Suchen und Finden im Collaborative Workspace

Raoul Pascal Pein

Department Informatik HAW Hamburg

24. November 2006

Einführung

Anwendungsszenarien

- Textbasiert („klassisch“)
- Ortsbezogen
- Bildsuche
- Low-Level Ansatz

Eigener Ansatz

- Architektur
- Skalierung

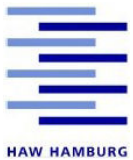
Ziele

- Masterprojekt

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.“

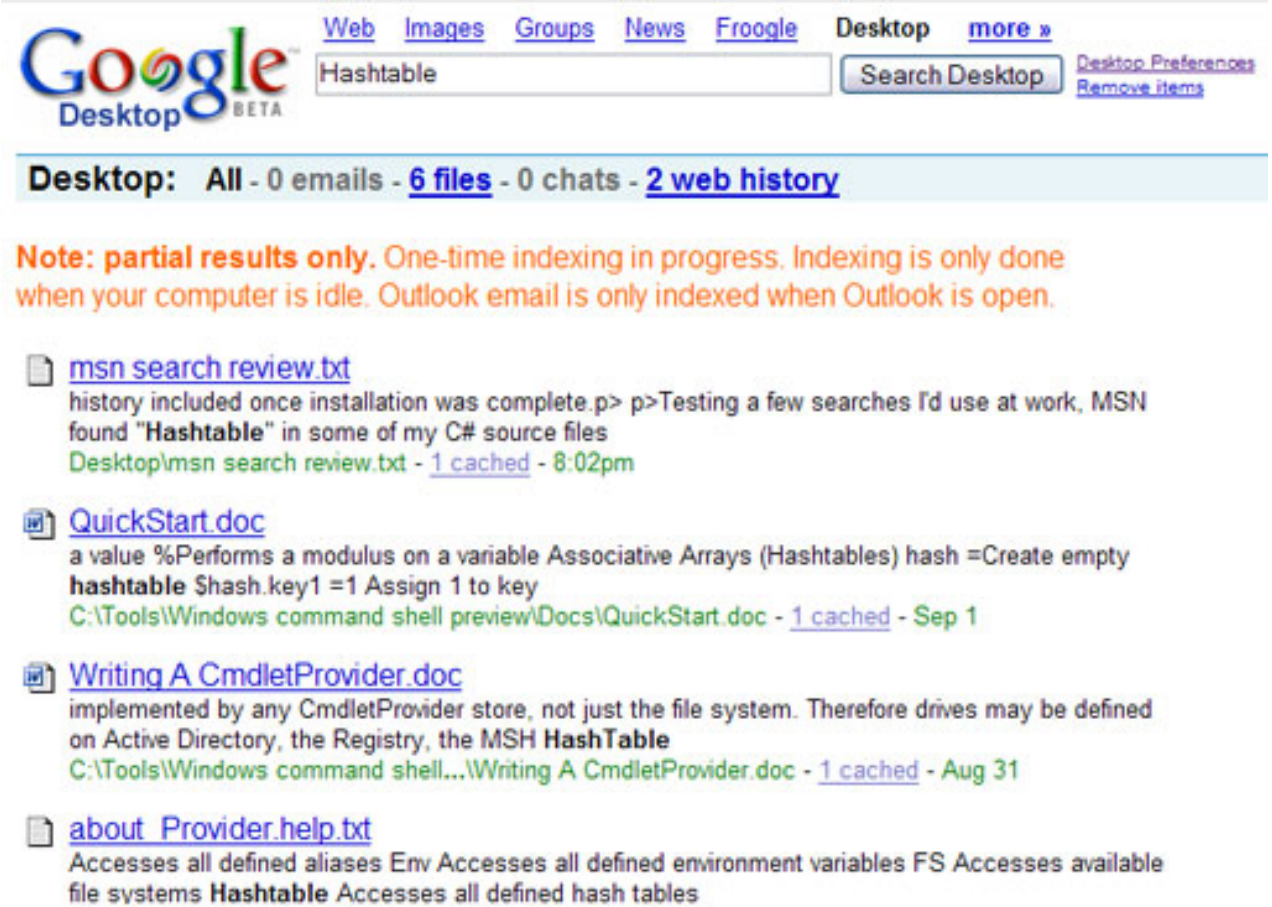
Fernando Pereira, Rob Koenen



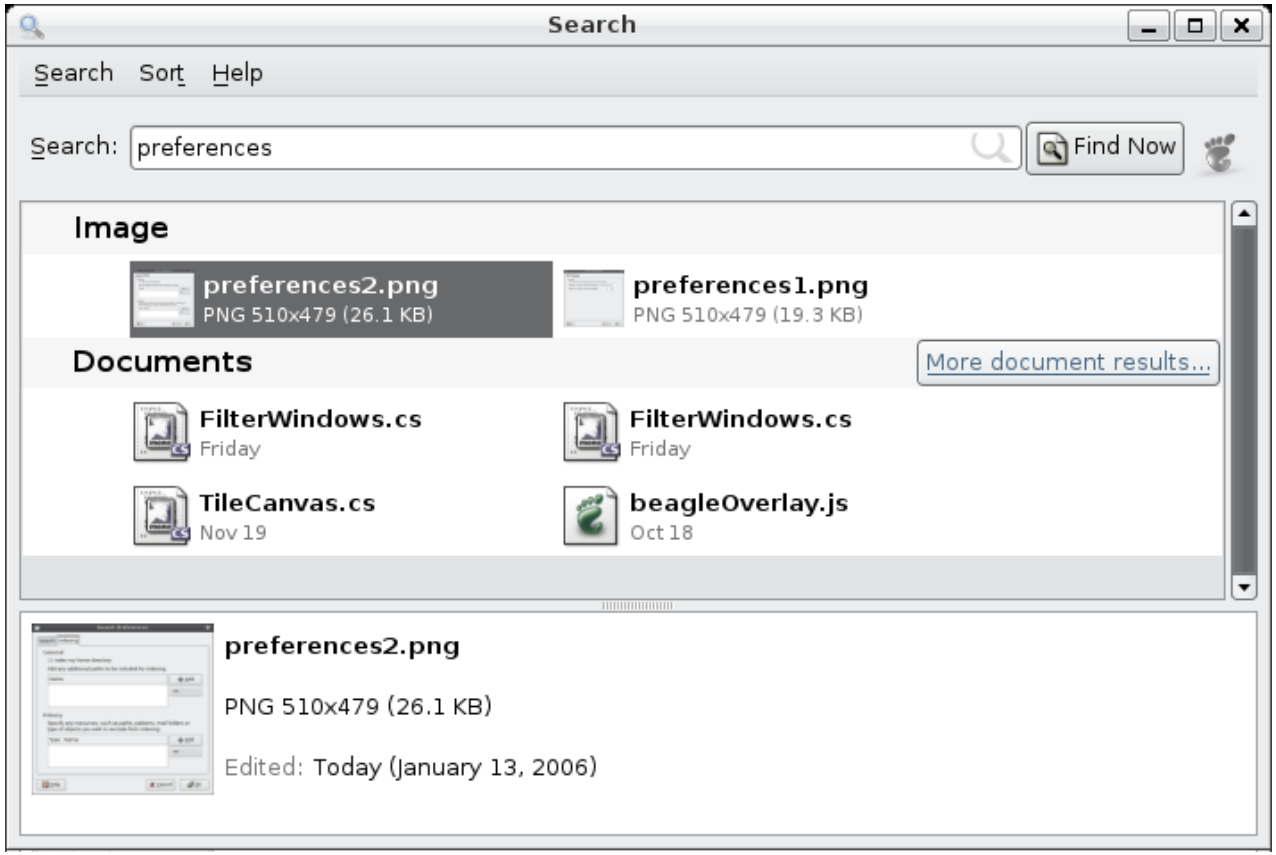
Problem im Collaborative Workspace

Wo sind die Informationen, die ich benötige?

