Freie Universität Berlin

Seminar Web Technologien
Information Extraction

WS 11/12

Prof. Dr.-Ing. Robert Tolksdorf
Freie Universität Berlin
Institut für Informatik
Netzbasierte Informationssysteme
mailto: tolk@ag-nbi.de
http://www.robert-tolksdorf.de

Unstrukturierte Daten

Strukturierte Daten
Verarbeitbar

Mark Rothko
Retrospektive
8. Februar – 27. April 2008

In welchem Jahr wurde der Maler Mark Rothko geboren?

Mark Rothko
Nr. 14, 1951
Öl/Leinwand, 143,5 x 165,1 cm
Privatsammlung

Mark Rothko ist einer der bedeutendsten amerikanischen Künstler des 20. Jahrhunderts. Bekannt sind seine meist großformatigen Gemälde mit horizontal geschichteten Farbflächen. Solche meditativen Abstraktionen gelten heute als Synonyme für den Abstrakten Expressionismus. Unter dem Schlagwort New York School übernimmt die amerikanische Kunst nach dem Zweiten Weltkrieg analog zur Vormachtstellung der Vereinigten Staaten im politischen und wirtschaftlichen Bereich mit diesem Stil auch die Führung in der Welt der Kunst.

Rothko hat sich jedoch zeitlebens dagegen gewehrt, als Maler abstrakter Bilder vereinnahmt zu werden. 1903 als Marcus Rothkowitz in Russland geboren, kommt er als Zehnjähriger mit seiner Familie in die Vereinigten Staaten. Nach Studien in Yale und an der New School of Design in New York, beginnt er ab 1930 als Künstler zu arbeiten. Ganz allmählich tastet er sich von figurativen Anfängen über den Surrealismus zu reinen Farbkonstellationen heran. Man erkennt in seinen frühen städtischen Szenen bereits Experimente mit formalen Reduktionen und es wird deutlich, wie sich seine surreal-biomorphen Formen zu immer kompakteren Farbwolken, so genannten „multiforms“ bündeln. Diese markieren Ende der 1940er Jahre den Übergang zu seinem genuinen Bildschema, den sich überlagernden Farbschleiern, die durch das Durchscheinen der Schichten zu räumlicher Wirkung gebracht werden. Wenn auch von allem Abbildlichen befreit, verbindet der Künstler seine Bilder weiter mit inhaltlichen Vorstellungen.

Auch vollkommen ungegenständliche Gemälde will er daher stets als etwas Konkretes verstanden wissen. Seine aus dem Rechteck entwickelten Formgefüge sorgen durch ihre Proportionen zueinander

Union List of Artist Names® Online
Full Record Display

New Search | Help

Click the icon to view the hierarchy.

ID: 500014869 **Record Type: Person**

Rothko, Mark (American painter, 1903-1970)

Note: Noted as one of the primary artists of Abstract Expressionism and color field painting.

Names:
Rothko, Mark (**preferred**, **index**, V)
Mark Rothko (**display**, V)
Rothkowitz, Marcus (V)

Nationalities:
American (**preferred**)

Roles:
artist (**preferred**)
painter

Gender: male

Birth and Death Places:
Born: Daugavpils (Daugavpils district, Latvia)
Died: New York City (New York state, United States)

Related People or Corporate Bodies:
parent of .... Rothko, Kate
..... (American, born 1950) [500069309]

List/Hierarchical Position:
.... Person
..... Rothko, Mark

Biographies:
(American painter, 1903-1970) ..... [VP Preferred]
(American painter, 1903-1970) ..... [BHA]
(American painter, 1903-1970) ..... [GRLPSC]
(American artist, 1903-1970) ..... [WCP]
(American artist, 1903-1970) ..... [WCI]

