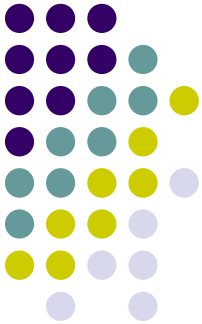
Web Mining und Farming

Shenwei Song



Gliederung

- **Übersicht über Web Mining und Farming**
- **Web Mining**
  - **Klassifikation des Web Mining**
  - **Wissensbasierte Wrapper-Induktion**
- **Web Farming**
  - **Übersicht über Web-Farming-Systeme**
  - **Verfeinerung der Web-Informationen**
- **Zusammenfassung**

Übersicht über Web Mining und Farming

- **Was ist Web Mining?**
  - Anwendung der Data-Mining-Technik zur automatischen Entdeckung und Extraktion der Informationen und Muster aus Web-Datenressourcen.
- **Warum Web Mining?**
  - Auffinden von neuen Kunden oder Geschäftsmöglichkeiten
  - Analysieren und Beobachten der Konkurrenz
  - Bedürfnisse web-basierter Applikationen besser erfüllen
- **Aufgaben des Web Mining**
  - Lokalisierung von Ressourcen
  - Extraktion von Informationen
  - Generalisierung von Mustern
  - Analyse der Muster

Übersicht über Web Mining und Farming

- **Was ist Web Farming?**
  - Systematische Verfeinerung der web-basierten Informationen für Bussiness Intelligence
- **Warum Web Farming?**
  - Verbesserung der Leistung der Geschäftsaktivtäten
  - Optimierung des Workflow der Geschäftsaktivitäten
- **Aufgabe des Web Farming**
  - Verbesserung der Data Warehouses

Übersicht über Web Mining und Farming

- **Unterschiede** zwischen Web Mining und Web Farming

	Web Mining	Web Farming
Schwerpunkte	Entdeckung und Extraktion der Informationen und Muster	Nachbearbeitung der entdeckten Informationen aus Web-Daten
Arbeitsweise	Kurzzeitig und diskontinuierlich	Systematisch Langzeitig und kontinuierlich

Klassifikation des Web Mining

- Klassifikation des Web Mining nach benutzten Datentypen:
  - **Web Content**
    - Daten direkt aus Web-Seiten, z.B. HTML, Multimediadateien und Graphiken
  - **Web Structure**
    - Hyperlink-Struktur zwischen Web-Seiten
  - **Web Usage**
    - Beschreibung der Nutzung von Web-Ressourcen, z.B. Log-Dateien von Proxy-Servern und Web-Servern

Web Content Mining

- Entdeckung nützlicher Informationen aus Web Content
- Zerlegung in zwei unterschiedliche Sichten
  - **Information-Retrieval-Sicht**
  - **Datenbanksicht**

Web Content Mining

- **Information-Retrieval-Sicht**
  - Unterstützung des Auffindens der Informationen oder der Filterung der Informationen
  - **Hauptdaten** → **Darstellung**
    (Hyper-)Textdokumente → Ausdrücke, Beziehungen, Wortmengen
  - **Applikation**:
    - Dokument-Kategorisierung
    - Auffinden der Muster in Texten
    - Auffinden der Extraktionsregeln

Web Content Mining

- **Datenbanksicht**
  - Modellierung der Daten für Netzanwendungen und Integration der Daten im Netz, um **komplizierte** Abfragen auszuführen
  - **Hauptdaten** → **Darstellung**
    Hypertext-Dokumente → Beziehungen, Edge-Labeled Graph
  - **Applikation**:
    - Auffinden von frequenten Substrukturen
    - Entdeckung von Web-Seiten-Schemata

