Introduction

Basics

Data Preparation Process

Preprocessing

Knowledge Discovery, Pattern Analysis

Apriori-Algorithmus

Behavior Models

Developer's Model

Users' Model

Transaction Identification

WEB MINER SYSTEM

Conclusion

Seminar Intelligente Anwendungen im Internet

Data Preparation for Mining World Wide Web Browsing Patterns

Robert Cooley Bamshad Mobasher and Jaideep Srivastava (1999)

University of Minnesota

Georg Neugebauer

18.12.2007

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Übersicht

1. Introduction
2. Basics
3. Data Preparation Process
   - Preprocessing
   - Knowledge Discovery, Pattern Analysis
4. Apriori-Algorithmus
5. Behavior Models
   - Developer's Model
   - Users' Model
6. Transaction Identification
7. WEB MINER SYSTEM
8. Conclusion

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Übersicht

1 Introduction
2 Basics
3 Data Preparation Process
- Preprocessing
- Knowledge Discovery, Pattern Analysis

4 Apriori-Algorithmus
5 Behavior Models
- Developer's Model
- Users' Model

6 Transaction Identification
7 WEB MINER SYSTEM
8 Conclusion

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Web Usage Mining

Definition & Ziele

Web Usage Mining beschäftigt sich mit der Analyse, wie eine Webseite benutzt wird.

- Topology verbessern
- Clickverhalten analysieren

Details

- Analyse der Log-Dateien von Web-Servern
- Häufigkeit von Zugriffen
- Typische Pfade
- association rule generation
- sequential pattern generation, clustering

Web Content Mining

Beschäftigt sich mit den Inhalten einer Webseite.

Frequent set mining

Frequent Sets spielen eine wichtige Rolle beim Data Mining

Frequent Set Mining Problem

Erstes Problem: Analyse von supermarket transaction data. Welche Items werden zusammen gekauft? Häufigkeit?

- Transaktion hier: Menge von Items, die gekauft wurden bei einem Einkauf
- Support von einem Item: Anteil des Items in Transaktionen (genaue Def. folgt)
- set ist **frequent**, falls $support > threshold$

Finde alle Sets deren Support größer als threshold ist.

Search Space

Die Größe des Suchraumes ist $2^{|I|}$. Bei großem I scheitert der naive Ansatz alle Itemmengen zu bestimmen.

Introduction

Basics

Data Preparation Process

Preprocessing

Knowledge Discovery, Pattern Analysis

Apriori-Algorithmus

Behavior Models

Developer's Model

Users' Model

Transaction Identification

WEB MINER SYSTEM

Conclusion

Übersicht

1 Introduction

2 Basics

3 Data Preparation Process
- Preprocessing
- Knowledge Discovery, Pattern Analysis

4 Apriori-Algorithmus

5 Behavior Models
- Developer's Model
- Users' Model

6 Transaction Identification

7 WEB MINER SYSTEM

8 Conclusion

Transaktion

Eine Transaktion t kann als Triple definiert werden:

$$t =< ip_t, uid_t, \{(l_1^t.url, l_1^t.time), ..., (l_m^t.url, l_m^t.time)\} >$$
$$for\ 1 \leq k \leq m, l_k^t \in L, l_k^t.ip = ip_t, l_k^t.uid = uid_t$$

mit L: Set von User Session File Entries mit IP-Adresse, Client-Userid, betretene URL und Verweildauer.

Support

Sei Datenbasis D gegeben mit $D = t_1, t_2, ...$ Menge von Transaktionen und X,Y sind Itemmengen.

$$support(X) = \frac{\{t \in D | X \subseteq t\}}{|D|} \quad support(X \rightarrow Y) = support(X \cup Y)$$

Confidence

Sei eine Assoziationsregel gegeben mit $X \rightarrow Y$ und X,Y sind Itemmengen.

$$confidence(X \rightarrow Y) = \frac{support(X \rightarrow Y)}{support(X)}$$

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Übersicht

1 Introduction

2 Basics

3 Data Preparation Process
- Preprocessing
- Knowledge Discovery, Pattern Analysis

4 Apriori-Algorithmus

5 Behavior Models
- Developer's Model
- Users' Model

6 Transaction Identification

7 WEB MINER SYSTEM

8 Conclusion

„Web Usage Mining“ Prozess besteht aus drei elementaren Schritten.