Freie Universität Berlin

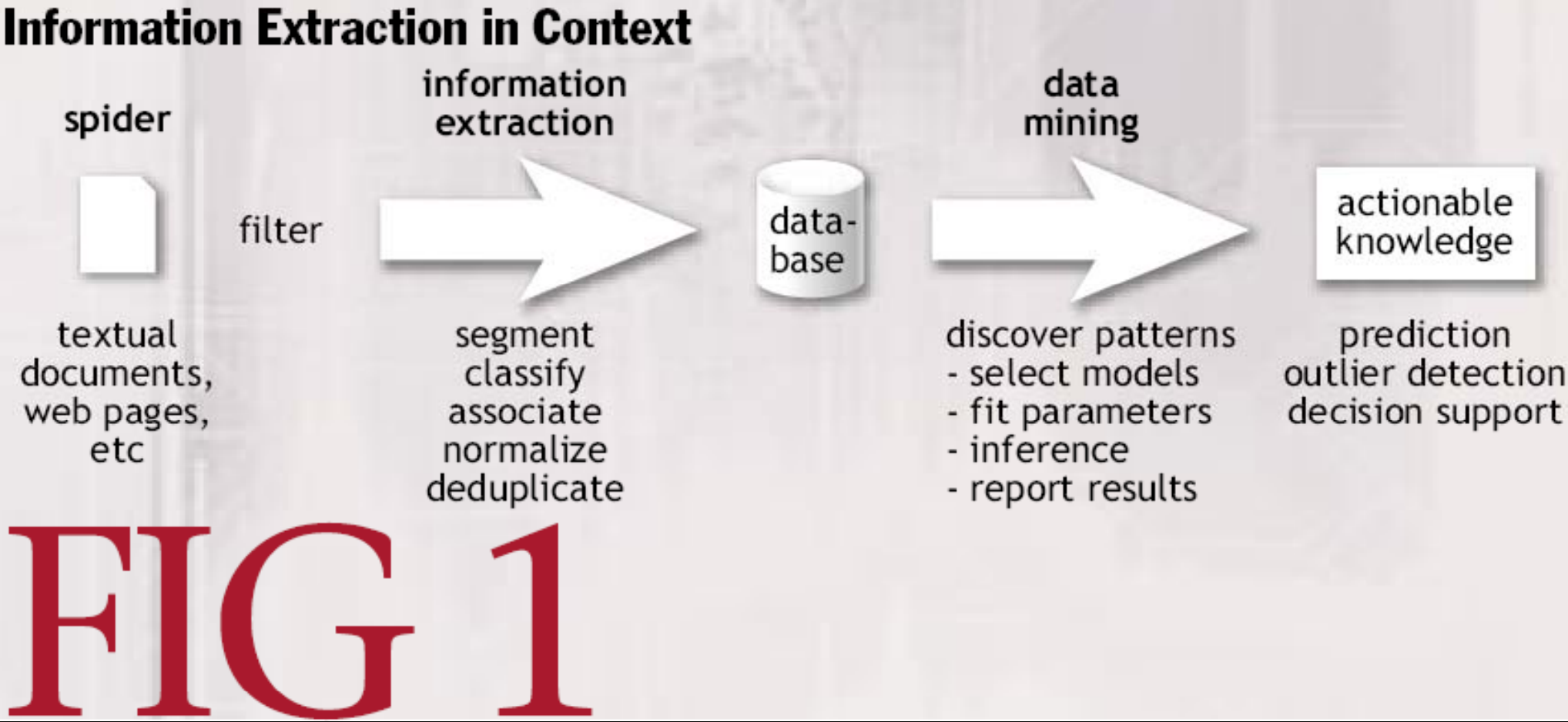
Information extraction

```
<table>
<tr>
<td valign="bottom">
<img align="right" src="rothko/a 460-300.jpg" width="210" he
</td>
<td ><img src="weiss.gif" width="5" height="9" border="0" al
<td valign="bottom">
<font face="Arial" color="#666666">
Mark Rothko <br>
Nr. 14, 1951<br>
&Ouml;l/Leinwand, 143,5 x 165,1 cm 
Privatsammlung<br>
</td>
</tr>

<tr>
<td colspan="3"><img src="weiss.gif" width="21" height="28"
<tr>
<td valign="bottom">
<img align="right" src="rothko/a 402-300.jpg" width="180" he
</td>
<td ><img src="weiss.gif" width="5" height="9" border="0" al
<td valign="bottom">
```

