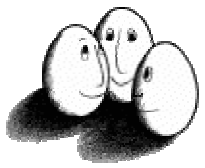
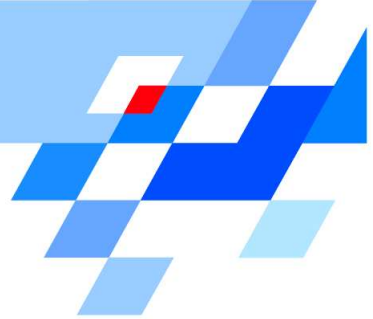
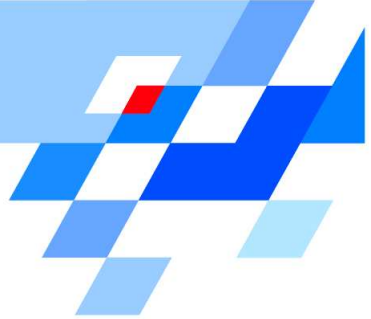




Seminar „Informationsextraktion“

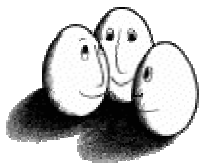
Structuring Domain-Specific Text Archives by Deriving a Probabilistic XML DTD

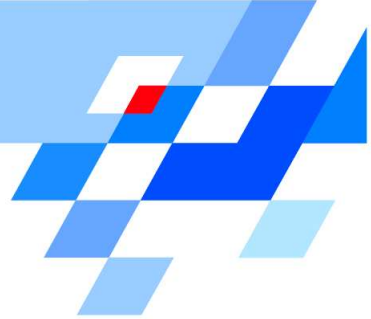
Karsten Winkler, Myra Spiliopoulou
Handelshochschule Leipzig
PKDD, Springer Verlag, 2002



Einordnung

- bisher im Seminar 2 Ansätze
  - NLP-basiert
  - Wrapper (nutzen Struktur der Dokumente)
- heute:
  - Wie kommt man von einem „normalen“ Text zu einem, dessen Struktur man dann nutzen kann?



