KSFE 2002

INFORMATION WORKS®

Unternehmensberatung & Informationssysteme

Web Mining bei der VICTORIA Versicherung

Denise Petrat

Information Works

sas e-Intelligence

VICTORIA

INFORMATION WORKS
Unternehmensberatung & Informationssysteme

Agenda

- Web Mining bei der VICTORIA
- Verfahren
- Auswertung
- Fazit

INFORMATION WORKS
Unternehmensberatung & Informationssysteme

Agenda

- Web Mining bei der VICTORIA
- Verfahren
- Auswertung
- Fazit

Ausgangssituation: e-Business

Die VICTORIA Versicherung präsentiert sich ihren Kunden im Internet.

Ziele:

- Vorstellung des Unternehmens, sowie der Produkte und Angebote.
- Umfassende Informationen und Serviceleistungen, um neue Kunden zu gewinnen und Alte an sich zu binden.
- Wettbewerbsfähigkeit im Internetzeitalter.

Die **Kontrolle dieser Ziele** ist für eine effektive Gestaltung der Internetseite sehr wichtig.

➔ **Web Mining der VICTORIA Homepage!**

Ziele von Web Mining

- **Strukturelle Optimierung** der Web Seite (Navigationspfade überprüfen, neue Links setzen, ...)
- Kennenlernen der eigenen Besucher (**Usersegmente**, bes. Interessen, ...)
- **Kundenbindung** erhöhen durch gezielte Ansprache verschiedener Usersegmente
- Erkennen von **Kunden unter den Internetusern** und Betreuung anpassen (Abwanderungstendenzen oder Interesse an neuen Versicherungen erkennen)
- **Potentielle Neukunden** im Internet erkennen und gezielt ansprechen
- **Technische Anforderungen** an die Seite überprüfen

