max planck institut
informatik

State of the Art and Trends in Search Engine Technology

Gerhard Weikum (weikum@mpi-inf.mpg.de)

Commercial Search Engines

Web search

→ Google, Yahoo, MSN

- simple queries, chaotic data, many results
- key is precision&importance @ top-10

good for „Britney Spears birthday“

+ open-source software: Lucene&Nutch, etc.

vs.

Enterprise search

→ Verity, Oracle, IBM, Fast, Google, etc.

- advanced queries, high-quality data, few results
- key is recall & saving human time

aiming at „recent conference papers by computer scientists on percolation theory with application of phase transition models to the analysis of Web graph dynamics“

DL search is here!

Thank You !

(End of Talk)

Why More Research?

What Google (& Verity) Can‘t Do

Killer queries (disregarding QA, multilingual, multimedia):

- ***recent conference papers by computer scientists on percolation theory, with application of phase transition models to the analysis of Web graph dynamics***

- by IT professionals, market analysts, IP rights lawyers, etc.:
  - ***peak load of Google***
  - ***effect of XML on IT industry in 2001***
  - ***first published record on search-engine spam countermeasures***
- by computer scientists, political scholars, etc.:
  - ***researcher who has worked on DB technology and astronomy***
  - ***articles that question the feasibility of the Semantic Web***
  - ***timeline of public debate on EU constitution***
- by kids:
  - ***negative reviews about the book „Lord of the Rings“***
  - ***next movie with Johnny Depp***

What is Beyond Google (& Verity)?

for Advanced Information Requests by „Power Users“
(librarians, market analysts, scientists, students, etc.)

- ***background knowledge***
  → ontologies & thesauri, statistical learning

- ***metadata + (semi-)structured & „semantic“ data***
  → XML, info extraction, annotation & classification

- ***humans in the loop***
  → feedback, collaboration, recommendation, peers

- ***context awareness***
  → personalization, geo & time, user behavior

- ***multimedia, cross-lingual, timelines, etc.***

Trends and Opportunities: Ontologies, Thesauri, Semantic Search

- Combine (Light-Weight) Ontologies with other Knowledge Sources (Corpus Statistics, Thesauri, Gazetteers)

  Example: „international organized crime“ automatically expanded into „mafia (0.98) yakuza (0.83) ... drugs (0.88) (money laundry) (0.72) ...“

- Mining Text/Web/Knowledge Sources for Concepts & Relations

  (EU projects, KnowItAll at UW Seattle, German SmartWeb project, etc.)

Trends and Opportunities: Metadata, Semistructured Data, XML IR

- Efficient Query Processing for XQuery Full-Text W3C Standard with Flexible Scoring for Content&Structure Similarity

  Example: /document [//toc „Hidden Markov Models“]
  [//sect[„Speech Recognition“]//equation]
  [//link/person „Kalman“]

- Strong Commercial Interest for Enterprise Search

  (Verity, Fast, MarkLogic, IBM, Oracle, etc.)

Trends and Opportunities: Information Enrichment

- ML Models and Toolkits for Info Extraction & Entity Matching (e.g. at CMU, Stanford, Sheffield, etc.)

  Example: tag all politicians and CEOs in today's newspaper articles and extract who met whom, and when and where

- Commercial NLP, Text-Mining, Extraction Tools by SMEs and Growing Commerical Interest for Enterprise Search to Overcome „Relational Envy" (Verity, Fast, etc.)

- Architectural Frameworks (e.g. IBM's UIMA)

Trends and Opportunities: Personalization

- Statistical Learning from Query Logs and Click Streams Emerges as Major Topic in WWW, SIGIR, etc.
- Personalized Search Embedded in Applications & Workflows (e.g. mobile phone services, job hunting, scholarly workbench, etc.)

Trends and Opportunities: Community Behavior, P2P Networks

➚ (Specialized) P2P Web Search (incl. Multimedia Search) has Great Potential and is Gaining Momentum

(e.g. projects at Berkeley, Stanford, CMU, EPFL, MPII, PlanetLab, etc.)

Trends and Opportunities: Multimedia Search

- Images, Video, Speech, Music, News Create Info Explosion
- Effective Search Can Leverage Metadata, Annotations, Speech-to-Text, Simple Features (e.g. Δ pitch in music) plus Richer Statistical and Semantic Features

Conclusion

short-term (≤1 year): use commercial enterprise-search engine

mid-term (1-2 years):

- consider existing, relatively mature research results as add-ons or for specialized services (e.g. personalized agents on client side or on top of engine)

long-term (> 2 years):

- information explosion continues, users more demanding
- pursue open service-oriented architecture and continuously innovate server-side technology (for XML / semantic / community / multimedia / cross-lingual search)