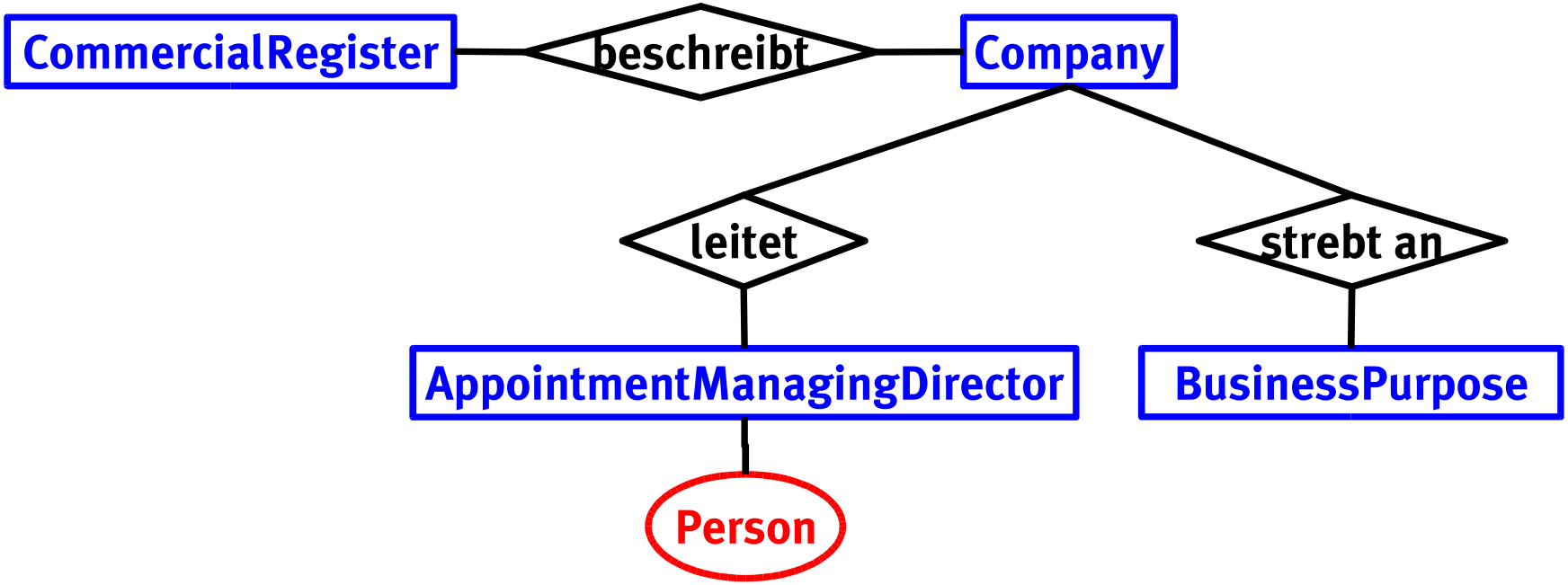


XML: semistrukturierte Texte

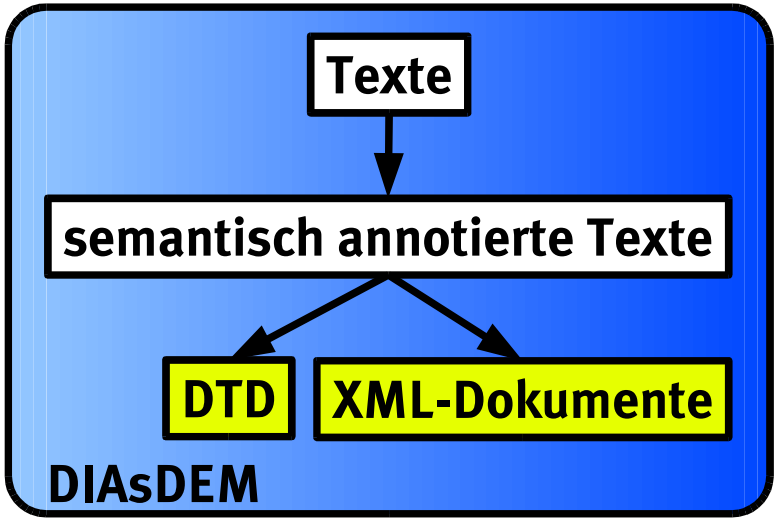
- Dokument wird mit „Tags“ ausgezeichnet
- ideal: Tags im Dokument entsprechen Slots im auszufüllenden Template
- DTD beschreibt
  - welche Elemente (Tags) gibt es?
  - optionale/notwendige Elemente
  - Schachtelungen/Reihenfolgen von Elementen

Domäne: Einträge ins Handelsregister

<BusinessPurpose>Der Betrieb von Spielhallen in Teltow und das Aufstellen von Geldspiel- und Unterhaltungsautomaten.</BusinessPurpose>
<AppointmentManagingDirector Person="Balski; Pawel"> Pawel Balski ist zum Geschäftsführer bestellt. </AppointmentManagingDirector>

