Content Mining

1. Semantic Text Mining
2. Content Based Retrieval
3. Visualisierung als Schlüssel

1 Semantic Text Mining

- Technologie und Datenquellen
- MindNet

07. März 2007 Drucken | Senden | Leserbrief | Bookmark

BUCH-DIGITALISIERUNG Schrift:

Google kooperiert mit Bayerischer Staatsbibliothek

Google und die Bayerische Staatsbibliothek in München haben eine Vereinbarung verkündet: Der gewaltige Bestand der Münchner wird digitalisiert und mit dem Google-Werkzeug Booksearch durchsuchbar und online lesbar gemacht.

03. Mai 2007 Drucken | Senden | Leserbrief | Bookmark

WIKIPEDIA-GRÜNDER WALES Schrift:

"Google ist nicht mehr allen überlegen"

Wikipedia-Gründer Jim Wales startet den Angriff auf Google. Im SPIEGEL-ONLINE-Interview verrät er seinen Plan für eine neuartige Suchmaschine - und verteidigt sich gegen den Vorwurf, dass Mitmach-Enzyklopädien notorisch fehlerhaft sind.

14. März 2007 Drucken | Senden | Leserbrief | Bookmark

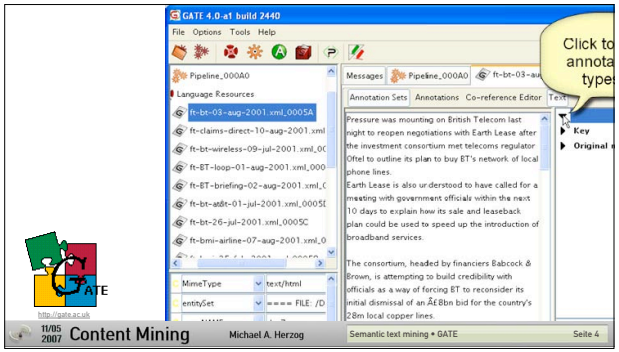
INTERNET-KRIMI Schrift:

David gegen Google

Aus San Francisco berichtet Marc Pitzke

In San Francisco bahnt sich ein Dotcom-Krimi an: Das kleine Internet-Startup Powerset, das noch weitgehend geheim arbeitet, will mit einer neuen Such-Technologie Google vom Thron stürzen. Erste Investoren sind begeistert.

SPIEGEL ONLINE



WORTSCHATZ UNIVERSITÄT LEIPZIG

Wort: Bank Suche! ? ☐ Beachte Groß-/Kleinschreibung

Wort: Bank
Anzahl: 47819
Häufigkeitsklasse: 8 (d.h. *der* ist ca. 2^8 mal häufiger als das gesuchte Wort)
Sachgebiet: Nachname
Morphologie: bank
Grammatikangaben: Wortart: Eigenname
Relationen zu anderen Wörtern:

- Synonyme: Bankhaus, Eckbank, Geldinstitut, Kasse, Kreditanstalt, Kreditbank, Kreditinstitut, Sandbank, Sitzbank, Sparkasse, Untiefe, Wechselstube, alle

Links zu anderen Wörtern:

- falls positiv bewertet Vollbank, Spezialbank, Superbank, Oberbank, Hauptbank, Spitzenbank, Sonderbank, Profibank
- Grundform: Bank
- Antonym von: Nichtbank
- -lich-Form von: banklich