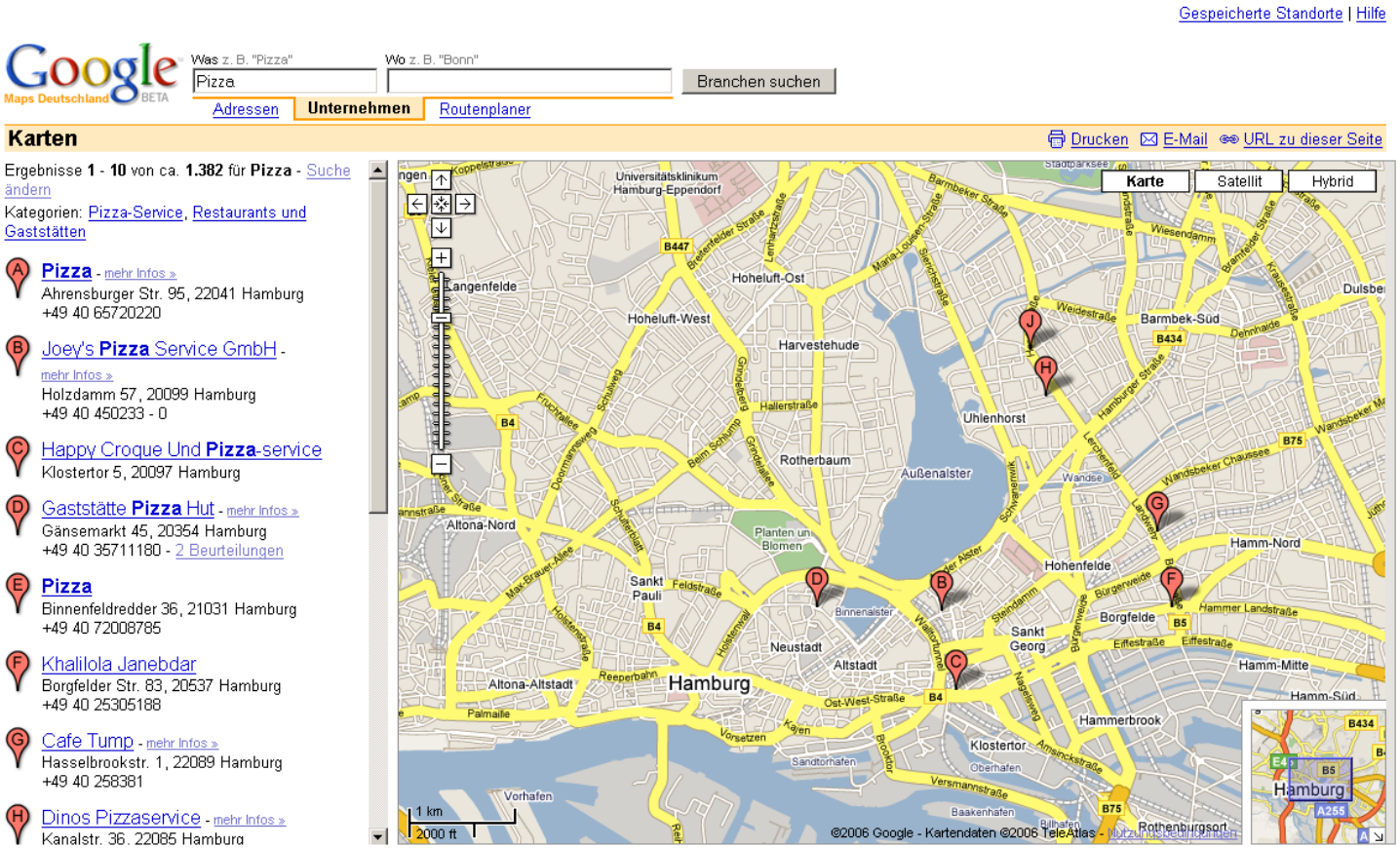
Google Desktop Search



Beagle Desktop Search



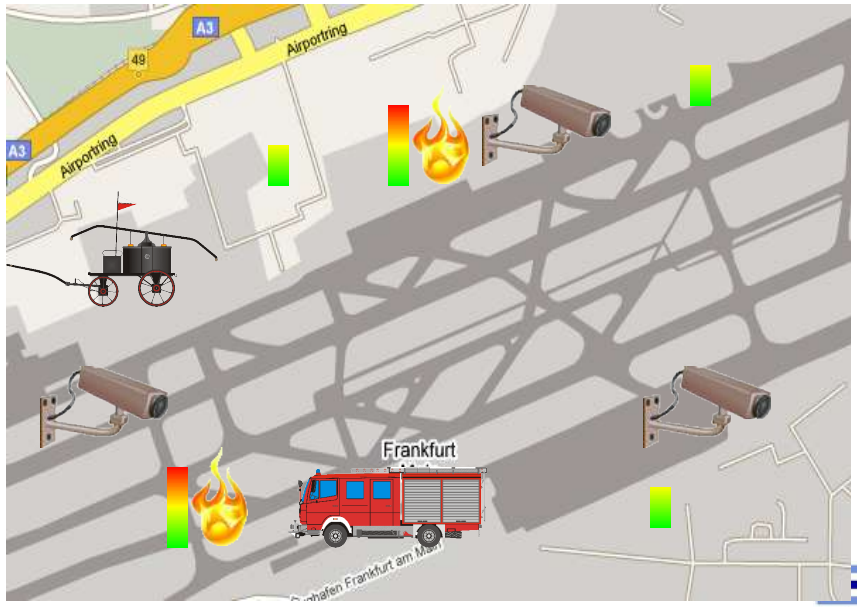
GoogleMaps


Gespeicherte Standorte | Hilfe
Google Maps Deutschland BETA
Was z. B. "Pizza"
Pizza
Wo z. B. "Bonn"
Branchen suchen
Adressen
Unternehmen
Routenplaner
Karten
Drucken
E-Mail
URL zu dieser Seite
Ergebnisse 1 - 10 von ca. 1.382 für Pizza - Suche ändern
Kategorien: Pizza-Service, Restaurants und Gaststätten
A Pizza - mehr Infos »
Ahrensburger Str. 95, 22041 Hamburg
+49 40 65720220
B Joey's Pizza Service GmbH - mehr Infos »
Holzdamm 57, 20099 Hamburg
+49 40 450233 - 0
C Happy Croque Und Pizza-service
Klostertor 5, 20097 Hamburg
D Gaststätte Pizza Hut - mehr Infos »
Gänsemarkt 45, 20354 Hamburg
+49 40 35711180 - 2 Beurteilungen
E Pizza
Binnenfeldredder 36, 21031 Hamburg
+49 40 72008785
F Khalilola Janebdar
Borgfelder Str. 83, 20537 Hamburg
+49 40 25305188
G Cafe Tump - mehr Infos »
Hasselbrookstr. 1, 22089 Hamburg
+49 40 258381
H Dinos Pizzaservice - mehr Infos »
Kanalstr. 36, 22085 Hamburg
Karte
Satellit
Hybrid
Hamburg
Altona-Nord
Sankt Pauli
Neustadt
Altstadt
Hohenfelde
Rotherbaum
Harvestehude
Uhlenhorst
Außenalster
Hoheluft-Ost
Hoheluft-West
Barmbek-Süd
Hamm-Nord
Hamm-Mitte
Hamm-Süd
Borgfelde
Sankt Georg
Klostertor
Hammerbrook
Altona-Altstadt
Universitätsklinikum Hamburg-Eppendorf
Langenfelde
Vorhafen
Rothenburgsort
1 km
2000 ft
©2006 Google - Kartendaten ©2006 TeleAtlas