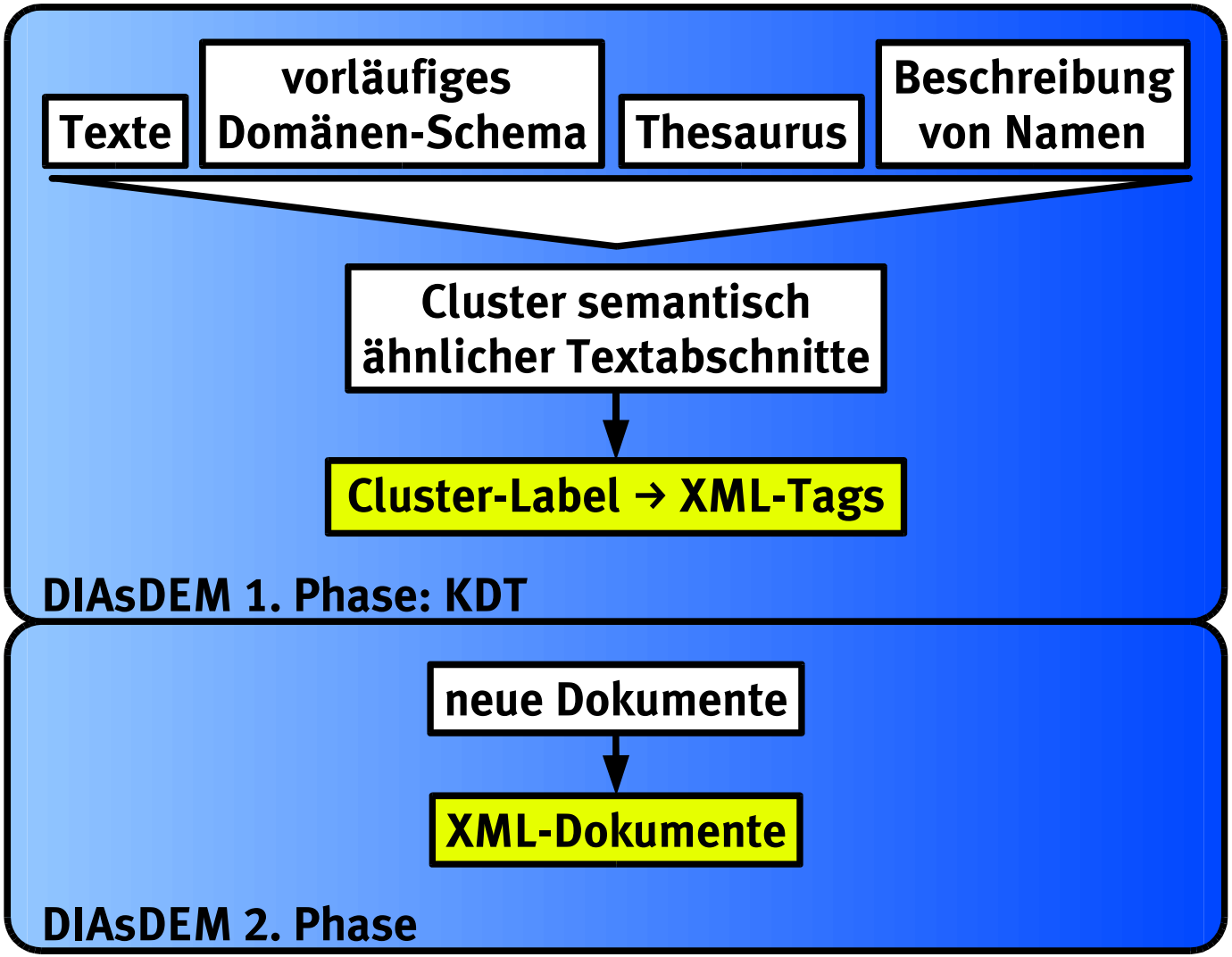


Einordnung in Arbeit der Autoren

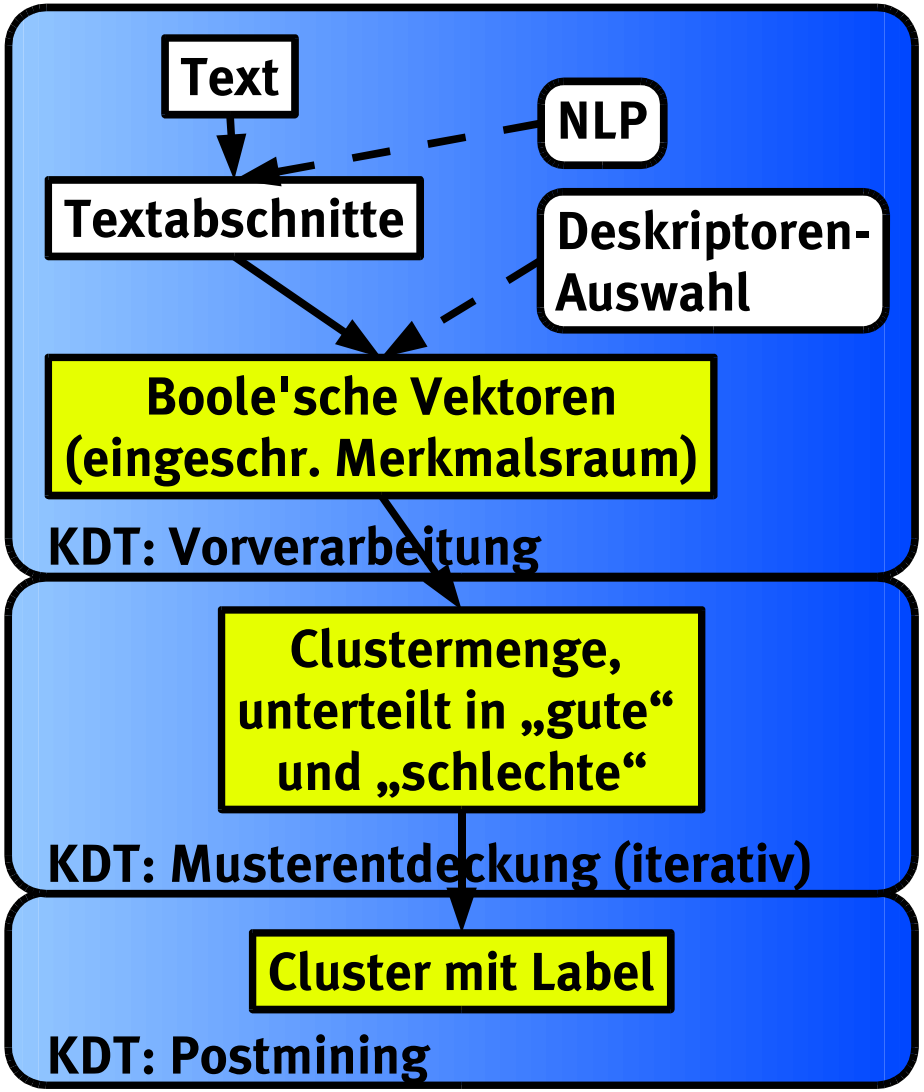


- DIAsDEM-DTD:
  - unstrukturiert
  - nicht-hierarchisch
- langfristig: Schema (XML/relational)
- Zwischenschritt: probabilistische DTD

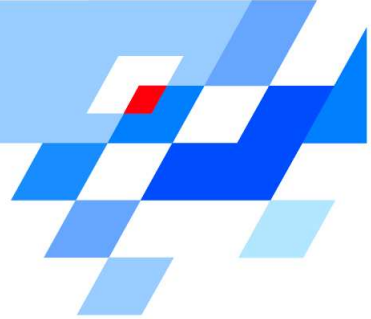
DIAsDEM: Arbeitsweise



DIAsDEM: KDT-Prozess



- „gutes“ Cluster von Textvektoren:
  - Kardinalität groß
  - Texteinheiten homogen
  - kann mit wenigen Deskriptoren beschrieben werden