Freie Universität Berlin

Web Information extraction



McCallum, A.: Information extraction: distilling structured data from unstructured text. ACM Queue, 3(9), 48-57. 2005.

Struktur des Seminars

- Bis Weihnachten: Theoretische Grundlagen
- 6 Referate

- Quellen:
  - McCallum, A.: Information extraction: distilling structured data from unstructured text. ACM Queue, 3(9), 48–57. 2005.
  - Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled F. Shaalan: A Survey of Web Information Extraction Systems. IEEE Transactions on Knowledge and Data Engineering, Volume 18 Issue 10, October 2006.
- Eigene Literaturrecherche ist mandatorisch!

- Themenvergabe heute

Struktur des Seminars

Freie Universität Berlin

- Im nächsten Jahr: Praktische Information Extraction, 7 Wochen
- Mehrere Teams treten gegeneinander an

- Aufgabe:
- Das Getty Research Institute bietet mir der Union List of Artist Names® eine Sammlung von Namen von Künstlern online an. Diese Namenssammlung ist urheberrechtlich geschützt und kann lizensiert werden. Man allerdings davon ausgehen, dass sich sämtliche Namen auch im freie Web finden lassen, allerdings eben nicht als qualitätsgesicherte Liste. Mit Hilfe von Web Information Extraction könnte man versuchen, diese Namen automatisiert aufzufinden und selber zu sammeln.
- Die Teilnehmer sollen dies versuchen indem sie gemeinsam eine Software schreiben, die ausgehend von der Liste der Sammlungen moderner oder zeitgenössischer Kunst bei Wikipedia versucht, die darüber verlinkten Sites die Namen der in den Sammlungen vertretenen Künstlern zu extrahieren.

Freie Universität Berlin

Themen

19.10.2011	Einführung und Themenvergabe	Tolksdorf
26.10.2011	Hinweise zur Gestaltung von Referaten	Tolksdorf
02.11.2011	Referat Überblick und Beispiele auf Basis McCallum2005	Dinh, Beraki, Jung
09.11.2011	Survey Abschnitt 1 bis 3 – Klassifikationsmöglichkeiten für IE	Große, Starroske, Schulz, Saenz
16.11.2011	Survey Abschnitt 4.1 – Handgefertigte Extraktoren	Rotar, Schröder, Dräger, Bischoff
23.11.2011	Survey Abschnitt 4.2 – Überwachte Extraktoren	Hermann, Kahl, Do, Schellenber, Bruns,
30.11.2011	Survey Abschnitt 4.3 und 4.4 – Halbüberwachte und unüberwachte Extraktoren	Sidykh, Hasan, Dahlke Bergunde,
07.12.2011	Survey Survey Abschnitt 5 – Vergleich	Wei, Agri, Siripanya
14.12.2011	Planung Projektarbeit, Getting Things Done	

04.01.2012	Projektarbeit Kickoff und Organisation	
11.01.2012	Projektarbeit Planungsstand und Status	
18.01.2012	Projektarbeit Planung für Zwischenmeilenstein	
25.01.2012	Projektarbeit Zwischenmeilenstein	
01.02.2012	Projektarbeit Planungsstand und Status	
08.02.2012	Projektarbeit Planung Abschlusspräsentation	
15.02.2012	Projektarbeit Abschlusspräsentation	

Freie Universität Berlin

Leistungen

- **Leistungsnachweis**
  - **Referatsbeitrag**
  - **Ausarbeitung**
  - *Eigene Literaturarbeit ist verpflichtend!*
  - **Mitarbeit an Softwareentwicklung**