Rescue

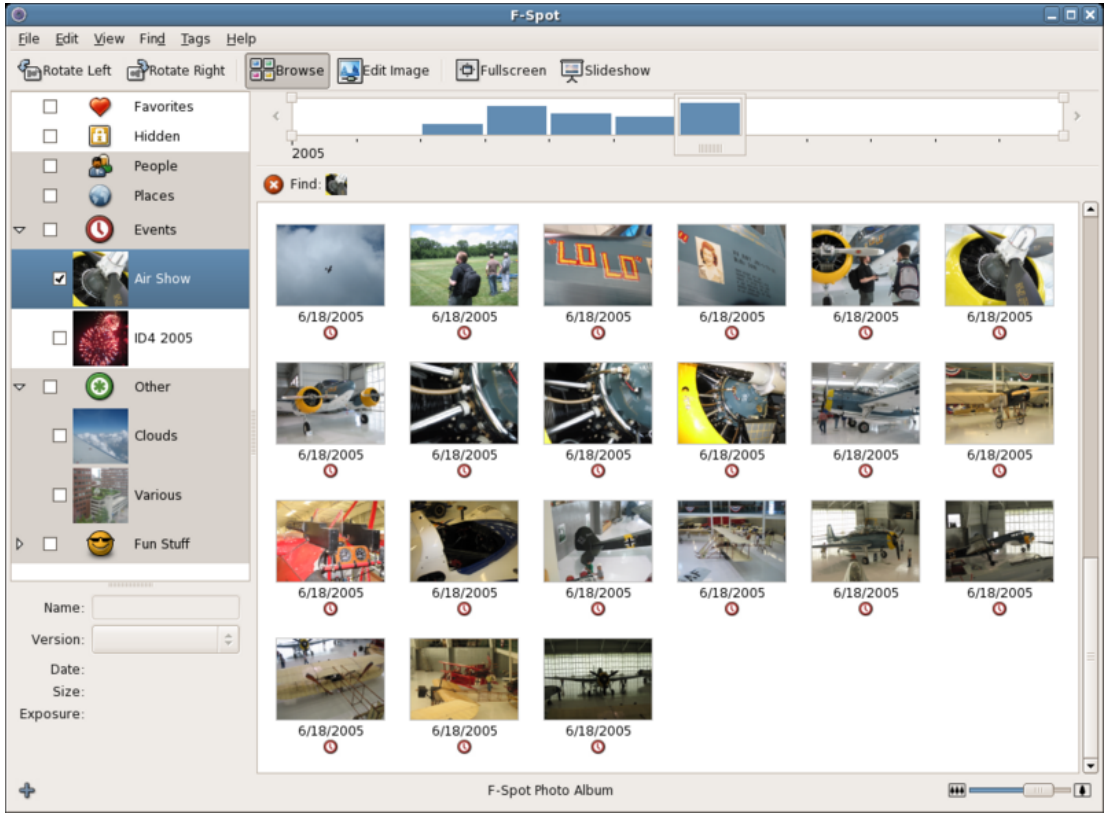
Arbeit mit interaktiven Übersichtskarten
Suche von:

- Sensordaten
  - Positionsabhängig
  - Überschreitung von Schwellwerten
- Kameras
- Einsatzkräften