- Preprocessing
- Knowledge Discovery
- Pattern Analysis

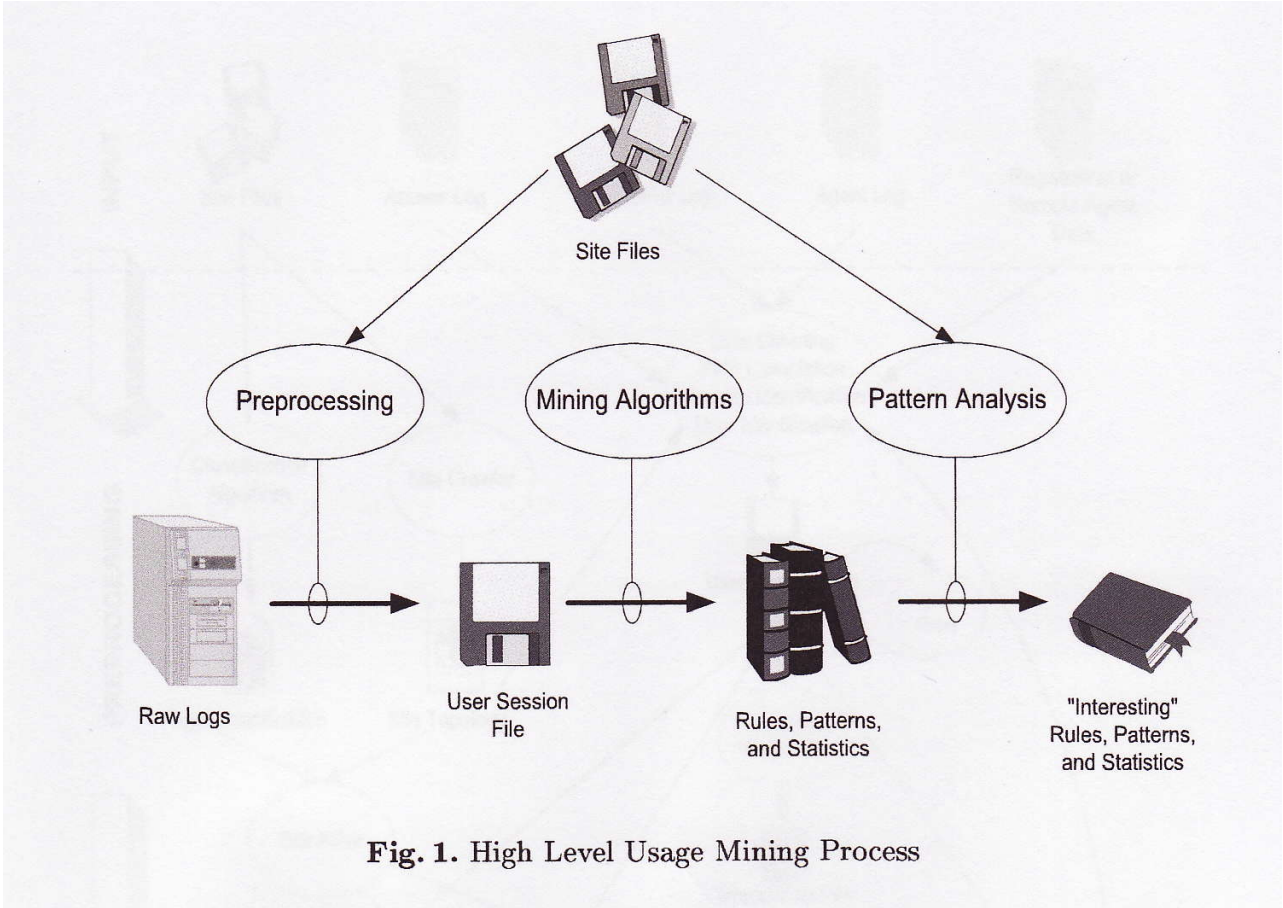


Fig. 1. High Level Usage Mining Process

Probleme

Inputgewinn ist oft nur erschwert möglich! Siehe folgendes Logfile:

- Cookies oder Cache Busting kann von User deaktiviert werden.
- Proxy Server
- Privatspähre

Introduction

Basics

Data Preparation Process

Preprocessing

Knowledge Discovery, Pattern Analysis

Apriori-Algorithmus

Behavior Models

Developer's Model

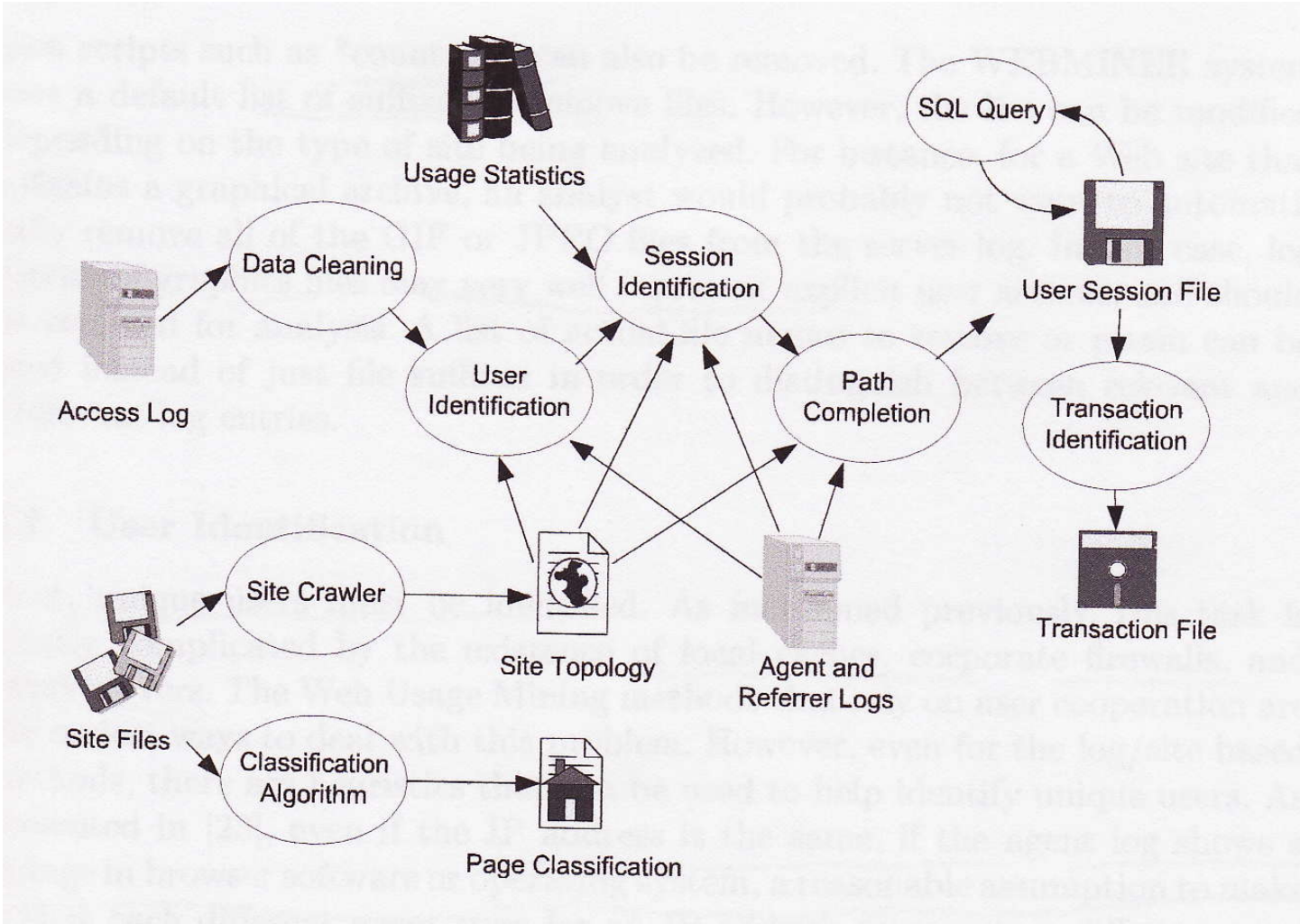
Users' Model

Transaction Identification

WEB MINER SYSTEM

Conclusion

Details des Web Usage Mining Prozess



Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Preprocessing besteht aus folgenden Schritten:

- data cleaning
- user identification
- session identification
- path completion
- formatting
- site topology, page classification, site filter (behandelt in Kapitel 4)
- transaction identification
  (separat behandelt in Kapitel 5)

Ziel: Erzeuge ein User Session File, welches als Input für die Knowledge Discovery Phase dient.