- **Bitte beachten Sie**
  - **Hinweise zu Ablauf und Leistungserbringung in Seminaren bei NBI http://www.ag-nbi.de/lehre/seminare.html**
  - **Die Hinweise zu Plagiaten http://www.ag-nbi.de/lehre/tipps/plagiate.html**

- **In Anlehnung an den Artikel Stefan Weber: Wissenschaft als Web-Sampling. Telepolis. 15.12.2006 . (http://www.heise.de/tp/r4/artikel/24/24221/1.html) legen wir für die Ausarbeitungen fest:**
  - **Direkte Zitate aus dem Internet nie zur Faktenvermittlung, sondern nur noch als illustrative Beispiele, wenn also das Zitat selbst thematisiert wird**
  - **Keine Zitate von der Wikipedia, außer zur kritischen Kommentierung**

- **Neben den Vorgaben *müssen* Sie weitere Literaturquelle verwenden**

Ablauf

- **Erster Termin:**
  - **Themen- und Terminvergabe**

- **Zwei Wochen vor Referatstermin:**
  - **Entwurf des Foliensatzes wird an Veranstalter geschickt und ein Termin für eine Vorbesprechung vereinbart (für die allerersten Referatstermine werden jeweils Sonderregelungen abgesprochen).**
  - *Ohne Vorliegen des Entwurfs und ohne Vorbesprechung muss das Referat ausfallen und es kann kein Schein erteilt werden*.

- **Referatstermin:**
  - **Referat :-)**

- **Ende der Vorlesungszeit:**
  - **Abgabe der Ausarbeitung**

Ablauf

- Mit der Themenvergabe kann die Arbeit am Referat und der Ausarbeitung beginnen.
- Gerade bei späten Referatsterminen sollte man keine Zeit verlieren, da dann der Abstand zwischen Referat und Ausarbeitung sehr kurz ist.
- Die Abgabe der Ausarbeitungen findet zum Ende der letzten Woche der Vorlesungszeit statt.
- Man sollte eventuelle Zusatzbelastungen durch Klausuren etc. am Ende der Vorlesungszeit durch rechtzeitige Fertigstellung der Ausarbeitung auffangen.

Referate

- Die einzelnen Seminartermine dauern 90 Minuten und sind für ein Thema reserviert. Ein möglicher Zeitplan ist
  - 5 Minuten Einleitung und Einordnung des Themas durch Veranstalter
  - 75 Minuten Referat einschließlich Nachfragen und Diskussion
  - 10 min Feedback durch die Teilnehmer zur Referatsgestaltung
- Bitte teilen Sie vorher dem Veranstalter mit, ob Sie Notebook, Beamer und/oder Overhead etc. benötigen. Bitte schicken Sie nach dem Referat Ihre Folien als PDF an den Veranstalter, damit sie ins Netz gehängt werden können. Falls Sie Ihre Folien nicht veröffentlicht sehen wollen - auch ok.

Ausarbeitung

- Die Ausarbeitung stellt den Inhalt des Referats als zusammenhängenden wissenschaftlichen Text dar. Mit ihm soll es jemanden, der nicht das Referat gehört hat, möglich sein, sich das behandelte Thema zu erschließen.

- Die Ausarbeit soll in der Regel einen Umfang von 5 Seiten pro Person haben. Mit "Seite" ist dabei eine handelsübliche Seite gemeint, also nicht in 12 Punkt Schrift mit riesigen Rändern. Es gibt keine weiteren Formatvorgaben, da es ja um den Gehalt der Ausarbeitung geht.

- Die Ausarbeitungen bitte unter Beachtung der Hinweise zu den Präsentations- und Ausarbeitungstechniken erstellen

- Elektronisch als PDF per Mail an Veranstalter schicken

Freie Universität Berlin

Fragen

Themenvergabe