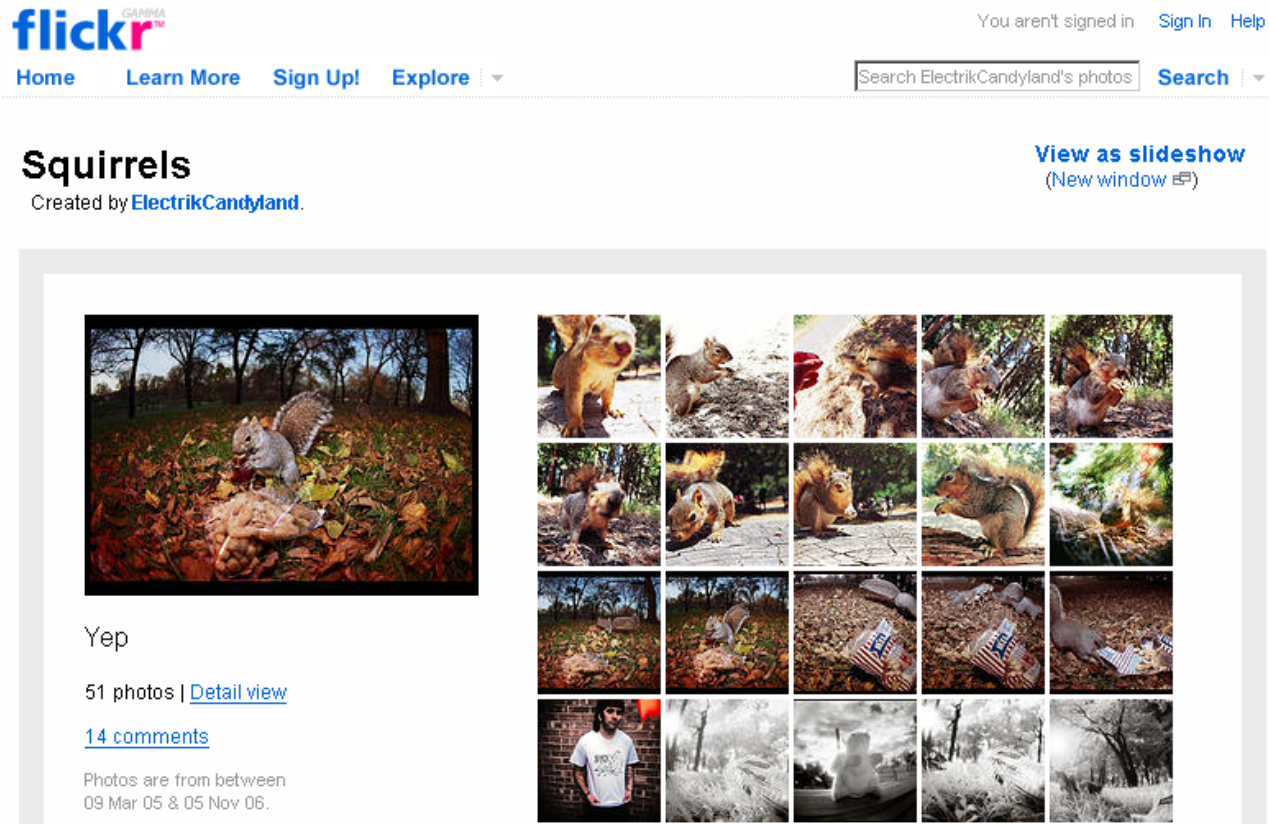
F-Spot

Tags, Timestamps



Flickr

Sets, Groups, Tags, Timestamps, Geotags

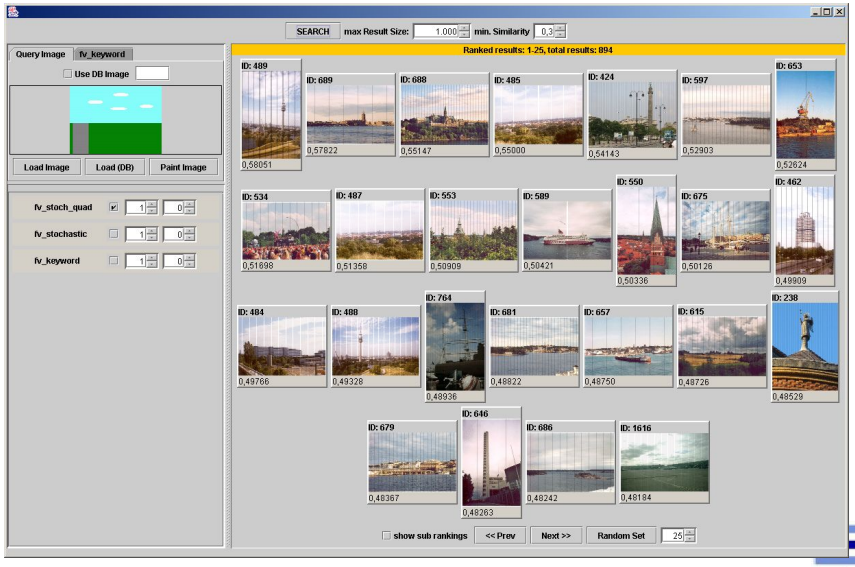


Content Based + Text/Tags

Benutzer kann Bilder in einem großen Bestand ansehen und auch gezielt nach diversen Kriterien suchen und filtern