- falls positiv bewertet Vollbank, Spezialbank, Superbank, Oberbank, Hauptbank, Spitzenbank, Sonderbank, Profibank
- Grundform: Bank
- Antonym von: Nichtbank
- -lich-Form von: banklich
- Teilwort von: Deutsche Bank, Dresdner Bank, Deutschen Bank, Berliner Bank, Bank Austria, Direkt Anlage Bank, Bank von England, Deutsche Bank AG, Deutscher Bank, Bank von Japan, Dresdner Bank AG, BfG Bank, Advance Bank, Bank of England, Deutschen Bank AG, Bank of Japan, Bank of America, Deutsche Bank Research, Federal Reserve Bank, Bank von Frankreich, Bank of New York, ABN Amro Bank, Barclays Bank, BfG Bank AG, West Bank, Europäische Bank für Wiederaufbau und Entwicklung, Europäischen Bank für Wiederaufbau und Entwicklung, Royal Bank of Scotland, Barings Bank, Chase Manhattan Bank, Bank of Scotland, Bank of Tokyo, Bank of China, Daiwa Bank, Credit Bank, Mitsubishi Bank, National Westminster Bank, Nippon Credit Bank, National Bank, Dai-Ichi Kangyo Bank, Baring Bank, Bank Austria AG, Bank von Spanien, Advance Bank AG, Deutsche Bank Bauspar AG, West Merchant Bank, Danske Bank, Bank von Italien, Santander Direkt Bank, Credit Bank of Japan, Midland Bank
- Form(en): Bank, Banken, Bänke, Bänken, Banks, Bankes, Bankn
- Abkürzung: Bk.
- Untereinträge: Bank (Geldinstitut), Bank (Sandbank), Bank (Sitzmöbel), Bank (Spielbank), Bank (Werkbank)
- Unterbegriffe: Deutsche Bank, Bundesbank, Dresdner Bank, Deutschen Bank, Commerzbank, Weltbank, Landesbank, Zentralbank, Notenbank, Datenbank, Postbank, HypoVereinsbank, Europäische Zentralbank, Volksbank, Investmentbank, Anklagebank, Hypo-Vereinsbank, US-Notenbank, Großbank, BHF-Bank, Berliner Bank, Spielbank, Deutschen Bundesbank, Citibank, Ersatzbank, Investitionsbank, Vereinsbank, Privatbank, Landeszentralbank, Nationalbank, Hausbank, Hypo-Bank, Hypothekenbank, Schulbank, Job-Datenbank, Westbank, Vereins- und Westbank, WGZ-Bank, Trainerbank, Bayerische Vereinsbank, Direktbank, Westdeutschen Landesbank, Westdeutsche Landesbank, Bayerischen Vereinsbank, Grundkreditbank, Staatsbank, Rückbank, Parkbank, Direkt Anlage Bank, Hypovereinsbank, Ausgleichsbank

Dornseiff-Bedeutungsgruppen:

- 4.18 Material, Vorrat: Akku, Arsenal, Aufbewahrungsort, Bank, Behälter, Boiler, Depot, Fundgrube, Fundus, Geheimfach, Kollektion, Konto, Kornspeicher, Lager, Lagerhaus, Magazin, Menagerie, Museum, Sammelstelle
- 15.11 Genossenschaft: Aktiengesellschaft, Bank, Firma, Gesellschaft, Handelsgesellschaft, Holding, Kartell, Kommanditgesellschaft, Kooperative, Personengesellschaft, Ring, Trust, Zweckverband, volkseigener Betrieb
- 19.3 Möbel: Bank, Hängematte, Sitz, Sitzbank, Sänfte, Thron
- 20.18 Verleihen: Bank, Kreditanstalt, Leihhaus, Pfandhaus, Pfandleihe
- 20.35 Bankwesen: Außenhandelsbank, Bank, Bankfiliale, Bankhaus, Bankinstitut, Bankkonzern, Depotbank, Direktbank, Direktbanktochter, Effektenbank, Filialbank, Finanzkonzern, Geldhaus, Geldinstitut, Geldinstitut, Genossenschaftsbank, Geschäftsbank, Girokasse, Girozentrale