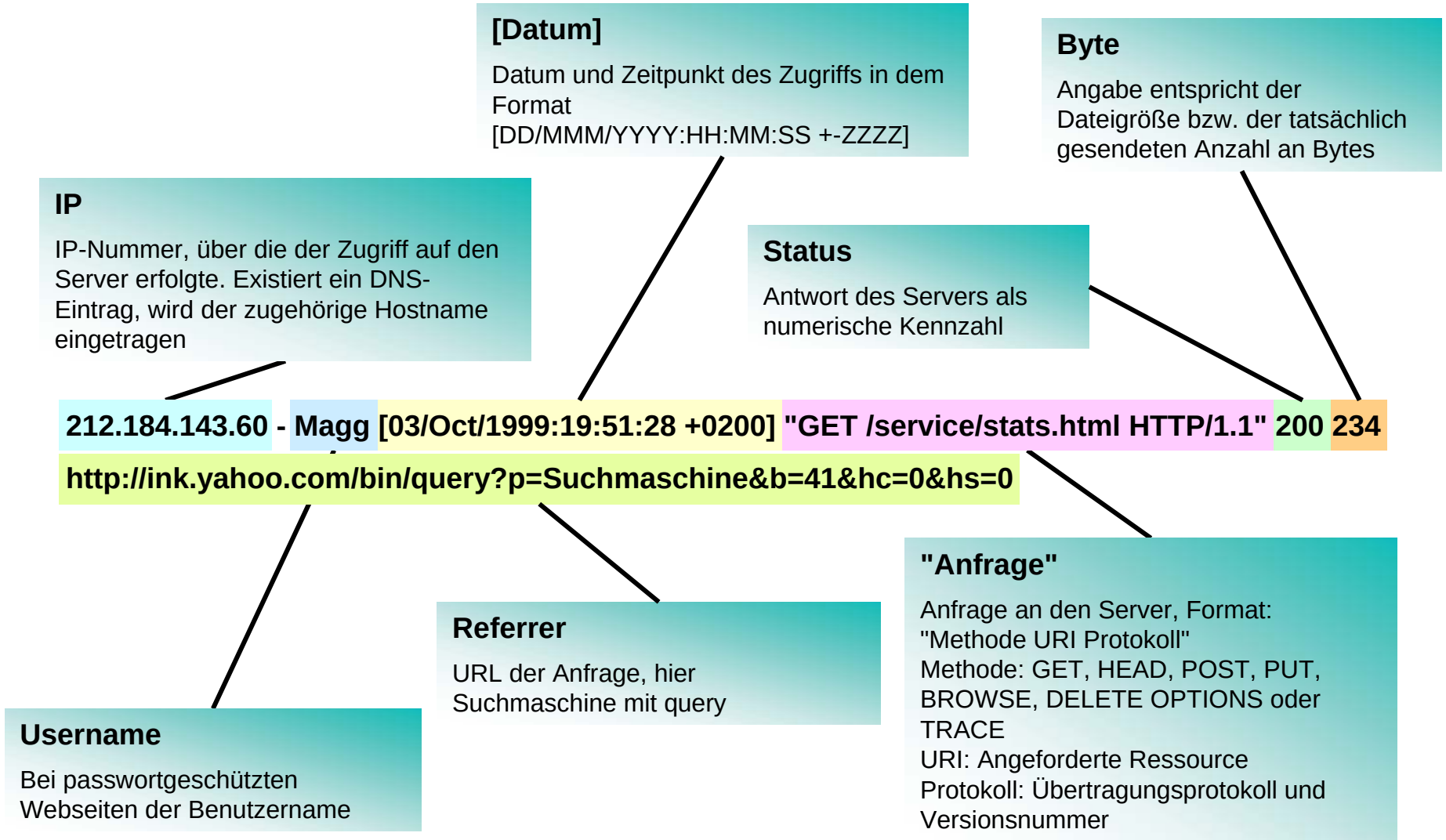
Web Mining bei der VICTORIA

Datenbeschreibung:

- Log Files der Victoria vom 17. Januar bis 30. April 2001 (104 Tage),
- Ca. 10 bis 40 MB Rohdaten pro Tag, insg. ca. 8.56 GB
- 705645 Clicks
- 107365 Besuche, dh ca. 1000 pro Tag (6054 Kontakte, 141 Online-Abschlüsse)

```
date time c-ip cs-username s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query sc-status sc-win32-status sc-bytes cs-bytes time-taken s-port cs-version cs(User-Agent) cs(Cookie) cs(Referer)

2001-01-01 00:52:09 62.54.240.112 - W3SVC9 ADA-ARC026 192.168.231.21 GET /images/victoria_logo_1.gif - 304 0 142 457 0 80 HTTP/1.1
Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+98) SITESERVER=ID=4770837c7abd7b83d3b72a467b5f2601http://www.victoria.de/menu/logo.asp

2001-01-01 00:52:09 62.54.240.112 - W3SVC9 ADA-ARC026 192.168.231.21 GET /home/home.asp - 200 0 8960 359 469 80 HTTP/1.1
Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+98) SITESERVER=ID=4770837c7abd7b83d3b72a467b5f2601http://www.victoria.de/home/frameset.asp

2001-01-01 00:52:11 62.54.240.112 - W3SVC9 ADA-ARC026 192.168.231.21 GET /images/victoria_logo_2.gif - 304 0 142 457 0 80 HTTP/1.1
Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+98) SITESERVER=ID=4770837c7abd7b83d3b72a467b5f2601http://www.victoria.de/menu/logo.asp

2001-01-01 00:52:11 62.54.240.112 - W3SVC9 ADA-ARC026 192.168.231.21 GET /images/victoria_logo_3.gif - 304 0 142 457 0 80 HTTP/1.1
Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+98) SITESERVER=ID=4770837c7abd7b83d3b72a467b5f2601http://www.victoria.de/menu/logo.asp
```

Die Elemente einer Logfile-Zeile



Vorbereitende Maßnahmen

- **Seiten Identifikation**: im Log File wird das Laden jedes einzelnen Objektes dokumentiert, deshalb müssen die echten Contentpages anhand der Dateinamenerweiterung identifiziert werden (hier: .asp)
- **Benutzer Identifikation**: jeder Benutzer hat eine IP-Adresse an der sein Weg durch die Internetseite verfolgt werden kann, diese ist aber nicht eindeutig (Proxy-Server), deshalb werden zusätzlich Cookies vergeben, die eine eindeutige Identifizierung ermöglichen
- **Variablenkreation**: zur mathematischen Handhabbarkeit müssen einige statistische Kenngrößen und neue Variablen kreiert werden: z. B. Verweildauern, Anzahl an Clicks, gesendeten Bytes usw., Tageszeit in Kategorien, Seiten mit den meisten Fehlern, ...