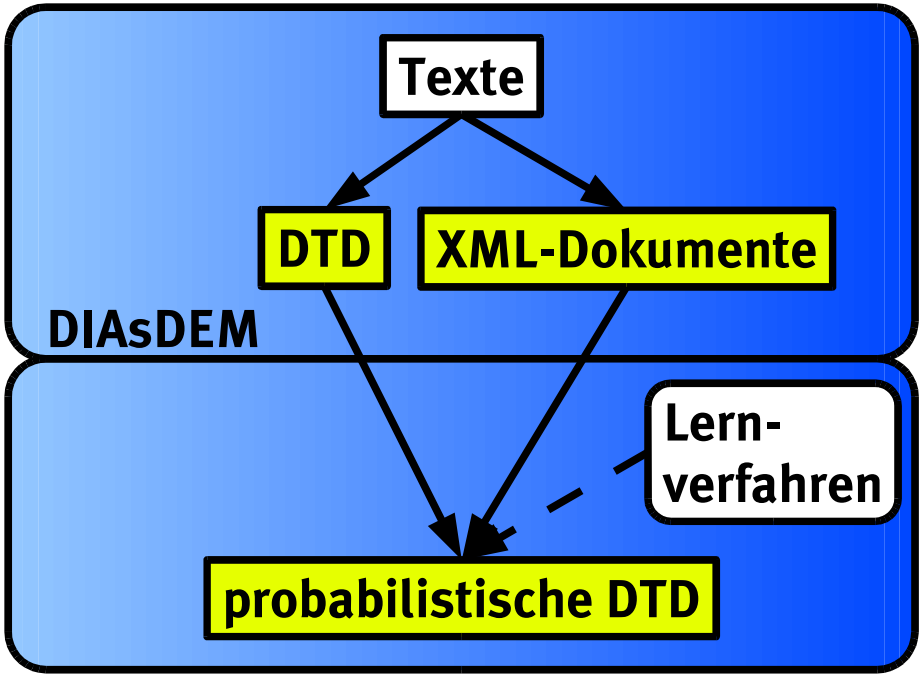


DIAsDEM: Ergebnis und Beispiel

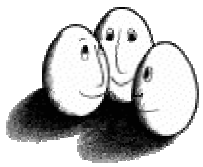
```
<!ELEMENT CommercialRegisterEntry ( #PCDATA | FoundationPartnership |
ShareCapital | AppointmentManagingDirector | (...) | ConclusionArticles |
BusinessPurpose | Owner )*>
(...)
<!ELEMENT Owner ( #PCDATA ) >
```

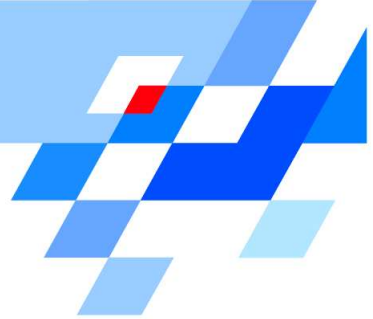
- unstrukturiert
- keine Informationen über
  - notwendige Elemente
  - optionale Elemente
  - Abhängigkeiten zwischen Elementen

Probabilistische DTDs: Motivation



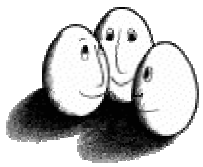
- auf DIAsDEM-Ergebnis Lernverfahren anwenden
  - → oft auftretende „Gruppen“ von Tags finden
- gesucht: Datenstruktur für das Ergebnis

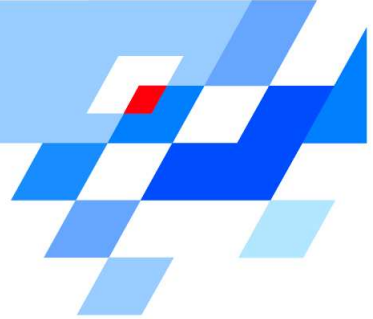




Probabilistische DTDs: Beschreibung

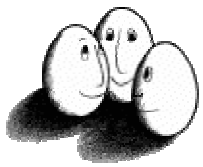
- Graph-basierte Datenstruktur
- beschreibt strukturelle Eigenschaften des XML-Archives
  - wahrscheinlichste Anordnungen von Elementen
    - Reihenfolgen von Tags
    - oft gemeinsam auftretende Tags
    - sich gegenseitig ausschließende Tags
  - statistische Eigenschaften der Elemente

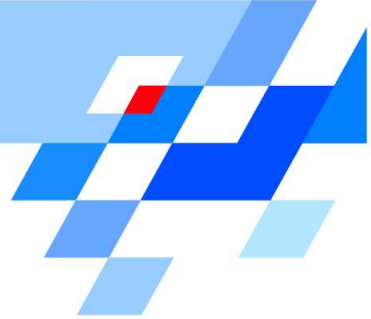




Probabilistische DTDs: Definitionen (1/2)

- d: XML-Dokument im Archiv $D = \{d_1,...,d_{|D|}\}$
- $T = \{t_1,...,t_{|T|}\}$: Menge der Tags in der abgeleiteten DTD
- $\{x,y_1,...,y_n\} \subseteq T$: Menge von Tags
- $\langle y_1 \cdot ... \cdot y_n \cdot x \rangle \in T^{n+1}$: Folge benachbarter Tags
- „Datentypen“: Tag, Assoziation von Tags, Folge von Tags
  Gruppe: Assoziation oder Folge





Probabilistische DTDs: Definitionen (2/2)

