daten mit komplexer Struktur – mit Anwendungen im Web Mining*. Frankfurt a. M., 2003.