Data Cleaning

Entferne irrelevante Daten aus der Server-Log Datei.
Ziel: Log-Datei spiegelt die User-Zugriffe der Webseite wieder.

- Eliminiere anhand des Suffix der URL: gif, jpeg, GIF, JPEG, jpg, JPG.
- Eliminiere bekannte Scripts: z. B. „count.cgi“
- Erstelle Liste von Files, die entfernt werden sollen.

Aber: Falls Webseite viele grafische Inhalte hat $\Rightarrow$ Eliminierung dieser Inhalte nicht sinnvoll!

User Identification

Eindeutige Erkennung von Usern. Dies wird erschwert durch:

- lokale Caches
- Firewalls
- Proxy Servers

Heuristiken werden benutzt um User eindeutig zu identifizieren. Anzeichen für einen neuen User sind:

- Wechsel der Browser Software
- Wechsel des Betriebssystems
- Page Request, der nicht mit einem Link von der aktuellen Seite erreicht werden kann

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Session Identification

Ziel: Splitte die Zugriffe eines Users in individuelle Sessions.

Timeout

Definiere einen Zeitwert, wo der Split stattfindet. (Default: 30 min)

Path Completion

Ziel: Finde wichtige Zugriffe, die nicht in der Server-Log Datei auftauchen.

Backtracking mittels Back-Button

- Cached Version der Webseite
- In Server-Log oder Site Topology nachschauen, ob Seite schon vorhanden und gegebenenfalls Hinzufügen
- Verweildauer für Seite festsetzen

Introduction

Basics

Data Preparation Process

Preprocessing

Knowledge Discovery, Pattern Analysis

Apriori-Algorithmus

Behavior Models

Developer's Model

Users' Model

Transaction Identification

WEB MINER SYSTEM

Conclusion

User, Session Identification und Path Completion

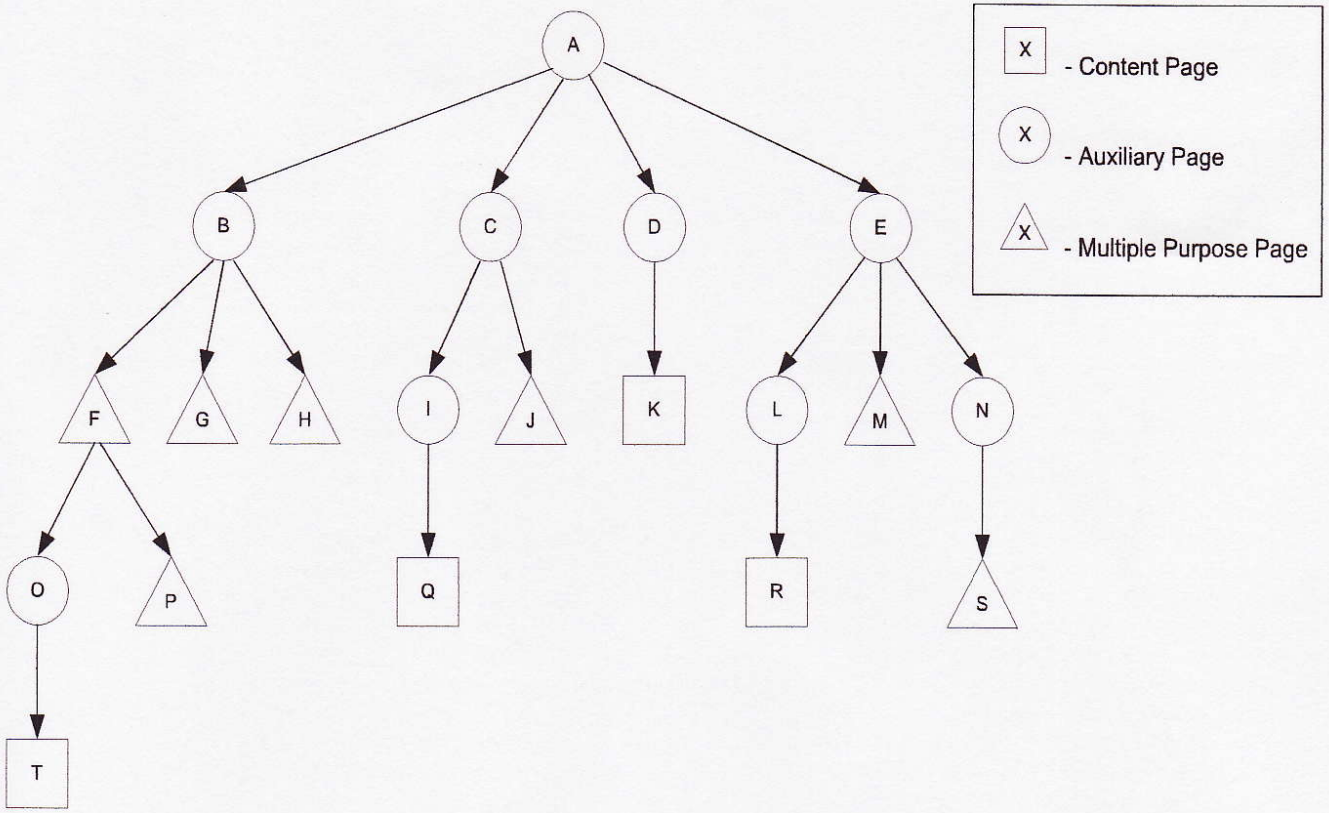


Fig. 5. Sample Web Site—Arrows between the pages represent hypertext links

User, Session Identification und Path Completion

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referred	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
12	123.456.78.9	-	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
13	123.456.78.9	-	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1680	A.html	Mozilla/3.04 (Win95, I)

Fig. 6. Sample Information from Access, Referrer, and Agent Logs (The first column is for referencing purposes and would not be part of an actual log).