Vorbereitende Maßnahmen

- **Seitenkategorisierung**: die ca. 2000 Seiten der VICTORIA Seite gehören zu verschiedenen Kategorien, zur effektiveren Analyse werden die Seiten an Hand des Menüs in Themenbereiche (z. B. ‚Kranken', ‚Leben', ‚Nähe', ‚Leistungen', ‚Berechnungen', ‚Mail' etc.) unterteilt.
- **Definition der Zielvariable**: der Schwerpunkt der Analyse richtet sich auf eine vorher bestimmte Zielvariable „Kontakt", die Sessions kennzeichnet, in denen eine Personalizierung des Internetusers nötig ist. (eMail, Vertragsänderungen, Online-Schadensmeldung, Online-Abschluß, ...)

Agenda

- Web Mining bei der VICTORIA
- Verfahren
- Auswertung
- Fazit

Verwendete Verfahren

Der Analyse mit dem SAS Enterprise Miner liegen verschiedene Verfahren zu Grunde:

- Visuelle Methoden (Histogramme, ...)
- **Clusterungen mit Selforganizing Maps**
- Sequenzanalysen
- Logistische Regression
- **Multilayer Perzeptron**
- Entscheidungsbäume

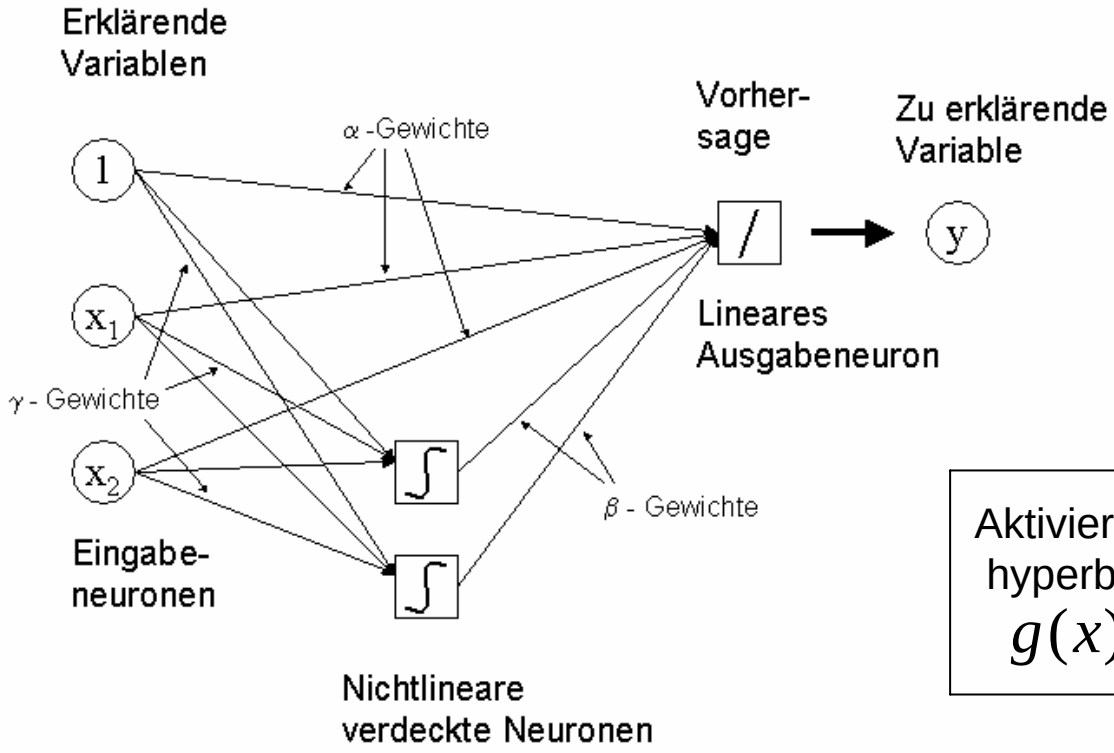
Neuronale Netze

Idee: Modelle der Gehirnfunktion

*„Sie versuchen in **Struktur und Funktionsweise Gehirnzellkomplexe** nachzubilden und dadurch eine tragfähige **Simulation komplexer menschlicher Denkvorgänge** zu erzielen. Sie sind **informationsverarbeitende Systeme,** die sich aus **primitiven, uniformen, miteinander kommunizierenden Verarbeitungseinheiten** in großer Zahl zusammensetzen.“* (Kratzer, 1990)

Leistung: **Musterassoziation und –rekonstruktion** durch adaptives ‚Lernen‘. Durch die **Präsentation von Eingangsmustern** in einem Trainingslauf erhält das neuronale Netz sein ‚Wissen‘.