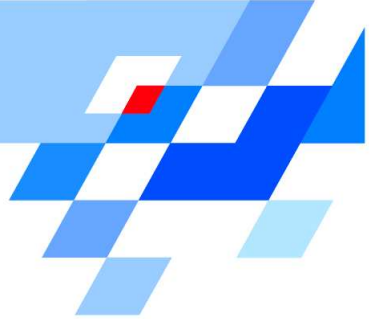
- Tags(d): Menge aller XML-Tags in d
- Seqs(d): Menge aller Folgen von benachbarten XML-Tags in d

d

<eins>...</eins>
<zwei>...</zwei>
<eins>...</eins>

Tags(d) = {eins,zwei}

Seqs(d) = {eins·zwei·eins,eins·zwei,zwei·eins}



Statistische Eigenschaften von Elementen (1/2)

- $$TagSupport(x) = \frac{|\{d \in D | x \in Tags(d)\}|}{|D|}$$

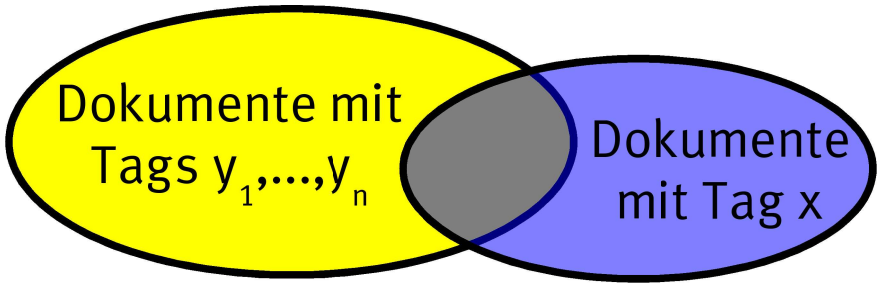
  → wie oft kommt Tag x vor?

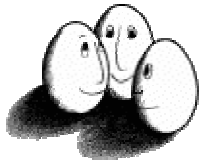
- $$GroupSupport(g) = \frac{|\{d \in D | g \in Tags(d) \cup Seqs(d)\}|}{|D|}$$

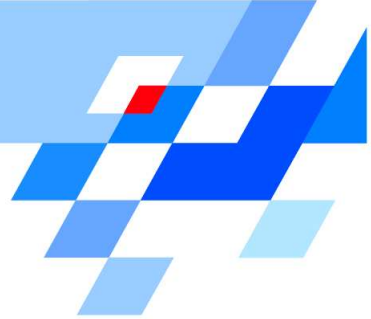
  → wie oft kommt Gruppe g vor?

Statistische Eigenschaften von Elementen (2/2)

	Association $y_1,\ldots,y_n \rightarrow x$	Sequence $y_1 \cdot \ldots \cdot y_n \cdot x$
Confidence	$\frac{\lvert\{d \in D \mid x, y_1, \ldots, y_n \subseteq Tags(d)\}\rvert}{\lvert\{d \in D \mid y_1, \ldots, y_n \subseteq Tags(d)\}\rvert}$	$\frac{\lvert\{d \in D \mid y_1 \cdot \ldots \cdot y_n \cdot x \in Seqs(d)\}\rvert}{\lvert\{d \in D \mid y_1 \cdot \ldots \cdot y_n \in Seqs(d)\}\rvert}$
Lift	$\frac{AssociationConf(y_1, \ldots, y_n \rightarrow x)}{TagSupport(x)}$	$\frac{SequenceConf(y_1 \cdot \ldots \cdot y_n \cdot x)}{TagSupport(x)}$





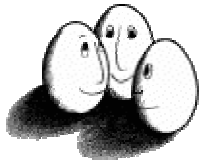


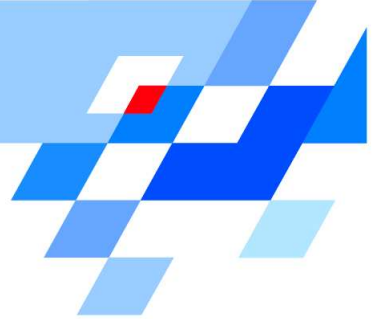
Ableiten einer DTD

- 2 Möglichkeiten:
  - aus Assoziationsgraph
  - aus Alternativenbaum
- Autoren probieren noch beides aus
- semi-automatisch, d.h. generiertes Ergebnis hilft Experten auf dem Gebiet

Assoziationsgraph: Definition (1/2)

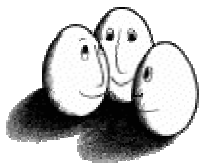
- Knoten: Tag(-Gruppe) + Statistik
- Knotenmenge $\mathcal{V} = V \cup V'$ mit
  - $V \subseteq T \times]0,1]$: Knoten der Form <TagName,TagSupport>
    - nur, wenn TagSupport > 0
  - $V' \subseteq (\wp(V) \times \{0,1\}) \times]0,1]$: Knoten der Form <TagGroup,GroupSupport>
    - nur, wenn GroupSupport > Schwellwert

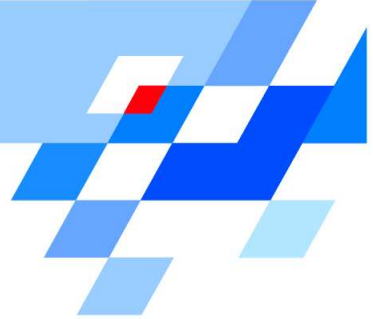




Assoziationsgraph: Definition (2/2)