User, Session Identification und Path Completion

Task	Result
Clean Log	• A-B-L-F-A-B-R-C-O-J-G-A-D
User Identification	• A-B-F-O-G-A-D • A-B-C-J • L-R
Session Identification	• A-B-F-O-G • A-D • A-B-C-J • L-R
Path Completion	• A-B-F-O-F-B-G • A-D • A-B-A-C-J

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Data Mining Techniken werden angewendet, um Regeln und Patterns zu erkennen.

Details

- Association Rule Mining (Apriori-Algorithmus)
- Sequential Patterns
- Page Clusters
- Usage Statistics (Number of Hits per Page, meistbesuchte Seite, Startseite, durchschnittliche Verweildauer per Page)

Pattern Analysis

Die gefundenen Regeln und Patterns werden analysiert, um die interessanten Regeln, Patterns und Statistiken zu erhalten.

- OLAP / Visualization
- Knowledge Query Mechanism

Introduction

Basics

Data Preparation Process

Preprocessing

Knowledge Discovery, Pattern Analysis

Apriori-Algorithmus

Behavior Models

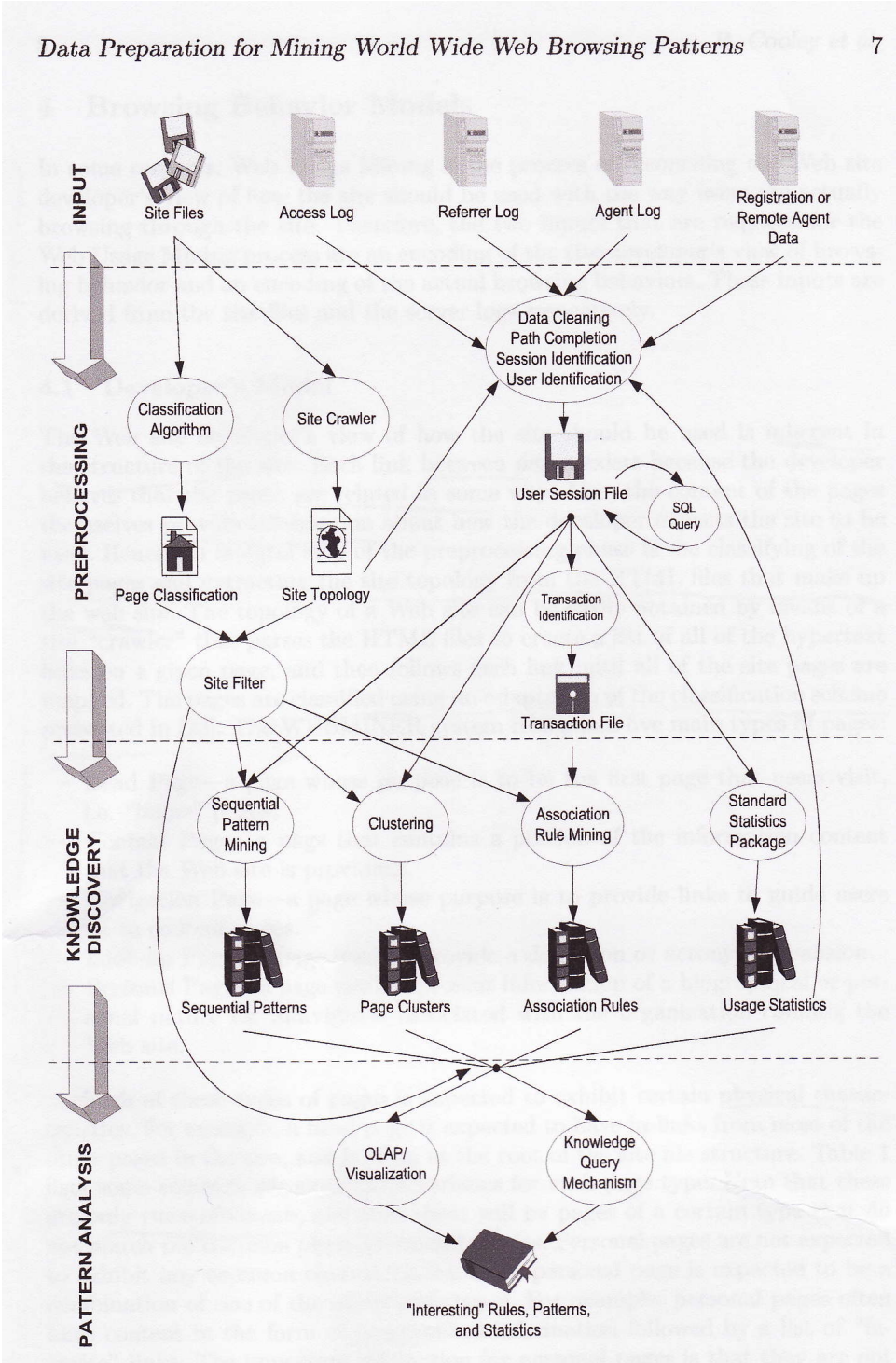
Developer's Model

Users' Model

Transaction Identification

WEB MINER SYSTEM

Conclusion



Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Übersicht

1 Introduction

2 Basics

3 Data Preparation Process
- Preprocessing
- Knowledge Discovery, Pattern Analysis

4 Apriori-Algorithmus

5 Behavior Models
- Developer's Model
- Users' Model

6 Transaction Identification

7 WEB MINER SYSTEM

8 Conclusion

Apriori-Algorithmus

Mit Einführung des Frequent Set Mining Problems auch erster Lösungsansatz $\Rightarrow$ Apriori-Algorithmus.