Web Structure Mining

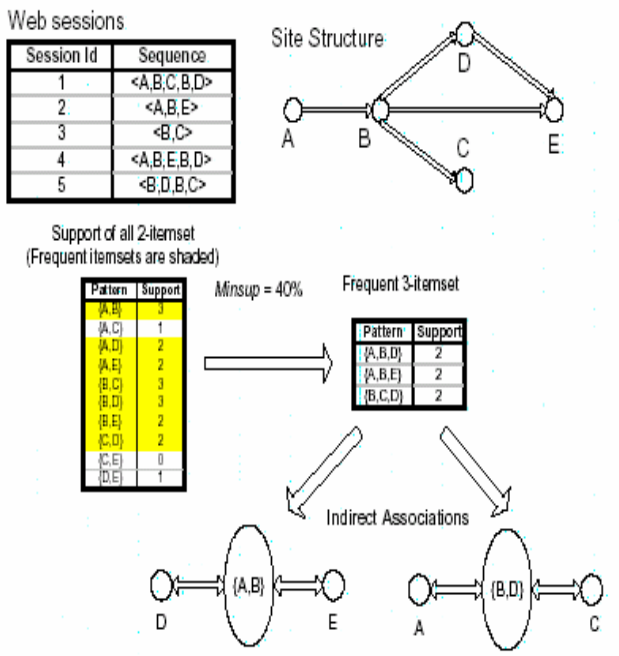
- Entdeckung des Modells der Verweisstruktur des Web
- Hauptdaten → Darstellung
  Linkstrukturen → Graphen
- Applikation
  - Generierung von Information über Unterschied und Ähnlichkeit zwischen den Web-Seiten
  - Maß der Vollständigkeit der Web-Seiten
  - Maß der Relevanz der Web-Seiten

Web Usage Mining

- Entdeckung der Zugriffsmuster
- Sammlung der Informationen über Benutzerverhalten
- Hauptdaten — Darstellung
  Server/Browser Logs ⟶ Graphen, relationale Tabellen
- Applikation
  - Konstruktion von Web-Seiten
  - Marketing
  - Benutzer-Modellierung

Web Usage Mining

- Beispiel der indirekten Beziehung aus Web-Usage-Daten



Wissensbasierte Wrapper-Induktion

- **Informationsextraktion**
  - Das Wiedererkennen und die Extraktion spezifischer Datenfragmente aus Dokumentsammlungen
  - Beispiel: Identifizierung von Datum, Ort und Temperaturwert aus Wettervoraussage
- **Wrapper**
  - Eine Regel oder eine Prozedur zur Extraktion von Information aus einer Informationsquelle
  - Auf Informationsquellen einer einzigen Art spezialisiert

Wissensbasierte Wrapper-Induktion

- **Kategorien der Wrapper-Generierung**
  - **Manuelle Wrapper-Generierung**
    - Festlegung von Extraktionsregeln von Menschen nach einer sorgfältigen Überprüfung von Beispiel-Web-Seiten
    - Hochpräzise Leistung, aber nicht automatisch anpassbar
  - **Wrapper-Induktion**
    - Automatische Bildung eines Wrappers durch Lernen aus Beispielseiten einer Informationsquelle
    - Zwei Typen:
      - **Heuristische Wrapper-Induktion**
      - **Wissensbasierte Wrapper-Induktion**

Heuristische Wrapper-Induktion

- Wrapper-Induktion auf Heuristiken basierend
- Einsatz in den meisten traditionellen Systemen
- Manchmal nicht effizient, da die Heuristiken einfach und naiv sein können.