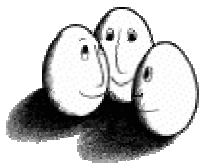
- Kante: Beziehung $y_1,\ldots,y_n \rightarrow x$ oder $y_1 \cdot \ldots \cdot y_n \cdot x$
- Kanten $E \subseteq (V \times V') \times x^4$; $x :=]0,1] \cup \{-\}$
  - Form:
    <Kante, AssociationConfidence, AssociationLift, SequenceConfidence, SequenceLift>
  - Folge: nur Sequence...;
    Menge: nur Association...
  - Eigenschaft nicht vorhanden: Wert „-“

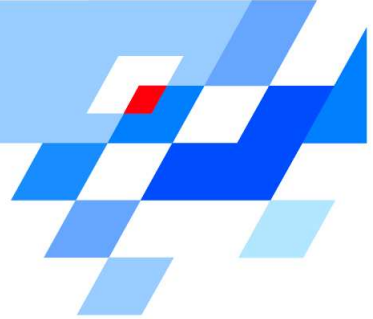




Assoziationsgraph: Pruning-Algorithmus

- lösche alle Kanten mit Lift < 1
- lösche alle Kanten mit Konfidenz < t
- lösche alle Gruppen-Knoten ohne ausgehende Kante zu einem Tag-Knoten
- für jeden Tag-Knoten:
  - behalte nur Gruppen-Knoten mit maximaler Konfidenz

