



Nutzung von Metainformation für die Navigation in (verteilten) Datenbeständen

Pascal Pein

1. Juni 2006



HAW HAMBURG





Einführung

Anwendungsszenarien

Referenzierung

Suche

Theorie

Informationen

Navigation und Suchanfragen

Architekturen

Ziele



Motivation

„Content that cannot be easily found is like content that does not exist, [...]. The easier it becomes to produce content, the faster the amount of content grows and the more complex the problem of managing content gets.“

Fernando Pereira, Rob Koenen

„[...] users of information systems do not need more information, but information which better corresponds to their interests and needs, i.e. information of higher quality.“

Hartmut Wittig



HAW HAMBURG



Einordnung

Verteilte Systeme enthalten...

- ▶ viele
- ▶ unstrukturierte
- ▶ flüchtige
- ▶ gemeinsam nutzbare

... Informationen

Wie kann man diese Informationen finden?





Feste Referenzen

- ▶ Hypertext
 - ▶ Texte mit Hyperlinks (HTML) stellen Beziehungen untereinander her.
 - ▶ Navigation zwischen diesen Seiten ist möglich
- ▶ Hypermedia (im engeren Sinne)
 - ▶ Hypertextvariante mit Schwerpunkt auf multimediale Daten
 - ▶ Sprünge an beliebige Stellen innerhalb einer Audiodatei, eines Videos, ...
 - ▶ Synchronisation verschiedener Multimediadokumente



Inhaltsbasierte Bildersuche

SEARCH max Result Size: 1.000 min. Similarity 0,3

Ranked results: 1.25, total results: 894

Query Image: Use DB Image

Load Image Load (DB) Paint Image

fv_stoch_quad 1 0
 $fv_stochastic$ 1 0
 $fv_keyword$ 1 0

ID: 489 0,58051	ID: 689 0,57822	ID: 688 0,55147	ID: 485 0,55000	ID: 424 0,54143	ID: 597 0,52903	ID: 653 0,52624
ID: 534 0,51698	ID: 487 0,51358	ID: 553 0,50909	ID: 589 0,50421	ID: 550 0,50336	ID: 675 0,50126	ID: 462 0,49909
ID: 484 0,49766	ID: 488 0,49328	ID: 764 0,48936	ID: 681 0,48922	ID: 657 0,48750	ID: 615 0,48726	ID: 238 0,48529
ID: 679 0,48367	ID: 646 0,48263	ID: 686 0,48242	ID: 1616 0,48184			

show sub rankings << Prev Next >> Random Set 25



Semantische Suche

Ontogator (Hyvönen) - Ontologien
TopicSEEK - TopicMaps

The screenshot displays the Ontogator search interface. The main window title is "Helsingin yliopiston museon promotionäyttely : :". The interface is divided into several sections:

- Query overview:** Located at the top left, it shows the current query and filters. A red box highlights the "Query overview" label.
- Selected picture:** A large central image of two men in suits, one holding a microphone. A red box highlights the "Selected picture" label.
- Recommendations:** Located on the right side, it displays a list of suggested images with titles and thumbnails. A red box highlights the "Recommendations" label.
- Query results:** Located on the left side, it shows a list of search results with thumbnails and titles. A red box highlights the "Query results" label.

Arrows point from the red boxes to their respective sections in the interface. The interface also includes navigation buttons at the bottom and a search bar at the top.



Kombinierte Suche und Navigation in verteilten Systemen

- ▶ Nutzung verschiedener Aspekte/Blickwinkel
 - ▶ Einschränkung des Suchraums
 - ▶ Filterung irrelevanter Daten
 - ▶ eigentlicher Typ der gesuchten Daten (Text, Bild, ...) zweitrangig
- ▶ Verteilte Suche über mehrere Systeme
 - ▶ Hyperlinks vergleichbar mit HTML
 - ▶ P2P Suchanfragen an Nachbarknoten (z.B. YaCy)
 - ▶ Verschmelzung der einzelnen Ergebnismengen

Was sind ähnliche Informationen?

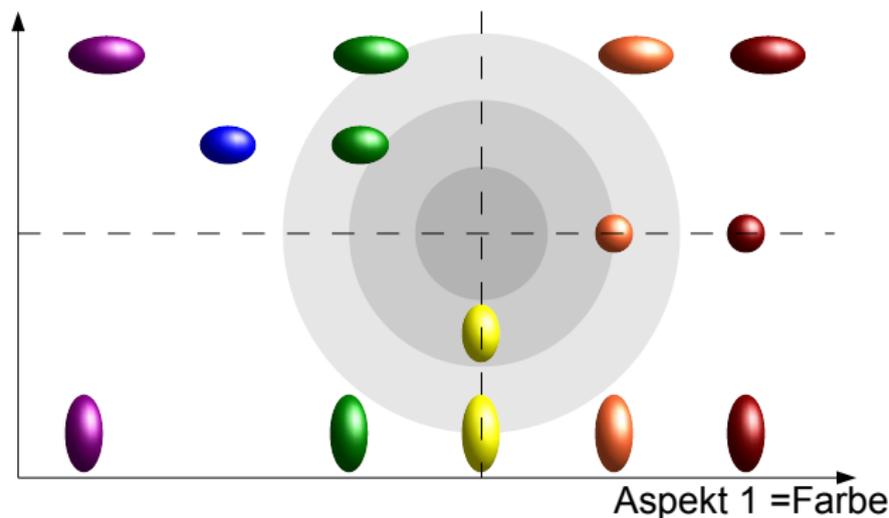
Ähnlichkeit kann auf verschiedene Weise definiert werden:

- ▶ Menschliche Wahrnehmung (subjektiv)
- ▶ Mathematische Distanzen (objektiv)
- ▶ Zugehörigkeit zu einer bestimmten Kategorie/Menge
- ▶ ...

Mehrdimensionale Suchräume

Beispiel: Suche nach einer gelben Kugel

Aspekt 2 = Form



Beispiele

- ▶ Schlüsselwörter
- ▶ Semantik
- ▶ Kategorien
- ▶ Erstellungsdatum/Zeitpunkt
- ▶ Histogramme (Bilder)
- ▶ Formen (Bilder)

3-Schichten-Modell für Bilder (J.P. Eakins, M.E. Graham):

Layer 1	primitives: color, texture, geometric shapes and spatial distribution
Layer 2	entities: object instances or individuals
Layer 3	abstract entities: activities, states, ...



Erfassung

- ▶ automatisch
 - + Für große Datemengen geeignet
 - + Objektiv
 - Nur für wenige Aspekte („Primitive“) nutzbar
- ▶ manuell
 - + Flexibel
 - + Hohe Qualität möglich (Spezialisten)
 - + Erfassung abstrakter Merkmale
 - Hoher Arbeitsaufwand
 - Subjektiv

Beschreibungsstandards

- ▶ textuell
 - ▶ Standard Generalized Markup Language (SGML)
 - ▶ Extensible Markup Language (XML)
 - ▶ Resource Description Framework (RDF)
- ▶ audiovisuell
 - ▶ Moving Picture Expert Group /
Multimedia Content Description Interface (MPEG-7)

Vorgehensweisen

- ▶ Navigation (gezielt)
 - ▶ „Entlanghangeln“ an vorhandenen Strukturen, um an die gewünschte Information zu kommen
 - ▶ Voraussetzung ist eine stimmige Aufbereitung/Katalogisierung der Daten
 - ▶ Benutzer kann vorher grob abschätzen, welche Ergebnisse er erhalten wird
- ▶ Suche (unscharf)
 - ▶ Herausfiltern weniger bestimmter Elemente aus einer schwer überschaubaren Menge an Daten
 - ▶ Gerade bei mangelhaft aufbereiteten Daten wichtig
 - ▶ Ergebnisqualität nur schwer vorherbestimmbar

Mögliche Anfragen an ein System

1. Filter (hart)
 - ▶ Kategorie
 - ▶ Zeitraum
2. Beispiele als Anfrage (unscharf)
 - ▶ Suchbild
 - ▶ Textdatei
3. Metainformationen (gemischt)
 - ▶ Schlüsselwörter
 - ▶ Suchstring

Relevance Feedback

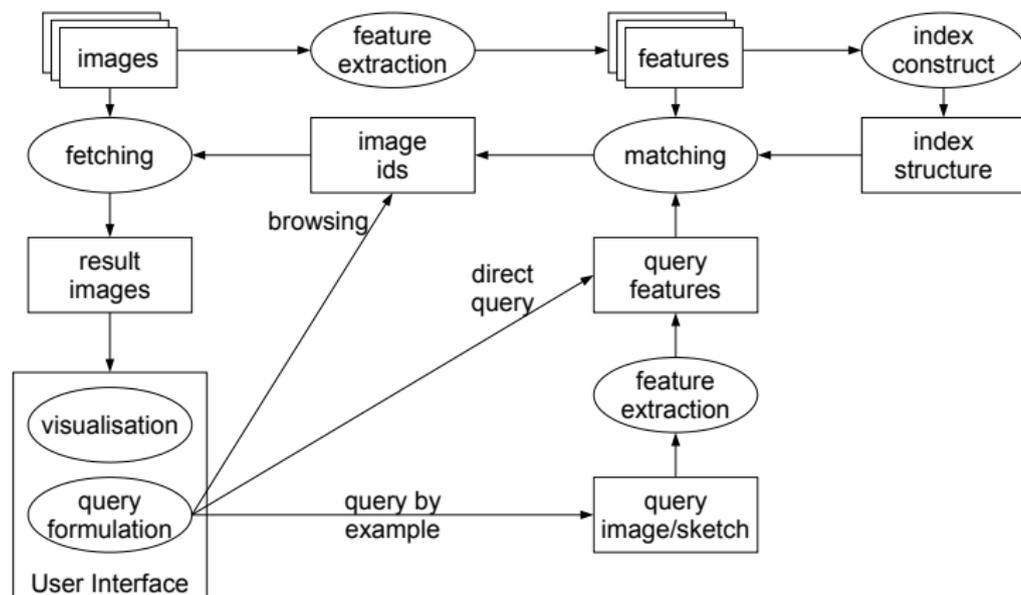
Wie lernt die Suchengine, welche Ergebnisse eine gute Qualität haben, mit der der Nutzer zufrieden ist?