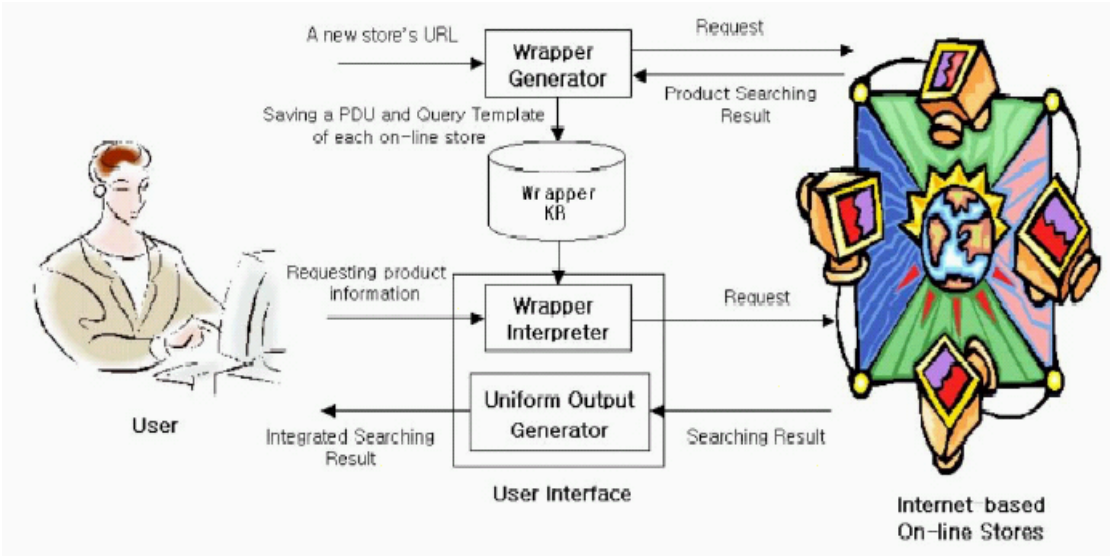
Wissensbasierte Wrapper-Induktion

- Manuelle Generierung und Repräsentation der **Domänenkenntnis**
  - Beschreibung von Terme, Konzept, und Beziehungen
  - Erkennung semantischer Fragmente eines Dokument
- Definition und Anwendung der Domänenkenntnis während Wrapper-Generierung
- Automatische Bildung eines Wrappers durch Benutzung der Domänenkenntnis

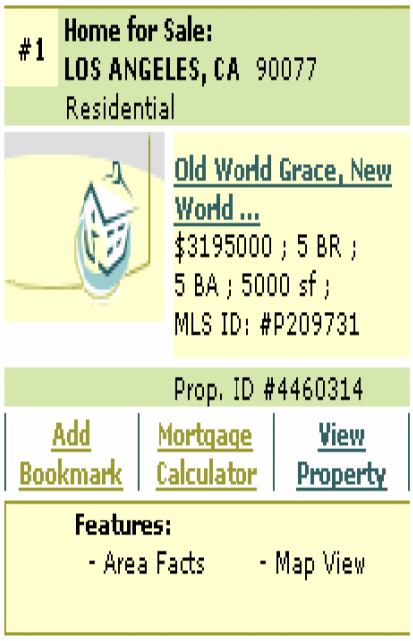
XTROS

- Wissensbasiertes Informationsextraktionssystem
- Repräsentation der Domänenkenntnis und Wrapper mit XML
  - Flexibilität und Modularität
  - Portabilität und Mitbenutzbarkeit
- Extraktion von Informationen aus Dokumenten

Übersicht über XTROS



Domänenkenntnis-Beispiel



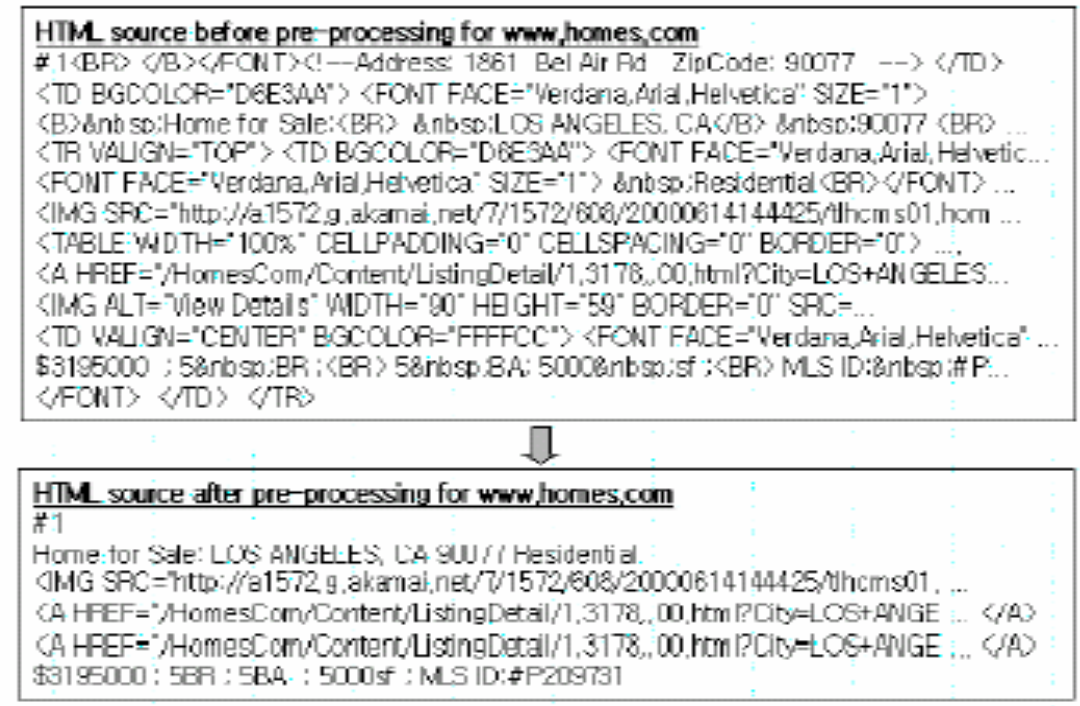

```
<KNOWLEDGE>

  <OBJECTS>
    <OBJECT>PRICE</OBJECT>
    <OBJECT>BED</OBJECT>
    <OBJECT>BATH</OBJECT>
    <OBJECT>CITY</OBJECT>
    <OBJECT>MLS</OBJECT>
    <OBJECT>DETAIL</OBJECT>
    <OBJECT>IMG</OBJECT>
  </OBJECTS>

  <PRICE>
    <ONTOLOGY>
      <TERM>PRICE</TERM>
      <TERM>$</TERM>
      <TERM>PRICE $</TERM>
    </ONTOLOGY>
    <FORMAT>
      <FORM>[ONTOLOGY] DIGITS</FORM>
      <FORM>DIGITS [ONTOLOGY]</FORM>
    </FORMAT>
  </PRICE>
```

```
<BED>
    <ONTOLOGY>
      <TERM>BEDROOM</TERM>
      <TERM>BED</TERM>
      <TERM>BEDS:</TERM>
      <TERM>BR</TERM>
      <TERM>BEDROOMS</TERM>
      <TERM>BD,</TERM>
    </ONTOLOGY>
    <FORMAT>
      <FORM>DIGITS [ONTOLOGY]</FORM>
      <FORM>[ONTOLOGY] DIGITS</FORM>
    </FORMAT>
  </BED>

  ------ omitted ------------

</KNOWLEDGE>
```