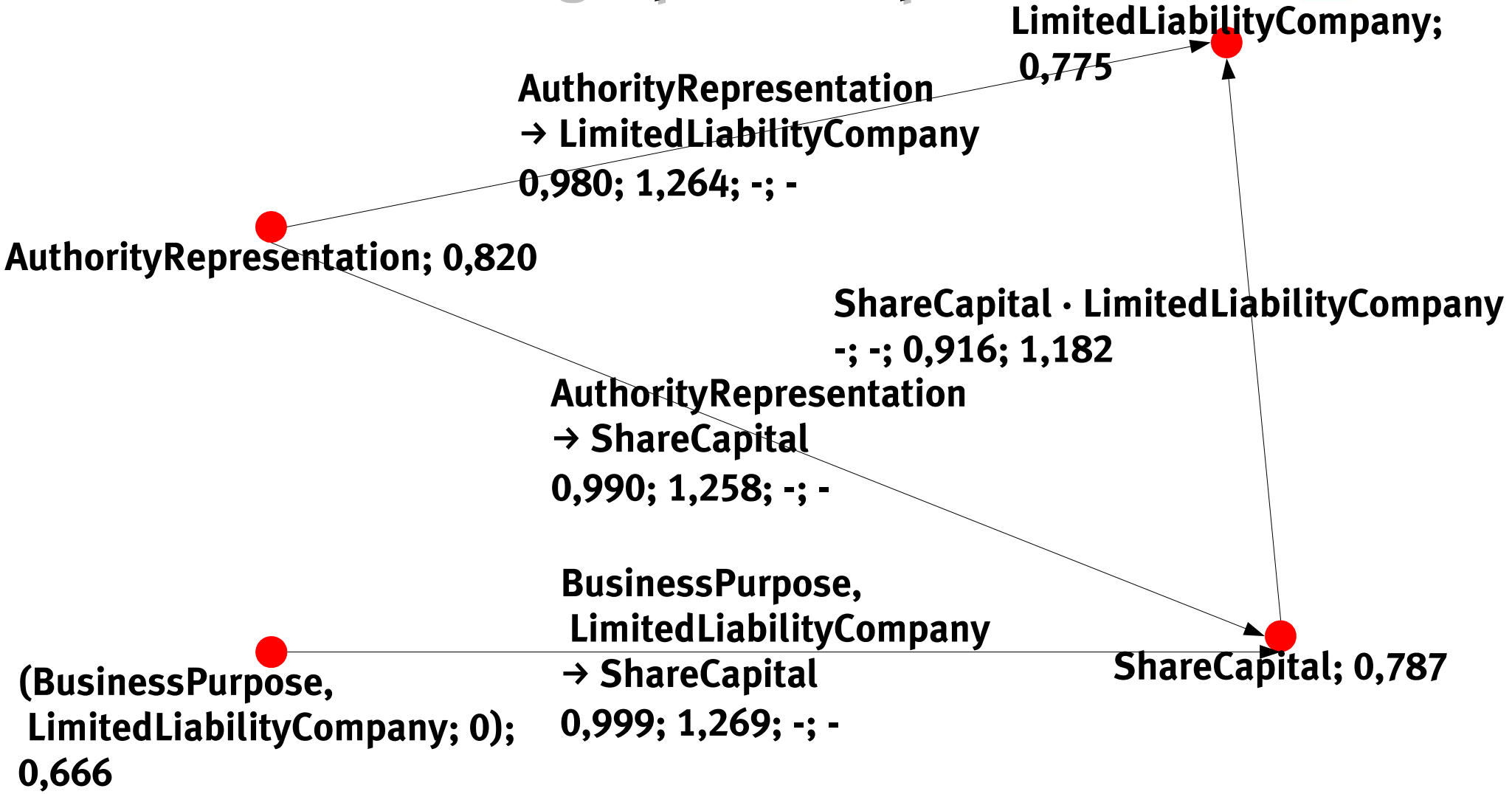


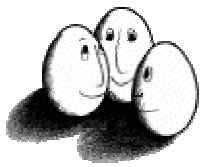


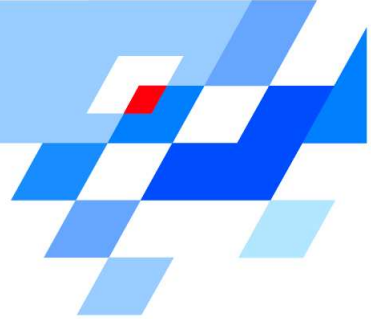
Assoziationsgraph: Ergebnis

- Eigenschaften gelöschter Gruppen:
  - falsche Assoziationen
  - führen zu Tags mit niedriger Konfidenz
  - können durch Gruppen ersetzt werden, die zu häufigen Tags mit höherer Konfidenz führen
- Ausgabe: Sammlung häufiger DTD-Elemente (keine wohlgeformte DTD!)