- Schlagworte
- Kategorien
- Inhaltsbasiert
- Eigene Zeichnung

Zusätzliche Annotation während der Benutzung denkbar

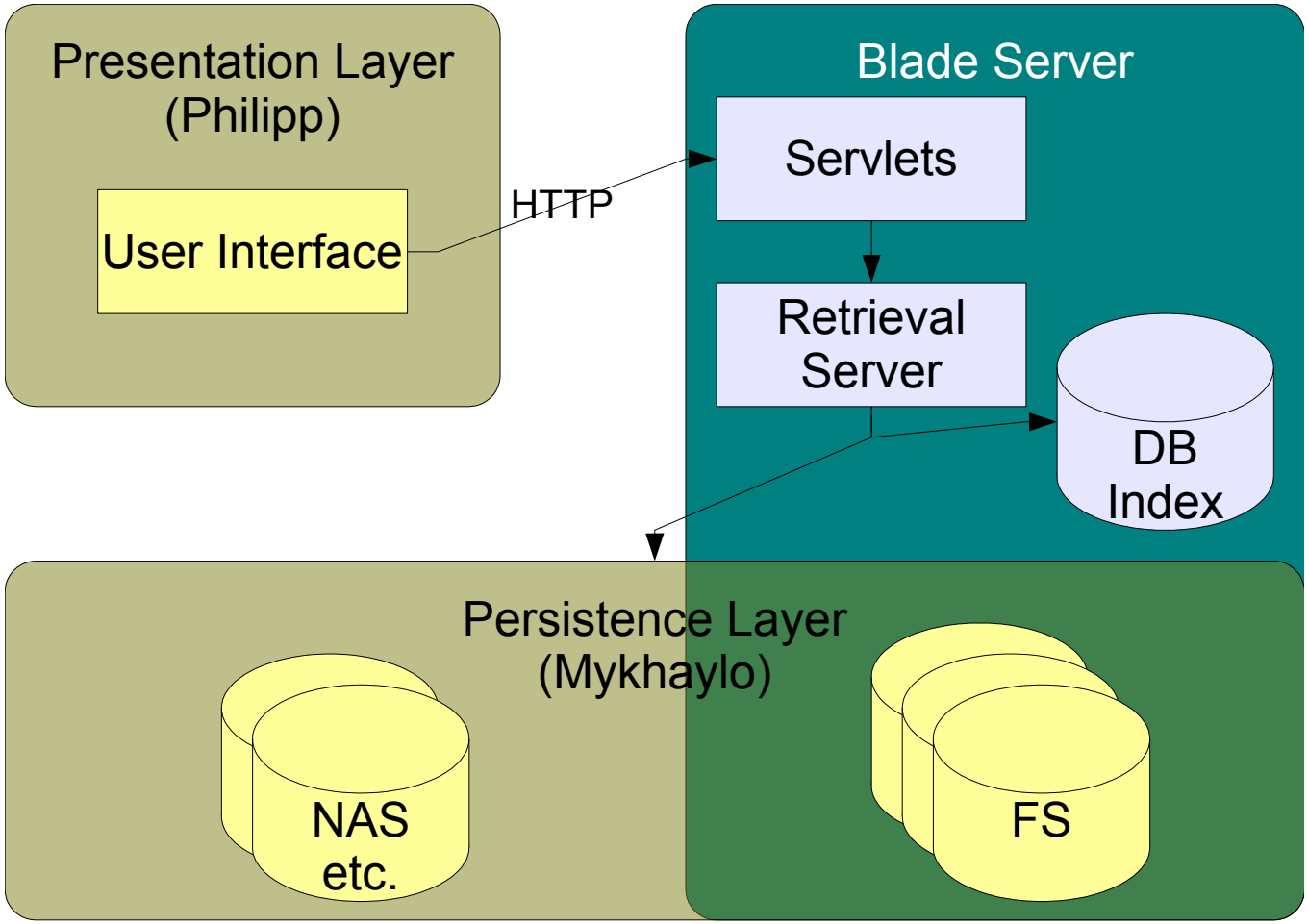


Virtuelle Verzeichnisse

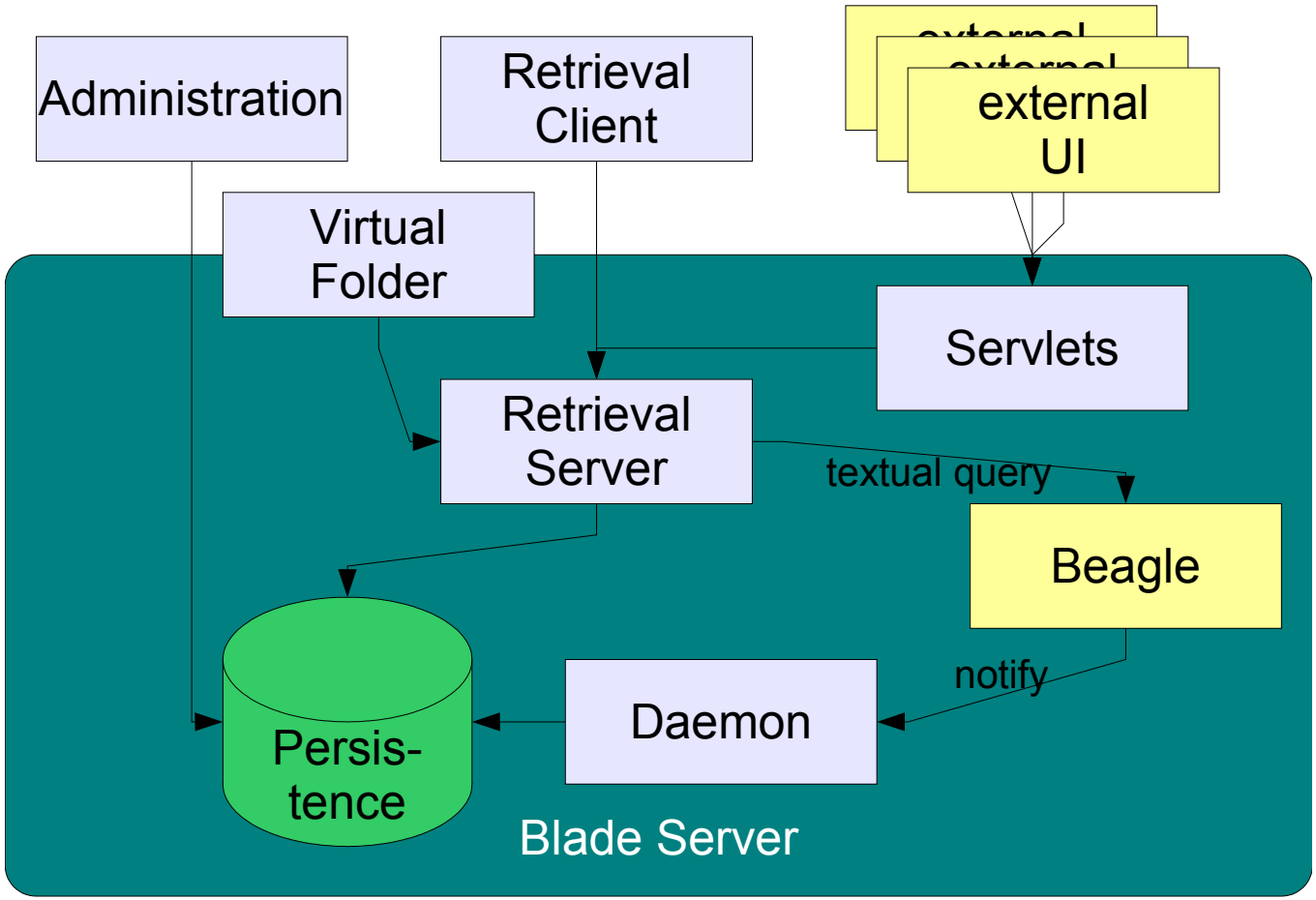
- Auf Dateisystemebene (z.B. Samba Share)
- Vordefinierte Suchanfrage für ein Verzeichnis
- Inhalt repräsentiert Suchergebnis
- Für Anwendungen vollkommen transparent
- Bei Änderungen der Datenbasis automatische Aktualisierung