- Löst frequent set mining problem und association rule mining problem.
- Schlüsselidee: Teilmengen häufiger Itemmengen ebenfalls häufig, und Obermengen nicht häufiger Itemmengen ebenfalls nicht häufig.
- Klar da für Itemmengen X, Y mit $X \subseteq Y$ gilt: $support(Y) \leq support(X)$

Apriori Details

- Bestimme alle einelementigen häufigen Itemmengen. Suche in Obermengen weitere häufige Itemmengen.
- Benötigt m Iterationen, wobei m Kardinalität der größten häufigen Itemmenge.
- Ausgabe ist die Menge aller Itemmengen I mit $support(I) \geq minsupp$

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Apriori-Algorithmus

Eingabe: Datenbasis D
Ausgabe: Menge L_k häufiger Itemmengen I, mit $support(I) \geq minsupp$

- Berechnung häufiger 1-Itemmengen L_1
- $k-1 \rightarrow k$:
  - Berechnung von Kandidatenmengen C_k aus den (k-1)-häufigen Itemmengen mittels AprioriGen
  - Berechnung des tatsächlichen Supports der Kandidatenmengen
  - Aufnahme der Mengen mit genügend hohem Support (mindestens minsupp) in L_k
- Ausgabe von $\bigcup L_k$

AprioriGen-Algorithmus

Details

- Nutzt aus, dass jede Teilmenge einer häufigen Itemmenge wieder häufig sein muss.
- Items sind lexikographisch geordnet.

AprioriGen

Eingabe: Menge häufiger (k - 1)-Itemmengen L_{k-1}
Ausgabe: Obermenge C_k der Menge häufiger k-Itemmengen

- Je zwei häufige (k - 1)-Itemmengen $p, q \in L_{k-1}$ mit denselben ersten (k - 2) Elementen werden zu einer k-Itemmenge vereinigt $\rightarrow C_k$
- Entferne aus C_k diejenigen Itemmengen, deren (k - 1)-Teilmengen nicht alle in L_{k-1} liegen.

AprioriGen-Beispiel

Items: $A < B < C < ...$
$L_3 = \{\{A, B, C\}, \{A, B, D\}, \{A, C, D\}, \{A, C, E\}, \{B, C, D\}\}$

- $C_4 : \{A, B, C, D\}, \{A, C, D, E\}$
- Entferne $\{A, C, D, E\}$, da $\{A, D, E\} \notin L_3$.

Ausgabe: $C_4 = \{\{A, B, C, D\}\}$
Beachte: $\{A, B, C, E\}$ wird nicht in C_4 aufgenommen, da $\{A, B, C\}$ und $\{A, B, E\}$ nicht beide in L_3 enthalten sind.

Assoziationsregeln

Algorithmus ebenfalls anwendbar auf Konfidenz, berechne alle Assoziationsregeln mit *confidence* $\geq$ *minconf*

- Nutze aus: Für Itemmengen X,Y mit $Y \subset X$ beträgt die confidence der Regel $(X - Y) \rightarrow Y$ mindestens minconf, dann gilt dies auch für jede Regel $(X - Y') \rightarrow Y'$, wobei $Y' \subseteq Y$ ist.
- $\{A\} \rightarrow \{B, C\} >$ *minconf*, dann ist $\{A, B\} \rightarrow \{C\} >$ *minconf*
- Berechne alle Regeln, deren Konklusion nur ein Item enthält. Erweitere Konklusion elementweise
- Benutze Apriori+AprioriGen

Fazit

Apriori-Algorithmus löst effizient frequent set mining problem und association rule mining problem.
Algorithmus braucht nur strukturierten Input $\Rightarrow$ Data Preparation wichtig!

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Übersicht

1 Introduction

2 Basics

3 Data Preparation Process
- Preprocessing
- Knowledge Discovery, Pattern Analysis

4 Apriori-Algorithmus

5 Behavior Models
- Developer's Model
- Users' Model

6 Transaction Identification

7 WEB MINER SYSTEM

8 Conclusion

Introduction

Basics

Data Preparation Process

Preprocessing

Knowledge Discovery, Pattern Analysis

Apriori-Algorithmus

Behavior Models

Developer's Model

Users' Model

Transaction Identification

WEB MINER SYSTEM

Conclusion

Ziel

Vergleich der Benutzung der Webseite unter den Gesichtspunkten des Entwicklers und des Users.

Sichten

- Entwicklersicht eindeutig festgelegt.
- Usersicht spezifiziert durch die Server-Log Datei. (Zugriffe auf die Webseite)

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Developer's Model

Details

- Struktur der Seite spiegelt die Sicht des Entwicklers wieder.
- Topology einer Webseite kann durch „Site Crawler" bestimmt werden.

Page Classification

Webseiten können in Typen klassifiziert werden.

- Head Page
- Content Page
- Navigation Page
- Look-up Page
- Personal Page

Developer's Model

Charakteristika von Webseiten

Page Type	Physical Characteristics	Usage Characteristics
Head	• In-links from most site pages • Root of site file structure	• First page in user sessions
Content	• Large text/graphic to link ratio	• Long average reference length
Navigation	• Small text/graphic to link ratio	• Short average reference length • Not a maximal forward reference
Look-up	• Large number of in-links • Few or no out-links • Very little content	• Short average reference length • Maximal forward reference
Personal	• No common characteristics	• Low usage

Page Classification entweder durch Classification Algorithmus lernen oder jeder Webseite Classification Tag geben.

Users' Model

Charakteristika von Webseiten

Page Type	Physical Characteristics	Usage Characteristics
Head	• In-links from most site pages • Root of site file structure	• First page in user sessions
Content	• Large text/graphic to link ratio	• Long average reference length
Navigation	• Small text/graphic to link ratio	• Short average reference length • Not a maximal forward reference
Look-up	• Large number of in-links • Few or no out-links • Very little content	• Short average reference length • Maximal forward reference
Personal	• No common characteristics	• Low usage

Betrachtete Typen

- Content Pages
- Andere Seiten: Auxiliary Pages

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Site Filter

Voraussetzung: Site Topology und Page Classification ist bekannt.