Das Multilayer Perzeptrons (MLP)



Aktivierungsfunktion: z.B. hyperbolischer Tangens

$$g(x) = tanh\ (x)$$

Funktionelle Form:

$$f(X,w) = X\alpha + \sum_{h=1}^{H} \beta_h g_h \left(\sum_{i=0}^{I} \gamma_{hi} x_i \right)$$

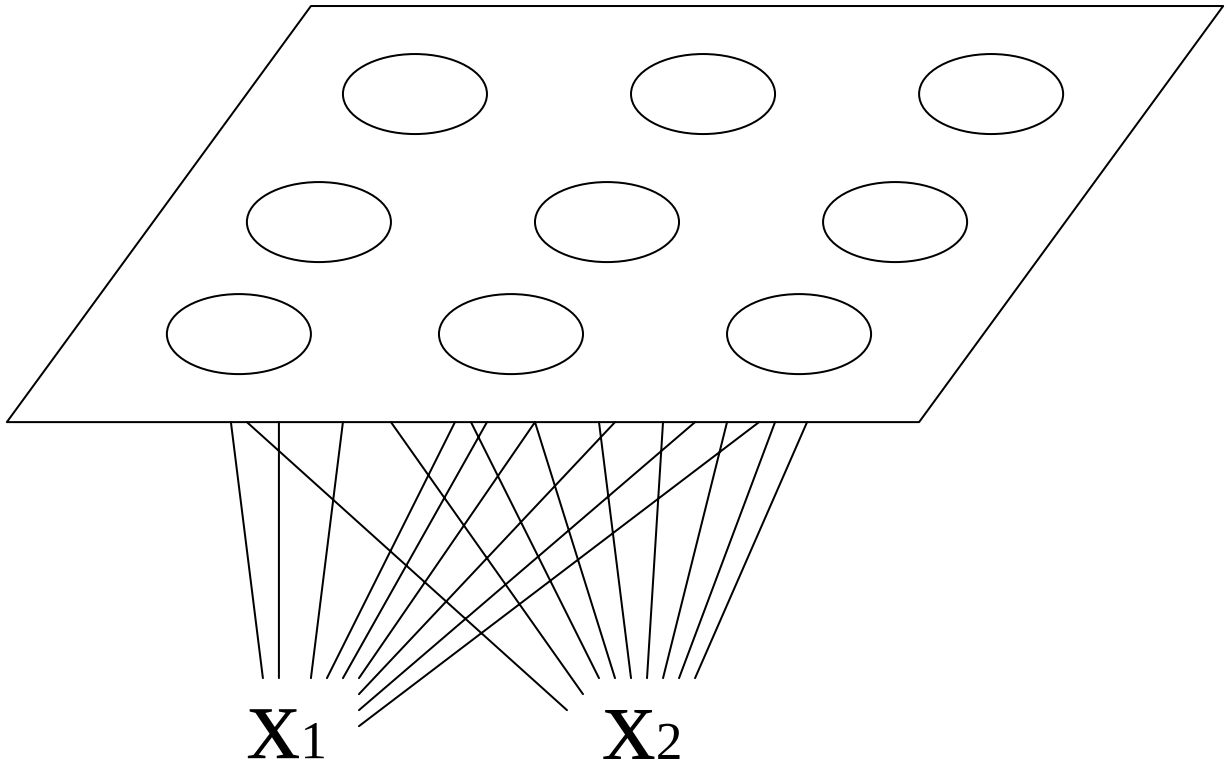
Selforganizing Maps (SOM)

*„**SOMs dienen der Clusterung, Visualisierung und Abstraktion von Daten** und sind nach dem Vorbild entstanden, wie verschiedenen sensorische Eindrücke im menschlichen Gehirn verarbeitet werden.“*
(Kohonen, 1995)

Es handelt sich um neuronale Netz, die ein **topologisches Mapping von Input Daten** in bestimmte Cluster durchführen, dabei werden (räumliche) **Beziehungen unter den Reizen** durch **räumliche Beziehungen unter den Neuronen** ausgedrückt.

