

# Technische Aspekte einer Videosuchmaschine

Björn Wilmsmann, CEO –  
MetaSieve GmbH

# Über MetaSieve

- <http://www.metasieve.com>
- Softwareentwicklung
- Internet Software
- Spezialisiert auf Suchmaschinentechnologie
- emTain.tv, Deine Multimedia Findemaschine

# Videosuche

- Metasuche für Onlinevideos vs. Suchfunktion auf den einzelnen Portalen
- Inhalte konzentrieren sich auf wenige Portale: YouTube, Dailymotion, Clipfish ...
- Videoinhalte kleiner Anbieter sind schwer zu finden
- Viele Portale stellen API zur Verfügung

# Unterschiede zur Textsuche

- Konzentration auf wenige Seiten vs. Gesamtes Internet
- Daher entweder gar kein oder nur begrenztes Crawling
- Meta-Informationen vs. Dokument
- Information Extraction vs. Information Retrieval

# Meta-Suchmaschinen

- emTain.tv
- Google Video
- Yahoo Video
- Truveo
- Bing

# Komponenten: APIs

- Programmierschnittstellen
- URLs, die XML / RSS / ATOM zurückliefern
- Vorteil: Strukturierte Daten, die sich direkt weiterverarbeiten lassen
- XSLT, um verschiedene APIs in ein einheitliches Format zu konvertieren

# Komponenten: Cache

- Für jede Anfrage müssen die APIs der einzelnen Portale abgefragt werden
- Anfragen an APIs machen Großteil der Ladezeiten einer Ergebnisseite aus
- Daher: Caching

# Komponenten: Crawler

- Analog zu Text-Suchmaschinen
- Durchsucht das Web nach neuen Dokumenten zur Indexierung
- Folgt Hyperlinks ausgehend von Seed URLs
- Teuer: Traffic und verteiltes Crawling
- Aber: Crawling ist begrenzt auf die für Videos relevante Teilmenge des Webs

# Komponenten: Index

- Crawler / API nachgeschaltete Komponente
- Datenstruktur für durchsuchbare Inhalten
- Vektorbasiertes Modell
- Relevanz nach  $tf*idf$
- Suffix Trees, Invertierter Index
- Lucene

# Komponenten: Index

- Merging
- Indexgröße
- Datenhaltung
- Antwortzeit

# Weitere Ansätze

- Nutzergenerierte Inhalte
  - Echtzeit
  - Ortsbasierte Suche
- Query By Example
  - Audio Fingerprinting
  - Bilderkennung, Spracherkennung

# Nutzergenerierte Inhalte

- Nutzergenerierte Inhalte zur Gewinnung semantischer Daten.
- Semantische Konnotationen ermöglichen besseres Auffinden von Videos.
- Auch Videos mit verwandten Inhalten werden gefunden.
- Playlists: Gruppierung von Videos

# Echtzeit

- Twitter, Facebook und Co.
- Activity Stream
- Videos gefiltert nach Zeitraum und Ort
- Interessant für aktuelle Informationen wie z.B. Nachrichten und aktuelle Trends

# Ortsbasierte Suche

- iPhone, Android, Blackberry
- Geolocation
- Suche nach Informationen zu einem bestimmten Ort

# Query By Example

- Suche anhand von Beispielen
- ‚Finde Videos, die so ähnlich sind wie Video X‘
- Empfehlungen anhand von Nutzerpräferenzen à la Amazon

# Audio Fingerprinting

- Abstrakte Merkmale des Audiosignals können genutzt werden, um Videos zu clustern
- Nutzer erhält Empfehlungen

# Bildererkennung

- Mustererkennung in Bilddaten
- Vorverarbeitung: Normierung, Noise Reduction
- Merkmalsgewinnung
- Merkmalsreduktion anhand von Trennfähigkeit
- Klassifizierung

# Spracherkennung

- Umsetzung gesprochener Sprache in textuelle Repräsentation oder Befehle
- Schwieriges Problem
  - Rauschen
  - Sprecherabhängigkeit
  - Nicht jedes Video enthält gesprochenen Text

# emTain.tv

- <http://www.emtain.tv>
- Deine Multimedia Findemaschine.
- Meta-Suchmaschine für Medieninhalte
- Aggregator für Videoinhalte
- Zusätzliche Informationsgewinnung durch Playlists

# emTain.tv

- Videos erhalten durch die Verknüpfung mit anderen Videos zusätzliche Bedeutung
- Neue Inhalte
- Beziehungen zwischen einzelnen Videos schaffen neue Meta-Informationen
- Netzwerk-Effekt
- Playlists als Vermarktungstool

# Die Technik hinter emTain.tv

- Java App Server
- Grails
- Rich Client Plugins
- Lucene
- Spring Cache
- Cloud Computing für Skalierbarkeit

# Die Technik hinter emTain.tv