- ▶ Relevance Feedback: direkte Rückmeldung während einer Suche
- ▶ Wikiprinzip: Nutzer können Annotation jederzeit frei editieren/erweitern

Dies gilt insbesondere für Aspekte, die nicht maschinell extrahiert werden können.

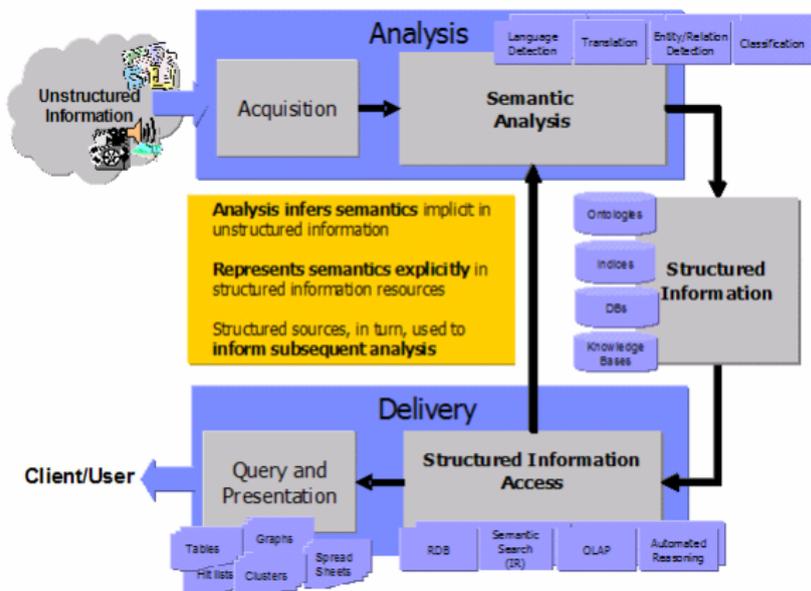
Standard CBIR-Architektur (Veltkamp, Tanase)

„Content Based Image Retrieval“



UIMA-Architektur (IBM Research)

„Unstructured Information Management Architecture“



Vision

- ▶ Flexible Navigation im verteilten Datenbestand
- ▶ Anschauliche Präsentation der Suchergebnisse
- ▶ Unabhängigkeit von eigentlichen Datentypen
- ▶ Verteilte, iterative, spontane Annotation (Wiki)
- ▶ Relevance Feedback
- ▶ Einbindung von Multimediadaten, wie z.B. Videos mit Einsprungstellen

Masterprojekt

Konzepte/Machbarkeitsstudie:

- ▶ Ähnlichkeitssuche über extrahierte Aspekte
- ▶ Semantische Beziehungen über Topicmaps
 - ▶ -> Kombination dieser Anfragen
- ▶ Erfassung von Meta/Indexdaten
- ▶ Manuelle Erweiterung/Verfeinerung der Daten
- ▶ Feste Verknüfungen zwischen Dokumenten
- ▶ Nutzung einer standardisierten Beschreibungssprache (z.B. MPEG-7)

Zusammenfassung

„The same digital technology that lowers the thresholds for producing and publishing content can also help in analyzing and classifying it, in extracting and manipulating features for specific applications and in searching and discovering content.“

Fernando Pereira, Rob Koenen

Weiterführende Literatur I



B. S. Manjunath, Philippe Salembier, Thomas Sikora
Introduction to MPEG-7

John Wiley & Sons Ltd., West Sussex, England, 2002,
ISBN 0-471-48678-7



Hartmut Wittig
Intelligent Media Agents

Vieweg & Sohn Verlagsgesellschaft mbH,
Braunschweig/Wiesbaden, ISBN 3-528-05706-8



J.P. Eakins, M.E. Graham
*Content-based Image Retrieval. A Report to the JISC
Technology Applications Programme*
University of Northumbria at Newcastle, 1999

Weiterführende Literatur II



Remco C. Veltkamp, Mirela Tanase

Content-Based Image Retrieval Systems: A Survey

Department of Computing Science, Utrecht University, 2002



Eero Hyvönen, Samppa Saarela, Kim Viljanen

Intelligent Image Retrieval and Browsing Using Semantic Web Techniques - A Case Study

Helsinki Institute for Information Technology (HIIT) /
University of Helsinki, 2003





Die letzte Seite

Vielen Dank für die Aufmerksamkeit