Beispiel(e):
Im Zuge der Neuausrichtung als starke Regionalbank würden zunächst die Standorte im übrigen Bundesgebiet überprüft, hieß es von Seiten der **Bank**. (Quelle: *Der Spiegel ONLINE*)
Das kam die **Bank** teuer zu stehen (siehe Kasten). (Quelle: *Der Spiegel ONLINE*)
Wenn also die Zinsen für das Baugeld steigen, dann bekommen ganz normale Sparer auch mehr Zinsen für das Geld, das sie monatlich zur **Bank** tragen. (Quelle: *Der Spiegel ONLINE*)
weitere Beispiele

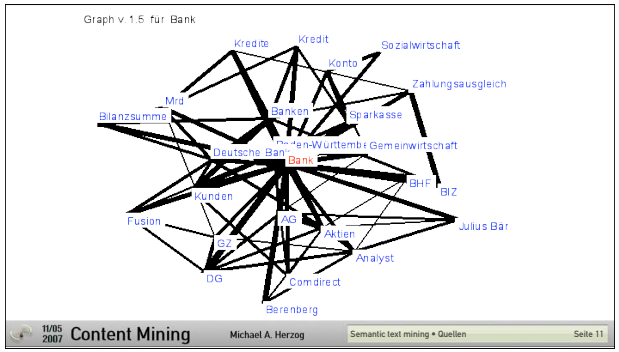
Signifikante Kookkurrenzen für Bank:
DG (7932), einer (1298), Kunden (1261), Berenberg (1103), Zahlungsausgleich (1002), eine (899), Euro (868), bei (747), Kredit (713), Prozent (694), Bilanzsumme (671), BHF (666), DM (662), Milliarden (561), BIZ (545), Millionen (541), Kredite (533), AG (508), Sparkasse (478), Baden-Württembergischen (470), Fusion (469), Aktien (456), Gemeinwirtschaft (444), Mrd (422), Banken (411), Comdirect (399), Sozialwirtschaft (399), Konto (398), Analyst (395), GZ (384), Geld (369), Mark (366), Frankfurt (363), Commerzbank (362), Risikovorsorge (352), Kunde (351), Aktie (339), größte (331), Internationalen (327), UBS (316), teilte (307), of (296), Filialen (293), Vorstandssprecher (286), überfallen (286), Baden-Württembergische (280)

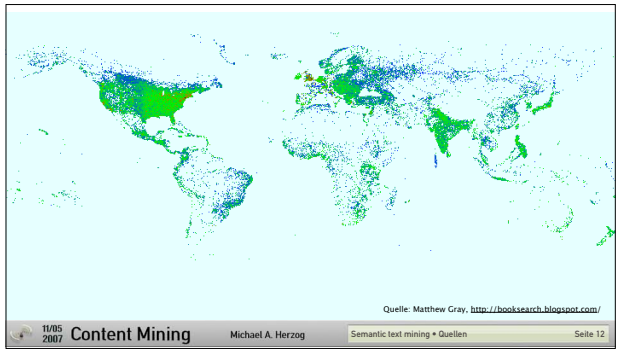
Mehrwortkookkurenzen:
Julius Bär (445), Deutsche Bank (384), Deutschen Bank (333), Dresdner Bank (278)

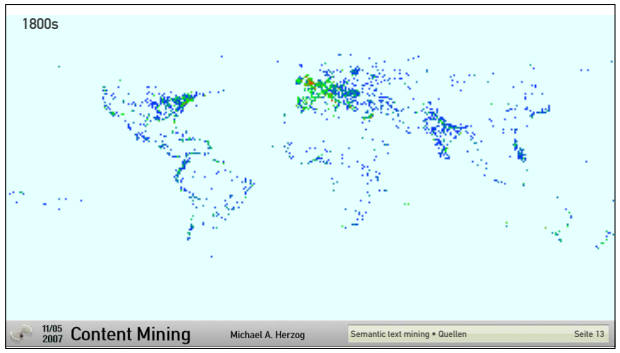
Signifikante linke Nachbarn von Bank:
DG (7370), Berenberg (1041), BHF (616), Baden-Württembergischen (460), Comdirect (296), Baden-Württembergische (253), Köpenicker (245), sichere (216), GZ (211), DSL (169), NordFinanz (160), Fuji (130), Sakura (129), Hinrich Donner (125), Development (117), Investment (106), SGZ (103), Schweizer (102), Aareal (100), Sumitomo (94), Tokai (90), Sanwa (89), Dresdener (89), LGT (87), Commercial (80), Jesper (71), Deka (71)