- Vergleich zwischen Developer's Model und User's Model.
- Falls Unterschiede bestehen $\Rightarrow$ Report.
- Beispiel: Webseite, die vom User als Startseite benutzt wird, aber vom Entwickler nicht als Startseite deklariert ist.
- Benutzt Site Topology, um uninteressante Regeln zu eliminieren.
- Beispiel: Developer definiert Link zwischen zwei Seiten, User benutzt den Link (irrelevante Regel)

Introduction

Basics

Data Preparation Process

Preprocessing

Knowledge Discovery, Pattern Analysis

Apriori-Algorithmus

Behavior Models

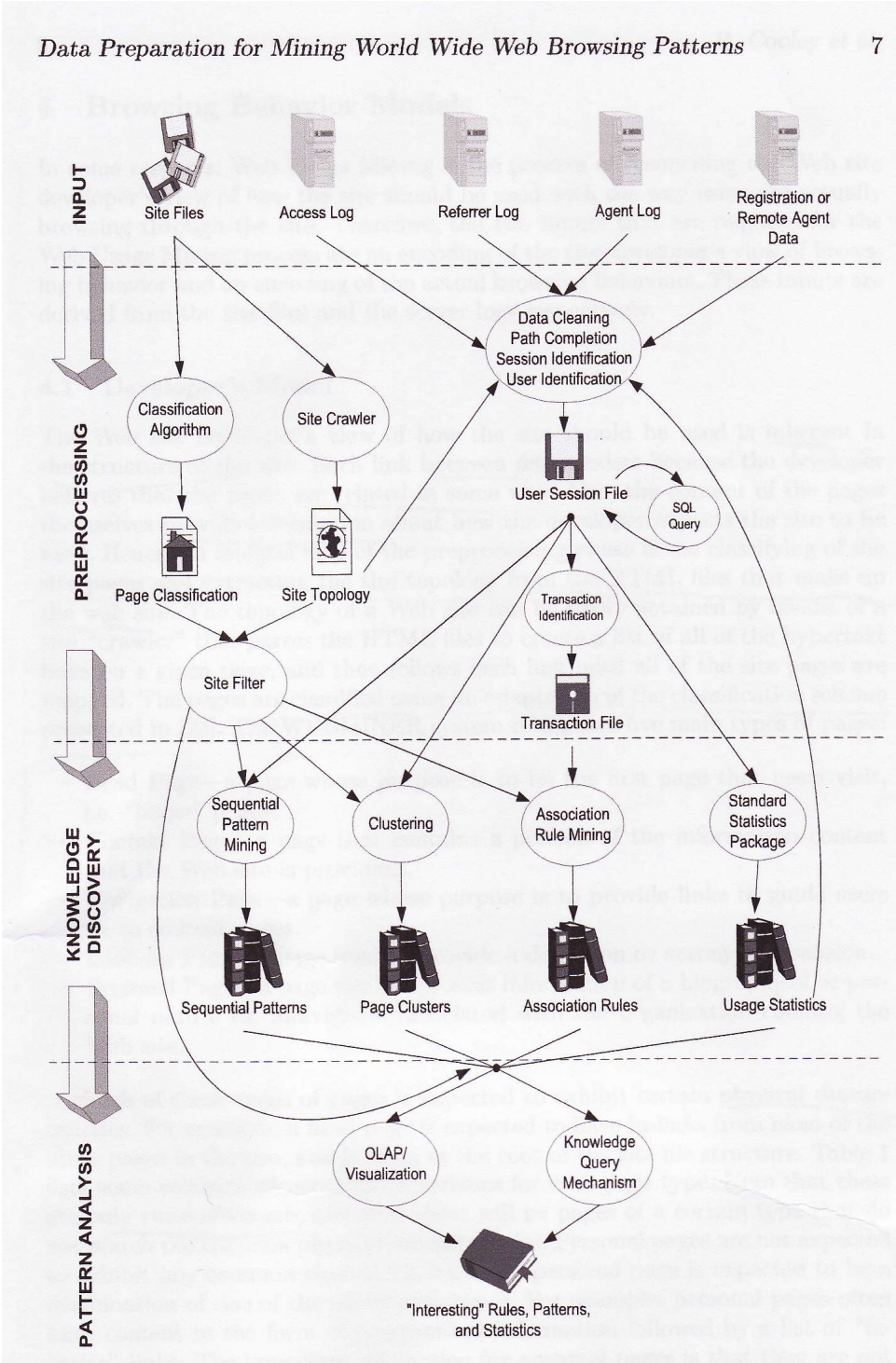
Developer's Model

Users' Model

Transaction Identification

WEB MINER SYSTEM

Conclusion



Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Übersicht

1 Introduction

2 Basics

3 Data Preparation Process
- Preprocessing
- Knowledge Discovery, Pattern Analysis

4 Apriori-Algorithmus

5 Behavior Models
- Developer's Model
- Users' Model

6 Transaction Identification

7 WEB MINER SYSTEM

8 Conclusion

Voraussetzung: User Sessions wurden identifiziert.
Ziel: Erzeuge sinnvolle Cluster von Page References für jede User Session.

Zwei Ansätze

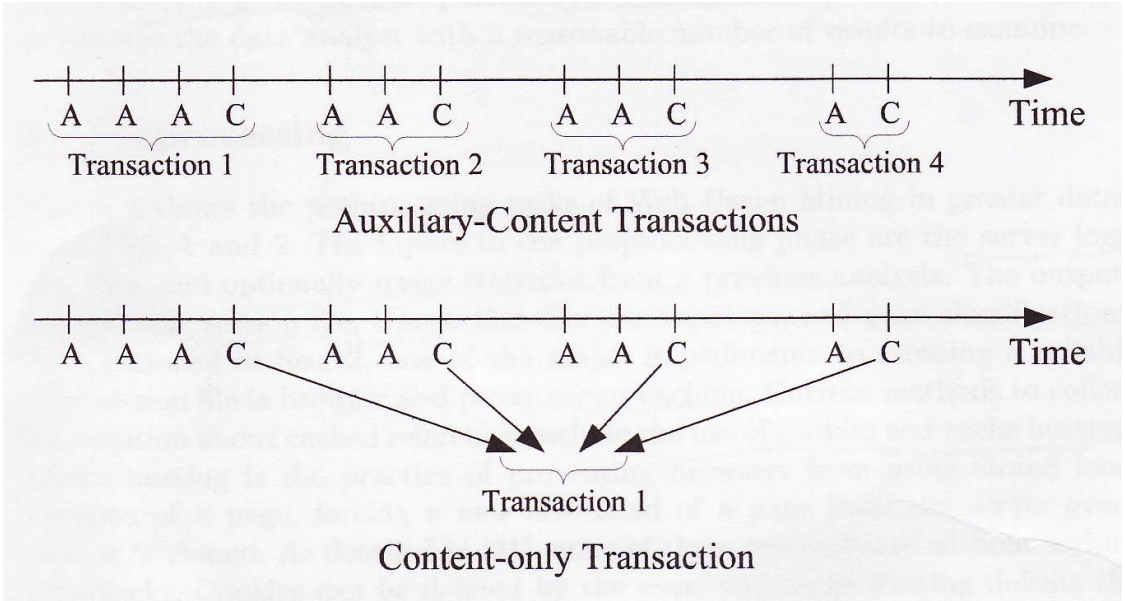
- User Session ist eine Transaktion mit vielen Page References (Divide-Verfahren)
- User Session beinhaltet ganz viele Transaktionen mit jeweils einer Page Reference (Merge-Verfahren)