Wissensbasierter Wrapper-Generierungalgorithmus

- Erste Phase
  - Konvertierung der HTML-Quelle in **logische Linien**
- Zweite Phase
  - Bestimmung der Bedeutung der logischen Linien durch Einsatz der Domänenkenntnis
  - Kategorisierung der logische Linien
- Dritte Phase
  - Auffinden der **am häufigsten benutzten Muster** in **Sequenz der Katalognummern**

Konvertierung der HTML-Quelle in Logische Lininen



Bestimmung der Bedeutung logischer Linien

No.	object	line	cat	type	format
1	{{IMG}}	< A HREF=“ .. ”> {[<IMG[{ALT=“ .. ”}]>]} {{IMG}}	6	IMGURL	[ONTOLOGY] IMGURL
2	{{PRICE}}	{[[{$}]3195000]} {{PRICE}}	0	DIGITS	[ONTOLOGY] DIGITS
3	{{BED}}	; {[5[{BR}]]} {{BED}}	1	DIGITS	DIGITS [ONTOLOGY]
4	{{BATH}}	; {[5[{BA}]]} {{BATH}}	2	DIGITS	DIGITS [ONTOLOGY]
5	{{MLS}}	; 5000sf ; {[[{MLS ID:#}] P209731]} {{MLS}}	4	DIGITS	[ONTOLOGY] DIGITS
6	{{DETAIL}}	{[< A HREF=“ .. ”> [{View Property}]]} {{DETAIL}}	5	URL	URL [ONTOLOGY]