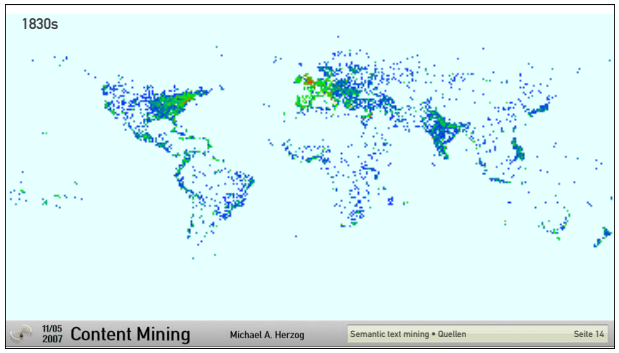
Signifikante rechte Nachbarn von Bank:

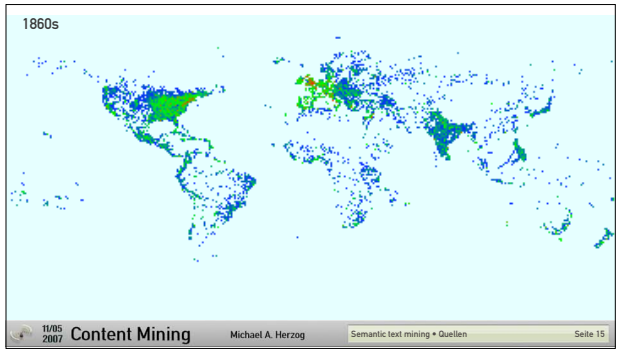
Julius Bär (426), Sarasin (191), of (164), Deutsche Genossenschaftsbank (138), Leu (130), Vontobel (127), AG (117), überfallen (100), ABN Amro (88), One (80), Gdanski (68), Paribas (66), direct (62), Handlowy (62), BNP Paribas (62), GiroTel (58), Crédit Lyonnais (57), Credit Lyonnais (56), UBS (55), Austria-Creditanstalt (49), Girotel (47), BW-Bank (46), eG (45), Slaski (43), verbannt (42), Credit Suisse (41), Menatep (40), Sal (39), Société Générale (38), KDB (34), GKB (33), Santander (30), Austria-Gruppe (30), Bernd Thiemann (29), Pekao (28), ausrauben (25), Ltd (24), Hapoalim (23), empfiehlt (22), Friedrich Steil (22), auszurauben (20), Centre (18), teilte mit (17), Mizuho (17), Leumi (17), HSBC (17), HBOS (17), Caisse (17), San Paolo (16), KEB (15), Crédit Suisse (15), Bipop-Carire (15), verbannte (14), rät (14), Sal Oppenheim (14), SBS-Agro (14), Przemyslowo-Handlowy (14), Michael Heise (14), ausgeraubt (13), UOB (13), Labouchere (13), Julius Baer (13), Credit (13), deponiert (12), Unicredito (12), Morgan Stanley (12), Kreiss (12), HSBC Trinkaus (12), Asset Management (12), UBS Warburg (11), Rozwoju (11), Internasional (11), Hofmann (11), Garantija (11), DnB (11), Credit Agricole (11), Caisse des Depots et Consignations (11), CIB (11), ABN-AMRO (11), gestürmt (10)