Verwendete Verfahren

- reference length
- maximal forward reference
- time window

Typen von Transaktionen

1. auxiliary-content transactions: Enthält alle Auxiliary Pages inklusive aller Content Pages in einem gewissen Zeitraum.
   - Liefern den Common Traversal Path für eine Content Page durch eine Webseite.
2. content-only transactions: Enthält alle Content Pages in einem gewissen Zeitraum.
   - Liefern Beziehungen zwischen Content Pages ohne den Pfad zwischen Seiten zu betrachten.



Transaction Identification by Reference length

Beziehung zwischen Verweildauer auf Webseite und Typ der Seite (auxiliary oder content Page)

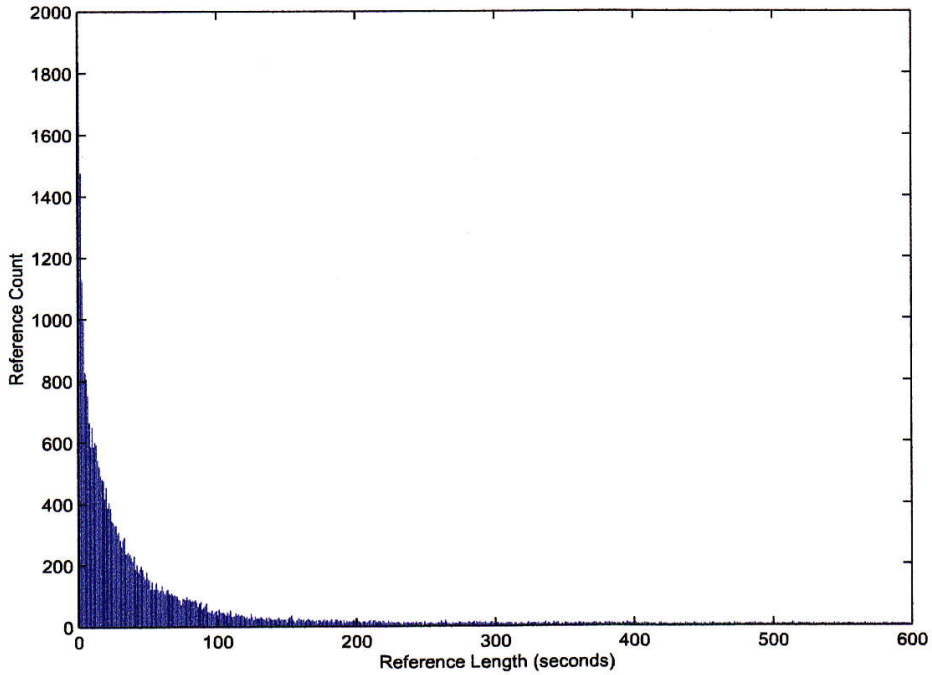


Fig. 7. Histogram of Web Page Reference Lengths (seconds)

1. auxiliary pages wenig Varianz
2. content references viel Varianz

Transaction Identification by Reference length

Details

- Mache Annahme über % der auxiliary references in einer Server-Log Datei
- Berechne reference length t, die den Split zwischen auxiliary oder content definiert
- Verwende maximum likelihood Schätzer um t zu berechnen:

$$t = \frac{-ln(1-\gamma)}{\lambda}$$

wobei $\gamma =$ % der auxiliary references (Annahme)
und $\lambda =$ Kehrwert des arithmetischen Mittels der beobachteten reference length

- Liefert einen brauchbaren Split-Wert
- Länge einer Referenz ist Differenz zwischen nächster Referenz und aktueller Referenz (ignoriere letzte Seite)

Transaction Identification by Reference length

Transaction

$$t =< ip_t, uid_t, \{(l_1^t.url, l_1^t.time, l_1^t.length), ..., (l_m^t.url, l_m^t.time, l_m^t.length)\}$$
$$for\ 1 \leq k \leq m, l_k^t \in L, l_k^t.ip = ip_t, l_k^t.uid = uid_t$$

Bei auxiliary-content transaction:

$$1 \leq k \leq (m-1) : l_k^t.length \leq C$$
$$k = m : l_k^t.length > C$$
C ist Splittime

Bei content-only transactions:

$$1 \leq k \leq m : l_k^t.length > C$$
C ist Splittime

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Transaction Identification by Maximal Forward Reference

Idee: Definiere jede Transaktion als Folge von Webseiten bis ein „Backtracking“ auftritt.

- Forward Reference: Webseite, die noch nicht in der aktuellen Transaktion ist.
- Backward Reference: Webseite, die schon in der aktuellen Transaktion ist.
- Starte neue Transaktion bei Forward Reference nach beliebig vielen Backward References.

Vorteile des Verfahrens: Es muss kein Inputparameter auf Grundlage des Datensatzes bestimmt werden.