Auffinden der am häufigsten benutzten Muster

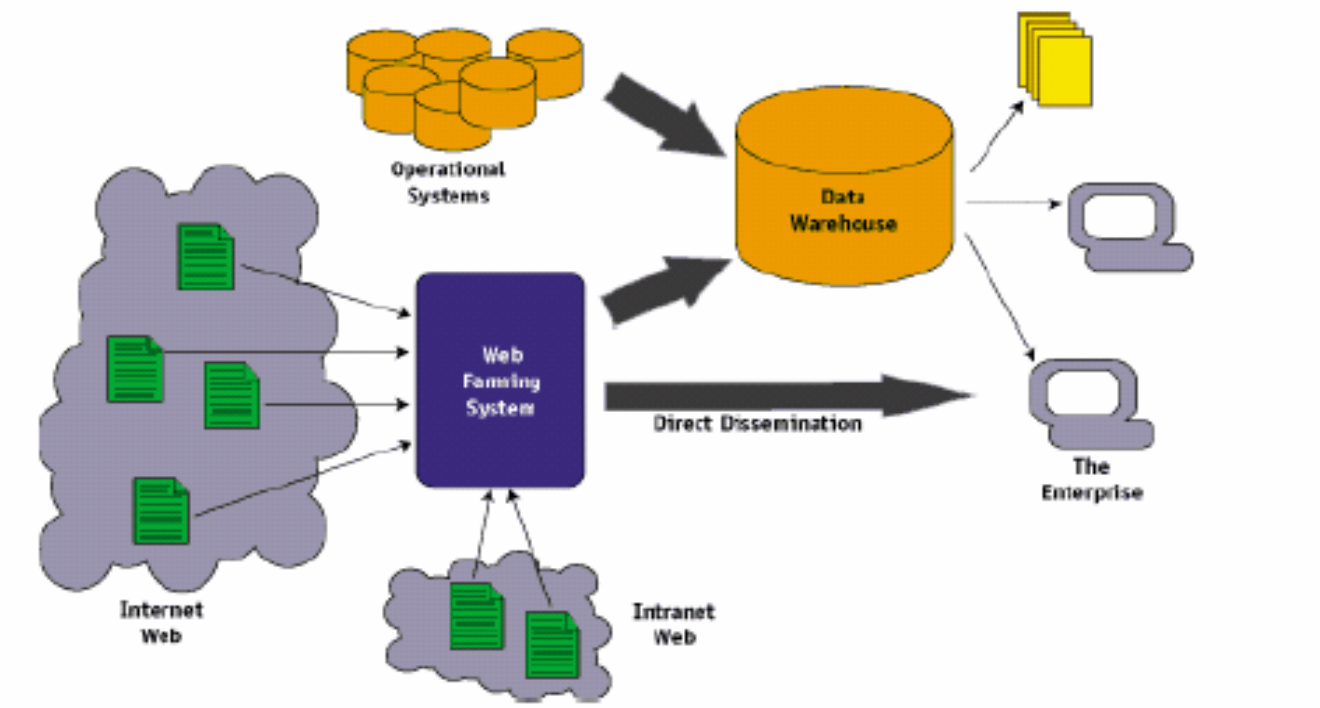
- Algorithmus:
  - Auffinden aller **Kandidatenmuster** aus **Sequenz** (maximale Substring)
    - Mindesten drei Attribute
    - Ohne duplizierte Attribute
  - **Fp:** Frequenz eines Kandidatenmusters in der Sequenz
  - **Ap**: Anzahl der Attribute in einem Muster
  - **Tp:** Gesamte Anzahl der Attribute des Musters
    - **Tp = Fp * Ap**

Beispiel der am häufigsten benutzten Muster

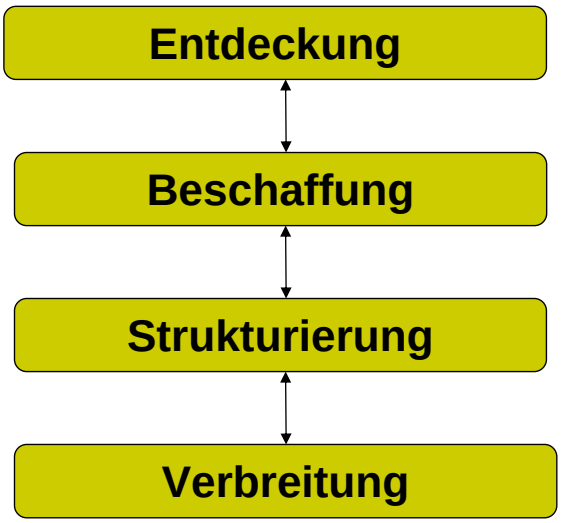
- **Sequenz**
  - **60124560124560124560124560124 5**

Kandidatenmuster	Muster Frequenz (Fp)	Anzahl der Attribute des Muster (Ap)	Gesamt Anzahl der Attribute des Muster (Tp=Ap*Fp)
601245	5	6	30
012456	4	6	24
124560	4	6	24
245601	4	6	24
456012	4	6	24
560124	4	6	24
56012	5	5	25
60125	1	5	5
01256	1	5	5
12506	1	5	5
25601	1	5	5
01245	5	5	25
1245	5	4	20
245	5	3	15

Übersicht über Web-Farming-Systeme



Verfeinerung der Web-Information



Verfeinerung der Web-Information

- Entdeckung
  - Lokalisierung von Individuen und Organisationen
- Beschaffung
  - Sammlung und Erhaltung von identifiziertem Inhalt
- Strukturierung
  - Analyse und Transformation des Inhalts in nützliche Form und Struktur
- Verbreitung
  - Verpackung und Lieferung der Informationen zum entsprechenden Verbraucher

Zusammenfassung

- **Übersicht über Web Mining und Farming**
- **Web Mining**
  - **Klassifikation des Web Mining**
  - **Wissensbasierte Wrapper-Induktion**
- **Web Farming**
  - **Übersicht über Web-Farming-Systeme**
  - **Verfeinerung der Web-Informationen**