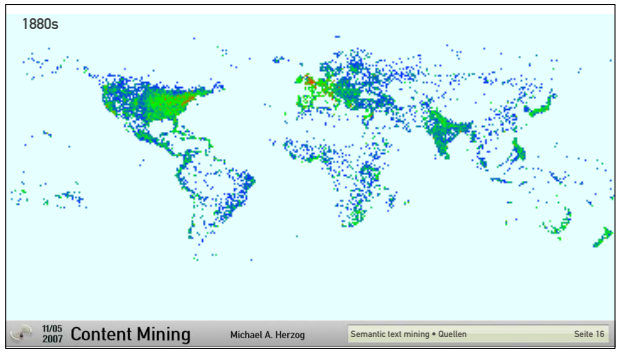


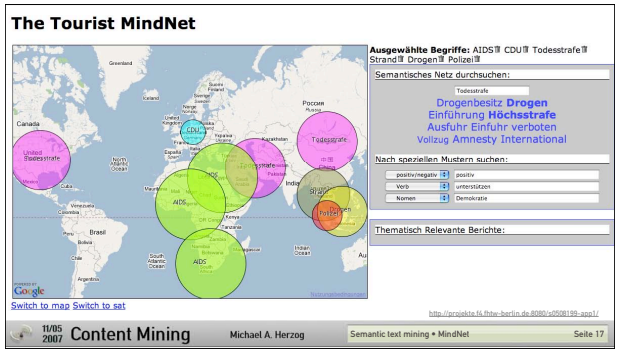


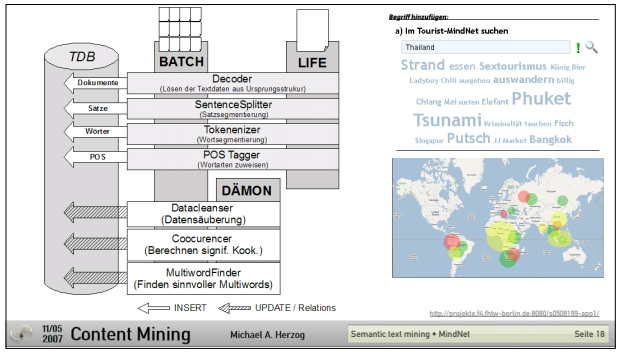












2 Content Based Retrieval

- Überblick und Beispiele
- MOCCA Media Repository

Problem

- Flut an digital verfügbaren Multimediadaten
  - Digitalfotos, Vektorgrafiken, XML-Dokumente, Filme, Audio-Daten, ...
- Problem: effektive inhaltsbasierte Suche
  - Bsp: Suche Fotos von einem Eisberg
  - Bsp: Suche Fotos vom TU-Hauptgebäude
  - Bsp: Suche Audioaufnahmen mit dem Sprecher „Matthias Trier"
- Datenbankanfrage: SELECT ... FROM ... WHERE ...
  auf Grund fehlender Attributwerte ungeeignet

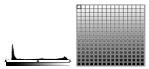
Bild: IBM Marvel, www.research.ibm.com/marvel/
Literatur: Ingo Schmitt. Ähnlichkeitssuche in Multimedia-Datenbanken. Retrieval, Suchalgorithmen und Anfragebehandlung. Oldenbourg 9/2005

Feature-Extraktion

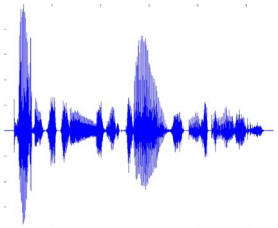
- High level feature:
  Eisberg, Fisch, Schloss Reichenow, Klaus Rebensburg

semantische Lücke

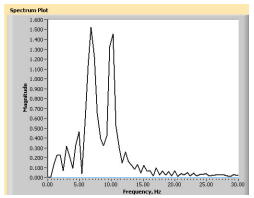
- Low level feature:
  Farbverteilung, Textur, Form