Es entstehen **zweidimensionale Gitter**, die auf Abstandsbildung basierenden Algorithmen entwickelt werden.

Schematische Darstellung von Selforganizing Maps zur Kundensegmentierung



INFORMATION WORKS
Unternehmensberatung & Informationssysteme

Agenda

- Web Mining bei der VICTORIA
- Verfahren
- Auswertung
- Fazit

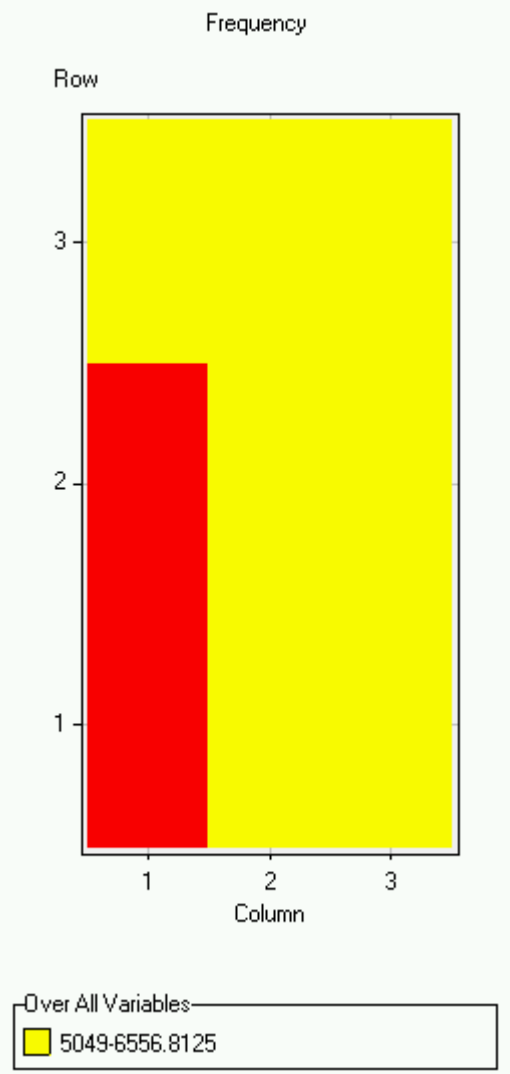
TOP 10 Seiten

Seite	Häufigkeit	Prozent
/home/home.asp	165402	23.44
/home/http.asp	37289	5.28
/berechnungen/kv/easy/go.asp	30067	4.26
/berechnungen/lv/bu/erwerb/ergebnis.asp	24043	3.41
/berechnungen/lv/bu/erwerb/default.asp	22213	3.15
/stellen/Default.asp	19464	2.76
/stellen/stellendetail.asp	16981	2.41
/naehe/betr-homepage.asp	16744	2.37
/berechnungen/lv/rente/privat/ergebnis.asp	15885	2.25
/naehe/Default.asp	13978	1.98

- Ca. ¼ aller Clicks werden von der **‚Home‘** Seite erzeugt.
- Die **Fehlerseite** ‚/home/http.asp‘ ist mit ca. 5% die 2. häufigste Seite.
- **Tarifrechner und Stellen-anzeigen** sind allgemein beliebt und werden häufig frequentiert.
- Informationen aus den Bereichen **‚Leben‘, ‚Kranken‘,** und **‚Nähe‘** werden von vielen Usern genutzt.

- **Kontakte** entstehen meist durch das **Schreiben von Infomails**, ansonsten werden auch Vertragsänderungen, Anregungen und Schäden online versendet.
- **Ausstiegsseiten und Seiten, die Fehler produzieren, unterscheiden sich nicht wesentlich von den TOP 10 Seiten**.