Transaction Identification by Time Window

Vorgehen: Splitte die User Session in Zeitintervalle mit bestimmter Größe.
Es existiert nur ein Transaktionstyp. Sei W die Länge des Zeitfensters. Dann gilt:

$$(l_m^t.time - l_1^t.time) \leq W$$

Transaction Beispiel

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referred	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
12	123.456.78.9	-	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
13	123.456.78.9	-	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1680	A.html	Mozilla/3.04 (Win95, I)

Fig. 6. Sample Information from Access, Referrer, and Agent Logs (The first column is for referencing purposes and would not be part of an actual log).

Transaction Beispiel

Task	Result
Clean Log	• A-B-L-F-A-B-R-C-O-J-G-A-D
User Identification	• A-B-F-O-G-A-D • A-B-C-J • L-R
Session Identification	• A-B-F-O-G • A-D • A-B-C-J • L-R
Path Completion	• A-B-F-O-F-B-G • A-D • A-B-A-C-J • L-R

Transaction Beispiel

Splittime: $78.4 = \frac{-ln(1-0{,}5)}{\frac{1}{113}}$

Approach	Transactions	
	Content-only	Auxiliary-Content
Reference Length	F-G, D, L-R, J	A-B-F, O-F-B-G, A-D L, R, A-B-A-C-J
Maximal Forward Reference	O-G, R, B-J, D	A-B-F-O, A-B-G, L-R A-B, A-C-J, A-D
Time Window	A-B-F, O-F-B-G, A-D, L-R, A-B-A-C-J	

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Übersicht

1 Introduction
2 Basics
3 Data Preparation Process
- Preprocessing
- Knowledge Discovery, Pattern Analysis

4 Apriori-Algorithmus
5 Behavior Models
- Developer's Model
- Users' Model

6 Transaction Identification
7 WEB MINER SYSTEM
8 Conclusion

System, um Web Usage Mining zu testen.

Eigenschaften

- Input-Gewinn durch Common Log Format.
- Data Cleaning: Liste von Suffixen von Dateien, die entfernt werden sollen.
- Association Rule und Sequential Pattern generation implementiert.

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Transaction Identification

Versuchsaufbau:

- Server-Log Datei mit erzeugten Daten um die drei Verfahren zu vergleichen
- Data Generator: Erhält Webseite als gerichteter Graph und ein paar Assoziationsregeln
- Site Filter sorgt dafür, dass nur Regeln ausgegeben werden, die man **nicht** aus der Seitenstruktur ablesen kann
- Log-Entries: Forward Reference, Backward Reference oder Exit. (Random Number generator erzeugt Wahrscheinlichkeiten für die drei Fälle)
- % auxiliary References ist bekannt
- Drei Webseitentypen: spärlich verbundener Graph, dicht verbundenener Graph und durchschnittlich verbundener Graph

Results

Gefundene Regeln

Approach	Parameter	Sparse	Medium	Dense
Time Window	10 min.	0/4	0/3	0/3
	20 min.	2/4	2/3	1/3
	30 min.	2/4	2/3	2/3
Reference Length	50%	4/4	3/3	3/3
	65%	4/4	3/3	3/3
	80%	4/4	3/3	3/3
M. F. R.		4/4	2/3	1/3

Table 5. Ratio of Reported Confidence to Actual Confidence

Approach	Parameter	Sparse	Medium	Dense
Time Window	10 min.	null	null	null
	20 min.	0.82	0.87	0.87
	30 min.	0.98	0.90	0.88
Reference Length	50%	0.99	0.95	0.96
	65%	1.0	0.99	0.96
	80%	0.97	0.99	0.96
M. F. R.		0.79	0.47	0.44

Table 6. Transaction Identification Run Time (sec) / Total Run Time (sec)

Approach	Parameter	Sparse	Medium	Dense
Time Window	10 min.	0.81/4.38	0.82/4.65	0.75/3.94
	20 min.	0.84/7.06	0.80/7.06	0.73/4.42
	30 min.	0.79/7.16	0.77/9.95	0.72/5.17
Ref. Length	50%	1.82/4.62	1.66/4.46	1.47/4.09
	65%	1.68/4.29	1.72/4.35	1.45/4.02
	80%	1.62/4.14	1.66/4.26	1.48/4.03
M. F. R.		1.26/3.98	1.30/3.95	1.20/3.87

Introduction
Basics
Data Preparation Process
Preprocessing
Knowledge Discovery, Pattern Analysis
Apriori-Algorithmus
Behavior Models
Developer's Model
Users' Model
Transaction Identification
WEB MINER SYSTEM
Conclusion

Übersicht

1 Introduction
2 Basics
3 Data Preparation Process
- Preprocessing
- Knowledge Discovery, Pattern Analysis

4 Apriori-Algorithmus
5 Behavior Models
- Developer's Model
- Users' Model

6 Transaction Identification
7 WEB MINER SYSTEM
8 Conclusion

Introduction

Basics

Data Preparation Process

Preprocessing

Knowledge Discovery, Pattern Analysis

Apriori-Algorithmus

Behavior Models

Developer's Model

Users' Model

Transaction Identification

WEB MINER SYSTEM

Conclusion

Fazit:

- Web Usage Mining ist ein interessantes Anwendungsgebiet.
- Data Preparation ist ein recht aufwändiger Prozess und **ohne** Data Preparation sind Logfiles wertlos.

Literatur

- R. Cooley, B. Mobasher, J. Srivastava (1999) Data Preparation for Mining World Wide Web Browsing Patterns, Knowledge and Information Systems journal, 1, 5-32
- Bart Goethals Departement of Mathematics and Computer Science, University of Antwerp, Belgium DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK Chapter 17 Frequent Set Mining
- Methoden wissensbasierter Systeme Christoph Beierle Gabriele Kern-Isberner (2006)
- Folien DVEW 2006 Kapitel 5 Gabriele Kern-Isberner
- Internet Lexikon http://www.at-mix.de/

Introduction

Basics

Data Preparation Process

Preprocessing

Knowledge Discovery, Pattern Analysis

Apriori-Algorithmus

Behavior Models

Developer's Model

Users' Model

Transaction Identification

WEB MINER SYSTEM

Conclusion

Vielen Dank für eure Aufmerksamkeit!
Frohe Weihnachten und guten Rutsch! :-)