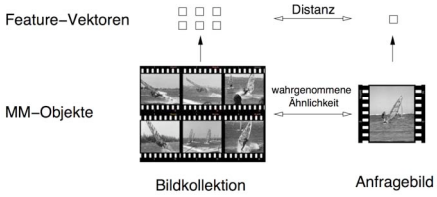
Bsp: AUDIO Merkmalsextraktion



DCT
STFT
DWT

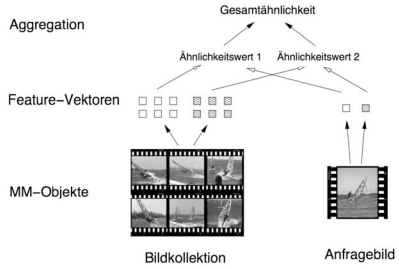


Feature Extraktion



Quelle: Ingo Schmitt. Ähnlichkeitssuche in Multimedia-Datenbanken. Retrieval, Suchalgorithmen und Anfragebehandlung. Oldenbourg 9/2005

Aggregation von Ähnlichkeitswerten



Quelle: Ingo Schmitt. Ähnlichkeitssuche in Multimedia-Datenbanken. Retrieval, Suchalgorithmen und Anfragebehandlung. Oldenbourg 9/2005

MPEG-7 Format

```
<AudioDescriptor hiEdge="16000.0" loEdge="62.5" octaveResolution="1/0"
xsi:type="AudioSpectrumBasisType">
  <SeriesOfVector hopSize="PT10N1000F" totalNumOfSamples="272" vectorSize="8">
    <Raw mpeg7:dim="1 34 8">0.15732187 -0.10239355 0.22149466 -0.071965046 0.14958718
    -0.09177902 0.050023418 -0.22242463
    </Raw>
  </SeriesOfVector>
</AudioDescriptor>
```