Sessionsegmentierung



Die Segmentierung mit Hilfe der SOMs ergibt 9 Cluster:

1. **Verirrte**: 1-Click-Sessions (25%)
2. **Hektiker**: kurze, schnelle Sessions
3. **Frustrierte**: kurze Sessions mit vielen Fehlern
4. **Jobsucher**: viele Clicks im Bereich ‚Stellen'
5. **Interessierte**: genaue Infos über div. Leistungen
6. **Rechner**: viele Tarifrechnernutzungen (Leben, Kranken, Unfall)
7. **Aktionäre**: genaue Infos über das Unternehmen
8. **Hartnäckige**: sehr lange Besuche mit div. Infos
9. **Kunden**: Sessions, die zu einem Kontakt führen

Clickstreamanalyse

Weiteres Analyseziel: **Identifikation typischer Navigationspfade**

- für die Benutzung der Bereiche
- für einzelne Seiten
- Vergleich Kontaktsessions – alle Sessions

	Ketten Länge	Support(%)	Confidence(%)	Anzahl der Transaktionen	Regel
1	2	36.45	56.08	1836	home ==> k_mail
2	2	16.44	25.29	828	home ==> lebensl
3	2	14.99	23.06	755	home ==> sc_vert
4	2	14.65	22.54	738	home ==> naehe
5	2	14.02	21.56	706	home ==> home
6	2	12.88	19.82	649	home ==> aktuell
7	2	12.31	18.94	620	home ==> suchen
8	2	10.28	55.70	518	naehe ==> k_mail
9	2	10.13	43.66	510	lebensl ==> k_mail
10	2	9.49	14.60	478	home ==> l_krank
11	2	8.44	79.29	425	l_krank ==> k_mail
12	2	8.32	15.26	419	k_mail ==> k_mail
13	2	8.16	12.55	411	home ==> l_haftp
14	2	8.04	57.86	405	suchen ==> k_mail
15	2	7.78	11.97	392	home ==> b_krank

Clickstreamanalyse

Ergebnisse: **Kontaktsessions**

- Kontaktsessions sind **homogener**: haben starken Support für identifizierte Navigationspfade.
- Typischerweise wird der direkte **ziel-strebige** Weg von **‚Home' zu ‚Mail'** gewählt.
- Ansonsten gelangen User von ‚Nähe', ‚Lebenslagen', ‚Leistungen Kranken', ‚Berechnungen Kranken', ‚Suchen' und ‚Service Center Vertrag' auf die ‚Mail' Seite.
- Lange **Clickstreams enden** immer wieder im Bereich **‚Mail'.**
- In ¾ aller Fälle wird nach Betreten des Bereichs ‚Mail' auch eine Mail versendet.

Ergebnisse: **alle Sessions**

- Typische Wege gehen jeweils von **‚Home'** aus.
- Lange Clickstreams enden meist auf **Tarifrechnern, Stellenanzeigen oder Fehlermeldungen.**
- Es gibt **wenige Wechsel** zwischen verschiedenen Kategorien.
- Viele Clickstreams folgen dem Seitenaufbau.

Clickstreamanalyse

2 Auffälligkeiten:

- Häufiges **hin-und-her-Clicken** zwischen den Bereichen **‚Leistungen Kranken‘,,Berechnungen Kranken‘** und **‚Suche‘.**

➔ Hinweis auf fehlende Infos auf der Berechnungsseite?