Assoziationsgraph: Beispiel



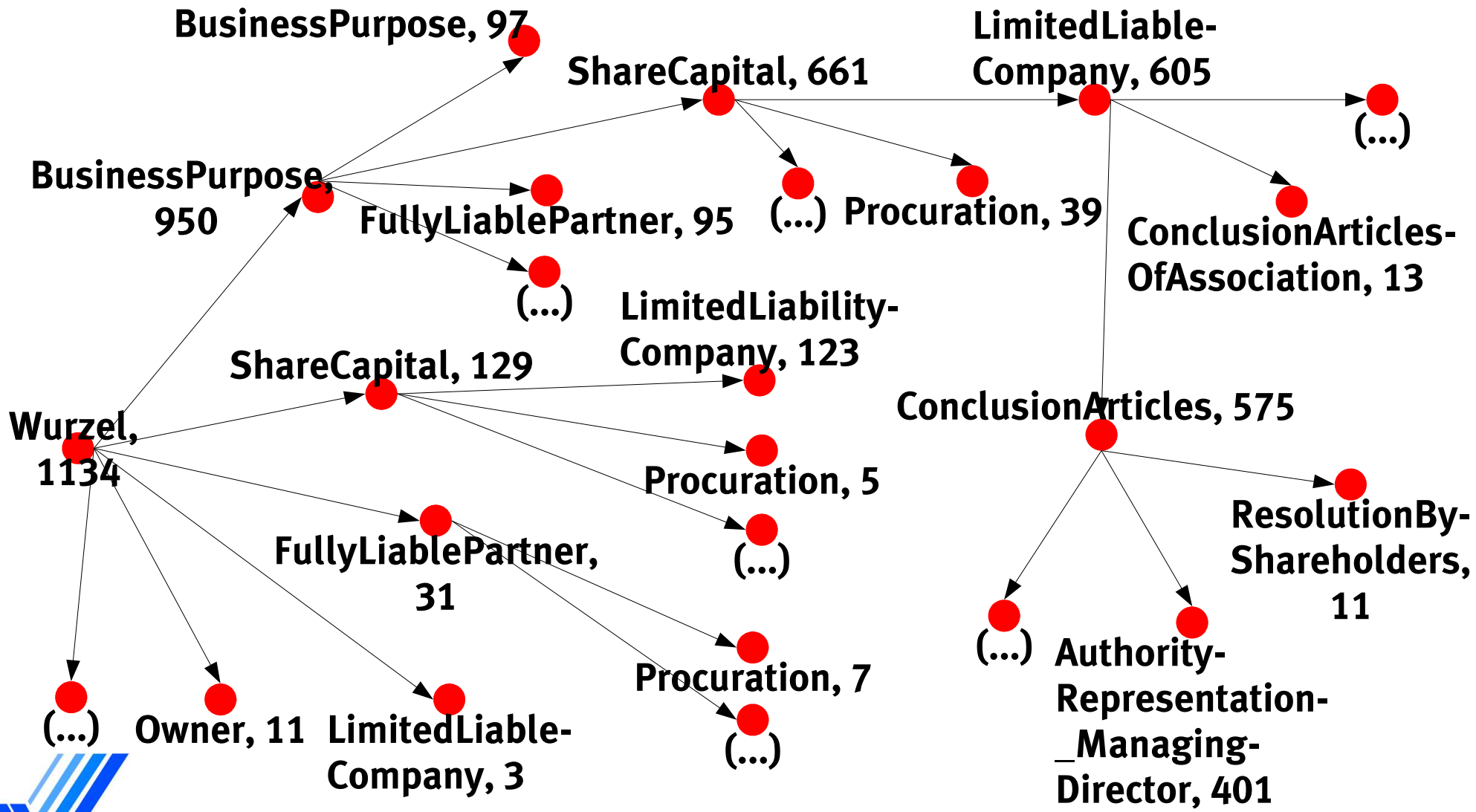




Alternativenbaum: Beschreibung

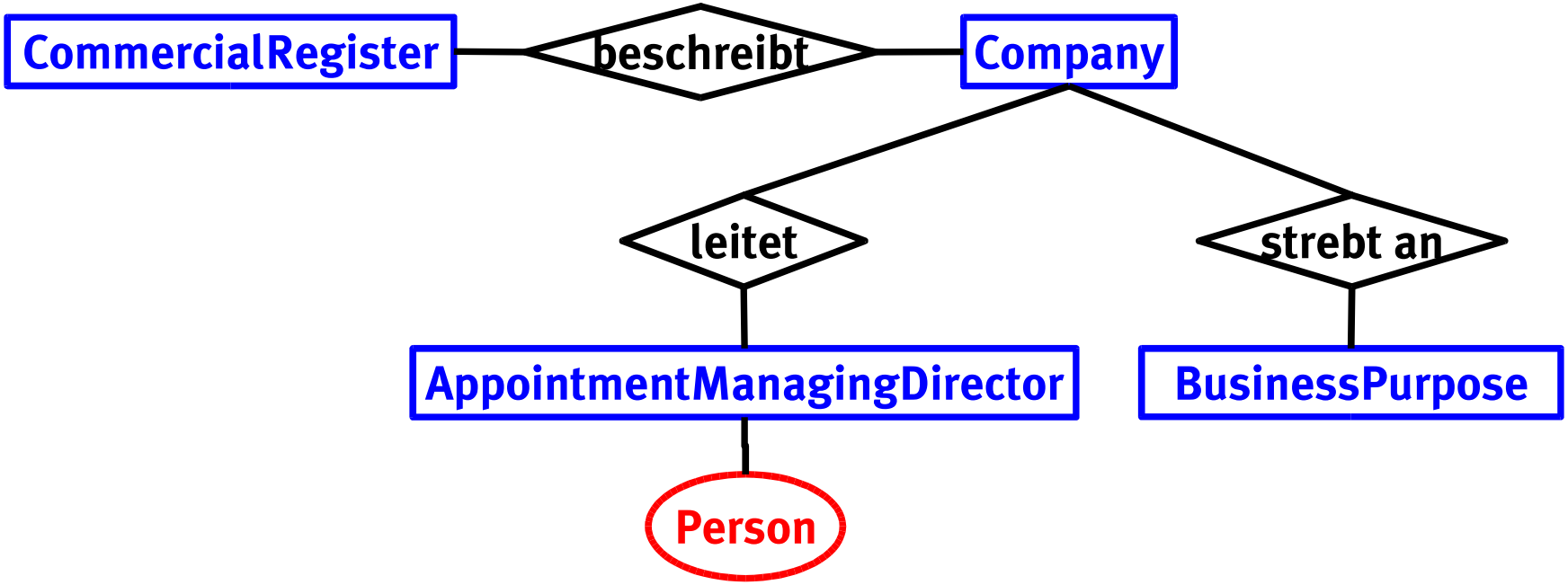
- ordnungserhaltende DTD: Baum von alternativen Tag-Folgen
- zu jedem Tag: Support in Bezug auf Folge, die zu diesem Tag führt
  - Anzahl Dokumente, die mit dieser Tag-Folge anfangen
- jeder Tag kann mehr als einmal vorkommen
- beschneiden (gleiche Kriterien wie eben)
- nur Sequence... und GroupSupport

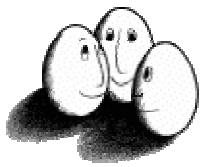
Alternativenbaum: Beispiel

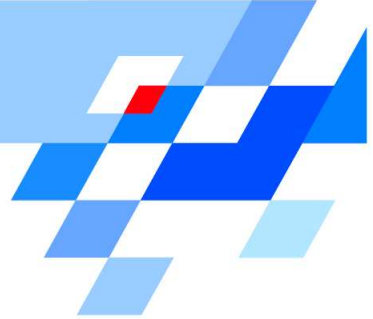


Domäne: Einträge ins Handelsregister

<BusinessPurpose>Der Betrieb von Spielhallen in Teltow und das Aufstellen von Geldspiel- und Unterhaltungsautomaten.</BusinessPurpose>
<AppointmentManagingDirector Person="Balski; Pawel"> Pawel Balski ist zum Geschäftsführer bestellt. </AppointmentManagingDirector>

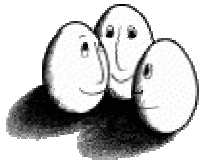


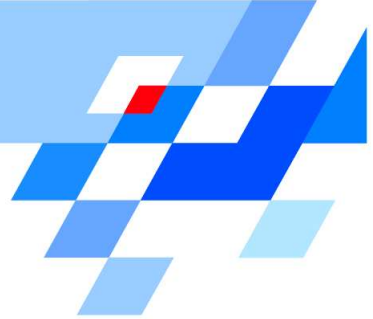




Probleme/offene Fragen

- Wie gut funktioniert das?
  - Autoren arbeiten an Methoden, das zu bewerten
- Möglichkeit für geschachtelte Tags fehlt
  - hierarchischer Clustering-Algorithmus für DIAsDEM geplant
- Schritt probabilistische DTD → Schema
- für viele Schritte Experte nötig





Vielen Dank für Eure Aufmerksamkeit!

- Noch Fragen?
  andrea.schweer@uni-dortmund.de