MEDIA RETRIEVAL • BEISPIELE

Findr for flickr
- Tag Explorer with live Preview

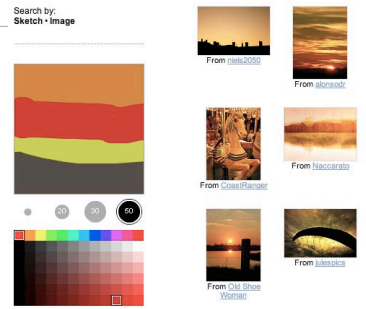
Riya
- Face recognition

Retrievr
- Bildsuche basierend auf Zeichnung

Tiltomo, Yotophoto, Picturefinder
- Visual search, MPEG 7

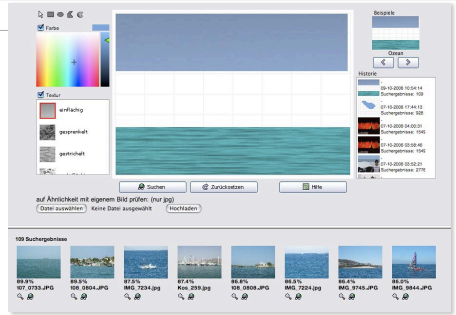
Marvel
- MPEG 7 Video Search Engine

retrievr for flickr

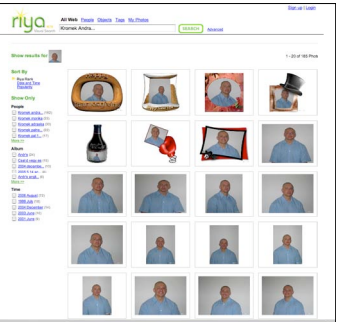


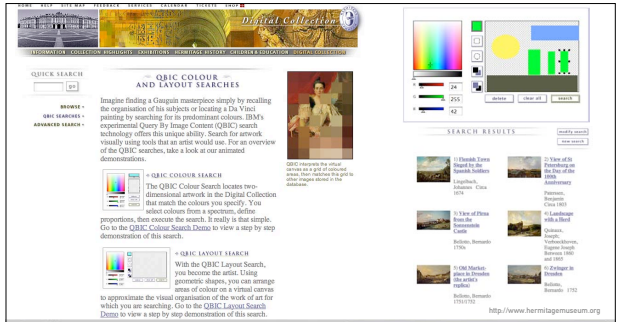
PictureFinder

- Uni Bremen



Riya face recognition





Tiltomo visual search

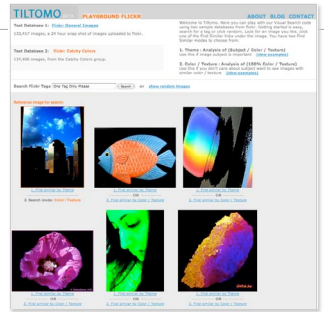
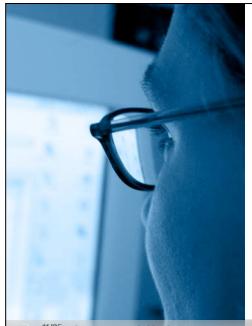


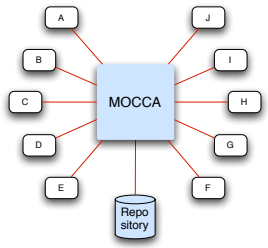
Photo Tourism

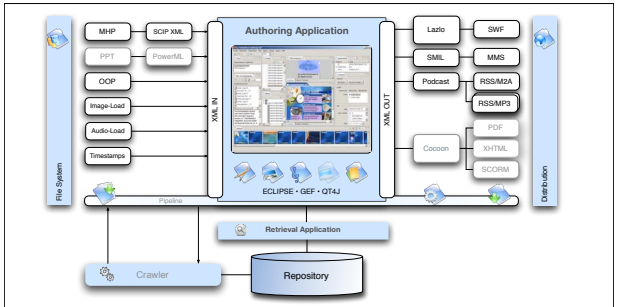
Exploring photo collections in 3D

Noah Snavely Steven M. Seitz Richard Szeliski
University of Washington *Microsoft Research*