Beispiel: „intelligente Wiedergabeliste“ bei iTunes

Position im Gesamtprojekt



Retrieval Architektur



Besonderheiten beim Content Based Image Retrieval

- Suche nach Ähnlichkeit statt Identität (über Aspekte wie Histogramme, Formen, ...)
- Keine scharfen Definitionen von Richtig und Falsch
  - Suchanfragen können nur näherungsweise gestellt werden
  - Ergebnisse sind „eher richtig“ und „eher falsch“
- Qualität hängt stark von den zu Grunde liegenden Aspekten ab

Wie erhält man ein brauchbares Ranking?

Beispielaspekte

Universelle Aspekte

- Schlüsselwörter
- Semantik
- Kategorien/Tags
- Erstellungsdatum/Zeitpunkt

Bildspezifische Aspekte

- Histogramme
- Formen
- Wavelets

Berechnung der kombinierten Ähnlichkeit

$$r_x = \frac{1}{\sum_{f=1}^{n} w^f} * \sum_{f=1}^{n} w^f * r_x^f$$

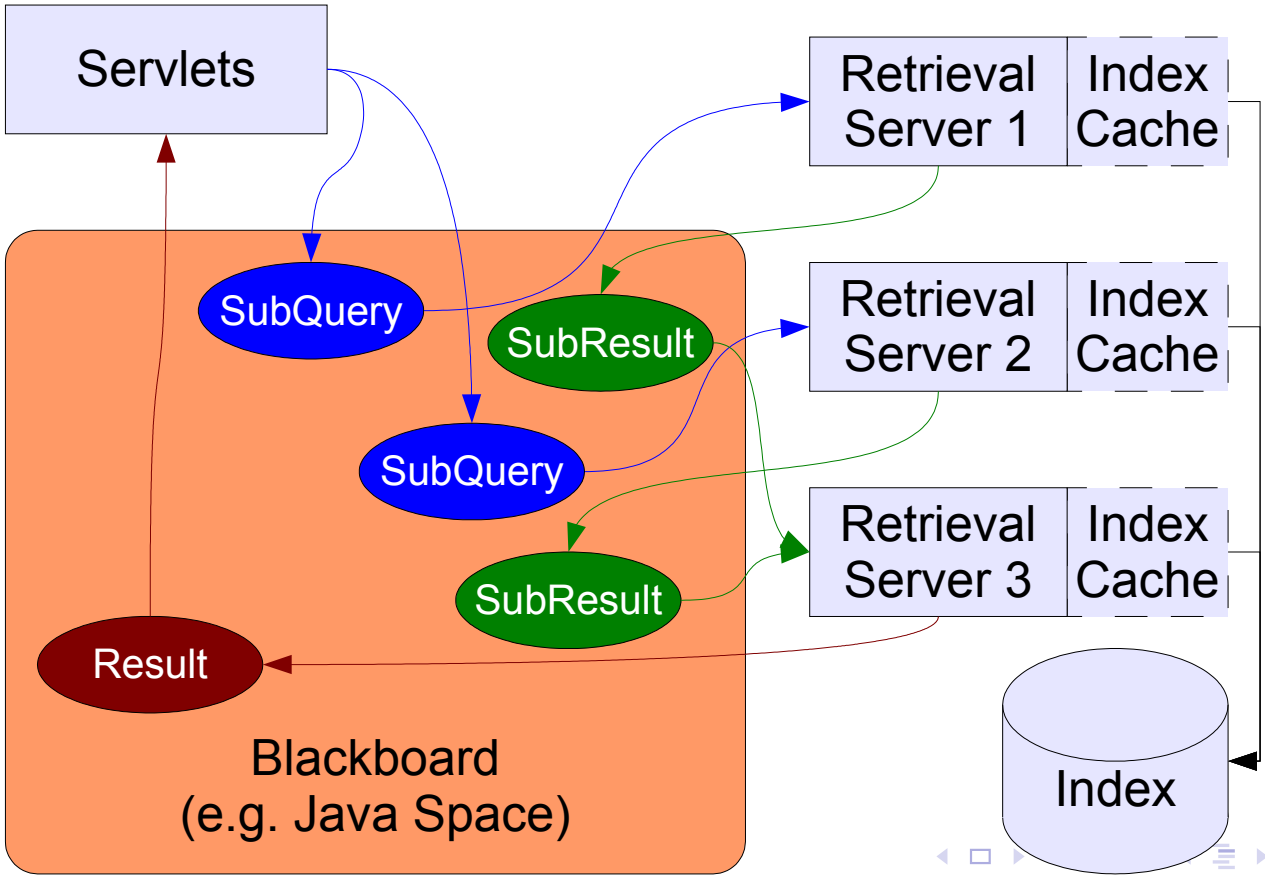
- n Anzahl der verschiedenen Bildaspekte
- x Ein Bild aus dem Datenbestand
- r_x „Ranking“ (Ähnlichkeitsmaß) zwischen Anfrage und Bild x
- f „Feature“ (Vergleichsaspekt)
- w^f Gewichtung eines Aspekts
- r_x^f Teilranking für Bild x bezogen auf Aspekt f