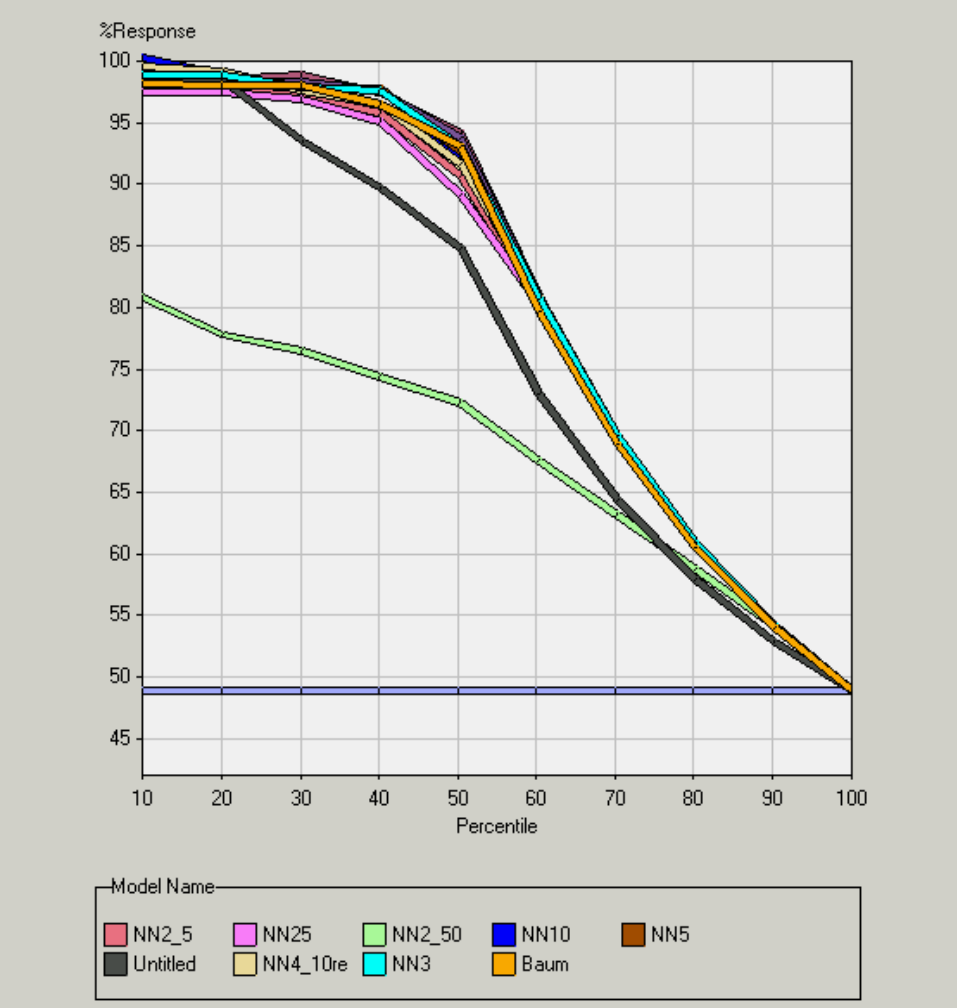
- Nach dem Betreten der Seite **‚/nähe/betr-homepage.asp‘** wird sehr häufig eine **Anregungsmail** versendet.

➔ Hinweis auf Verbesserungsmöglichkeiten in der Struktur und neue Links?

Identifikation der Haupteinflussvariablen auf die Zielvariable „Kontakt“

Modellierung mit Hilfe von

- logistischer Regression,
- neuronalen Netzen,
- Entscheidungsbäumen.



Auf Grund von **starken linearen Zusammenhängen** ergeben die logistische Regression und ein einfaches multilayer Perzeptron (MLP) mit nur 5 Neuronen die besten Ergebnisse:

Logistische Regression:

- Es werden 29 Einflussvariablen als signifikant bestimmt.
- **Die 3 wichtigsten sind ‚Anzahl der gesendeten Bytes', ‚Verweildauer' und ‚Berechnungen Leben'.**
- Die beiden erstgenannten haben einen positiven Einfluss auf die Kontaktaufnahme, die Letztgenannte wirkt sich negativ aus.

MLP:

- Stellt **gutes Prognosemodell** dar.
- ABER: Ergebnisse sind **schlecht interpretierbar**.
- Die Gewichte für die verdeckten Neuronen sind sehr niedrig, direkte Verbindungen zur Ausgabeschicht besitzen hohe Gewichte ➔ linearer Zusammenhang.

INFORMATION WORKS
Unternehmensberatung & Informationssysteme

Agenda

- Web Mining bei der VICTORIA
- Verfahren
- Auswertung
- Fazit

Fazit

Diese Log File Analyse zeigt beispielhaft verschiedene Möglichkeiten und Fragestellungen eines Web Mining Projektes.

Zentrale Bereiche einer **Web Analyse:**

- **Kundensegmentierung,**
- **Clickstreamanalyse,**
- **Identifikation von Einflussvariablen.**

Bis jetzt standen für die Analyse weder Stammdaten der VICTORIA-Datenbank noch Informationen aus den gesendeten Formularen und Emails zur Verfügung.

➔ **Chance: Identifizierung** von **VICTORIA-Kunden** und **Neukunden** unter den Internetusern durch diese zusätzlichen Daten und damit **Zugewinn wertvoller Informationen über die Identität der Users!**

Vielen Dank für Ihre Aufmerksamkeit!!!