SIGGRAPH 2006





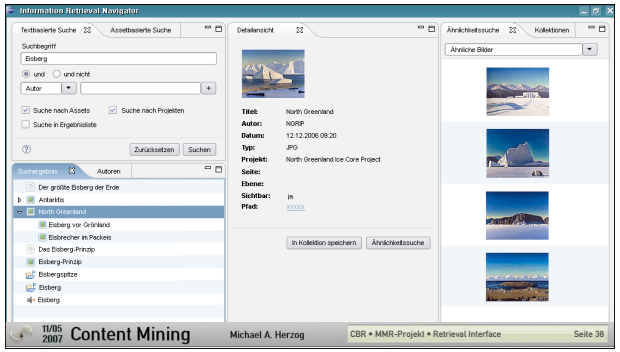


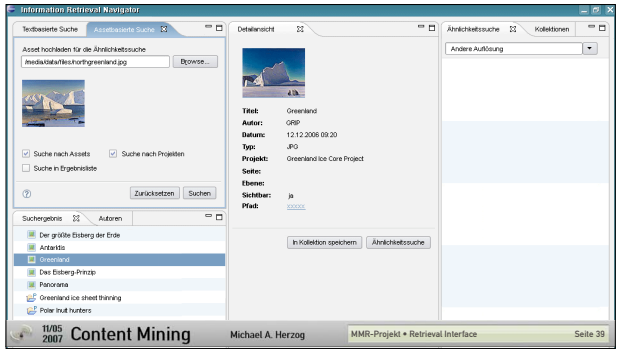
MOCCA-Repository-Konzepte

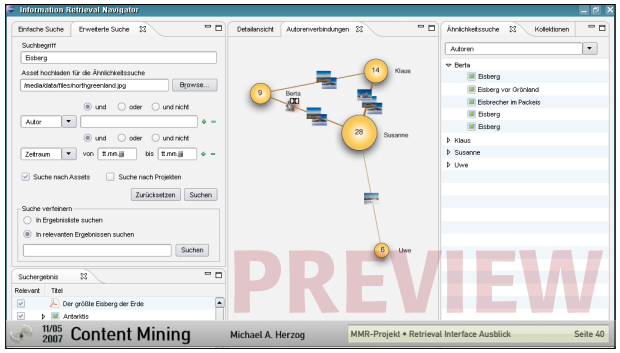
- Suche nach Assets und Projektdaten
  - Nutzung der Volltext-Daten aus Authoring-Projekten
  - Basis: Generische XML Datenstruktur
  - Vollautomatische Transcodierung in das Enterprise Repository (Crawler)
- CBR für das Auffinden von
  - Asset-Varianten (andere Auflösung, Bildausschnitt)
  - Ähnliche Bild- und Tondateien

CBR Basistechnologien

- apache LUCENE
  - Open-Source-Java-Bibliothek zum Erzeugen und Durchsuchen von Text-Indizes.
  - Volltextsuchen für beliebige Textinhalte
  - Hohe Performanz und Skalierbarkeit
- LIRE
  - Lucene Image REtrieval
  - Universität Graz, Know-Center
  - Implementiert verschiedene MPEG-7 Methoden
    - ScalableColor, ColorLayout and EdgeHistogram

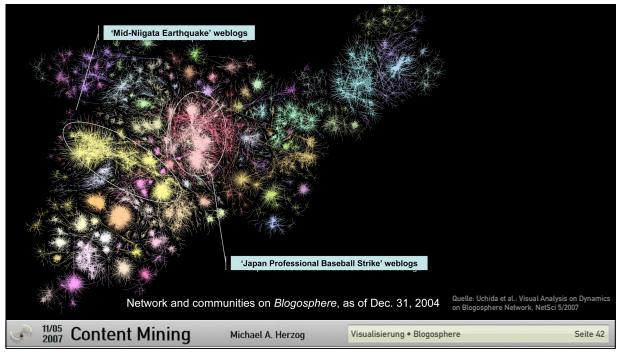


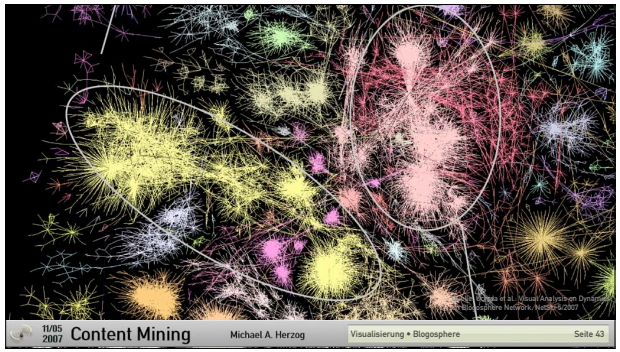




3 Visualisierung als Schlüssel

- Netzwerke und Geovisualisierung
- Commetrix-Projekt





Top 20 Wikipedia Most Actively Revised Articles (2001-2007)

1. Jesus
2. Adolf Hitler
3. October 2003
4. Nintendo revolution
5. Hurricane Katrina
6. RuneScape
7. India
8. Anarchism
9. Britney Spears
10. Playstation 3
11. Saddam Hussein
12. Japan
13. Albert Einstein
14. 2004 Indian Ocean Earthquake
15. New York City
16. Germany
17. Muhammad
18. Pope Benedikt XVI
19. Ronald Reagan
20. Hinduism

Quelle: Koblin, Flight patterns, NetSci 5/2007

COMMETRIX

application fields

- analyze social network dynamics
- map electronic communication
- search expert network maps
- find hidden communities
- find important actors
- observe merging networks

visit: www.commetrix.de

Content Mining

1. Semantic Text Mining
2. Content Based Retrieval
3. Visualisierung als Schlüssel