Probleme bei der Skalierung

- Indexerstellung gerade bei hochaufösenden Bildern teuer
- Eine perfekte Ähnlichkeitssuche benötigt einen kompletten Scan über alle Datenätze
- Rankingaufwand steigt linear mit der Anzahl der Datensätze

Load Balancing

Mehrere Server können sich Aufträge auf einfache Weise teilen:



Suchraum eingrenzen

Problem

Ähnlichkeitssuche lässt sich nicht direkt auf einfache Indexe abbilden. Praktisch jeder Vergleich liefert eine Ähnlichkeit $> 0{,}0$. Dadurch existieren keine klaren Grenzen, welches Objekt in die Ergebnismenge gehört.

Lösungsansätze

- Clustering
- Mehrdimensionale Suchbäume
- „harte“ Filter (Keywords, Kategorien, Tags)

Prototyp

- Integration in gemeinsames Projekt
- Ähnlichkeitssuche über extrahierte Aspekte
- Textsuche extern z.B. über Beagle
- Semantische Beziehungen z.B. über Topicmaps (TopicSEEK)
- Automatische Erfassung von Meta/Indexdaten
- Manuelle Erweiterung/Verfeinerung der Indexdaten
- Samba-Shares als low-level Schnittstelle

Risiken

- Evaluierung von CBIR generell schwierig, da es keine Referenzprojekte gibt
- Samba-Shares möglicherweise extrem aufwändig zu implementieren
- Offen zugängliche Systeme können mit „Müll“ geflutet werden (z.B. Wikis)

Evaluierung der kombinierten Suche

Möglicher Ablauf

1. Interne Auswahl eines zufälligen Bildes
2. Suche mit vorgegebenen Parametern (Einzelaspekt, Kombination)
3. Präsentation der Ergebnismenge
4. Versuchsperson selektiert Ergebnisse, die sie als ähnlich ansieht
5. Nächstes Bild

Die erhaltenen Daten können bei ausreichend vielen Testdurchläufen statistisch ausgewertet werden.

Zusammenfassung

Die zu bewältigenden Datenmengen wachsen ständig, auch im privaten Bereich

- Kurze Wege zu den Daten werden benötigt
- Einfache hierarchische Dateisysteme reichen nicht mehr aus
- Effiziente Suchstrategien werden immer wichtiger
- Systembedingt können bei der Suche Daten „verschwinden“
  - Auf Geschwindigkeit optimierte Indexe können inkonsistent sein
  - Inhaltbasierte Suche bei Bildern, etc. ist nicht eindeutig

„Content that cannot be easily found is like content that does not exist, [...].“

Fernando Pereira, Rob Koenen

Weiterführende Literatur I


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


Andreas Christensen
Semantische Anreicherung von Suchanfragen auf Basis von Topic Maps
Hochschule für Angewandte Wissenschaften Hamburg, 2005


Raoul Pascal Pein
Multi-Modal Image Retrieval - A Feasibility Study
Hochschule für Angewandte Wissenschaften Hamburg, 2006

Weiterführende Literatur II

- *Beagle Desktop Search*
  http://beagle-project.org/Main_Page
- *F-Spot - personal photo management*
  http://f-spot.org
- *Flickr*
  http://www.flickr.com/
- *Google Desktop Search / Google Maps*
  http://www.google.com
- *iTunes*
  http://www.apple.com/de/itunes/jukebox/playlists.html

Die letzte Seite

Vielen Dank für die Aufmerksamkeit