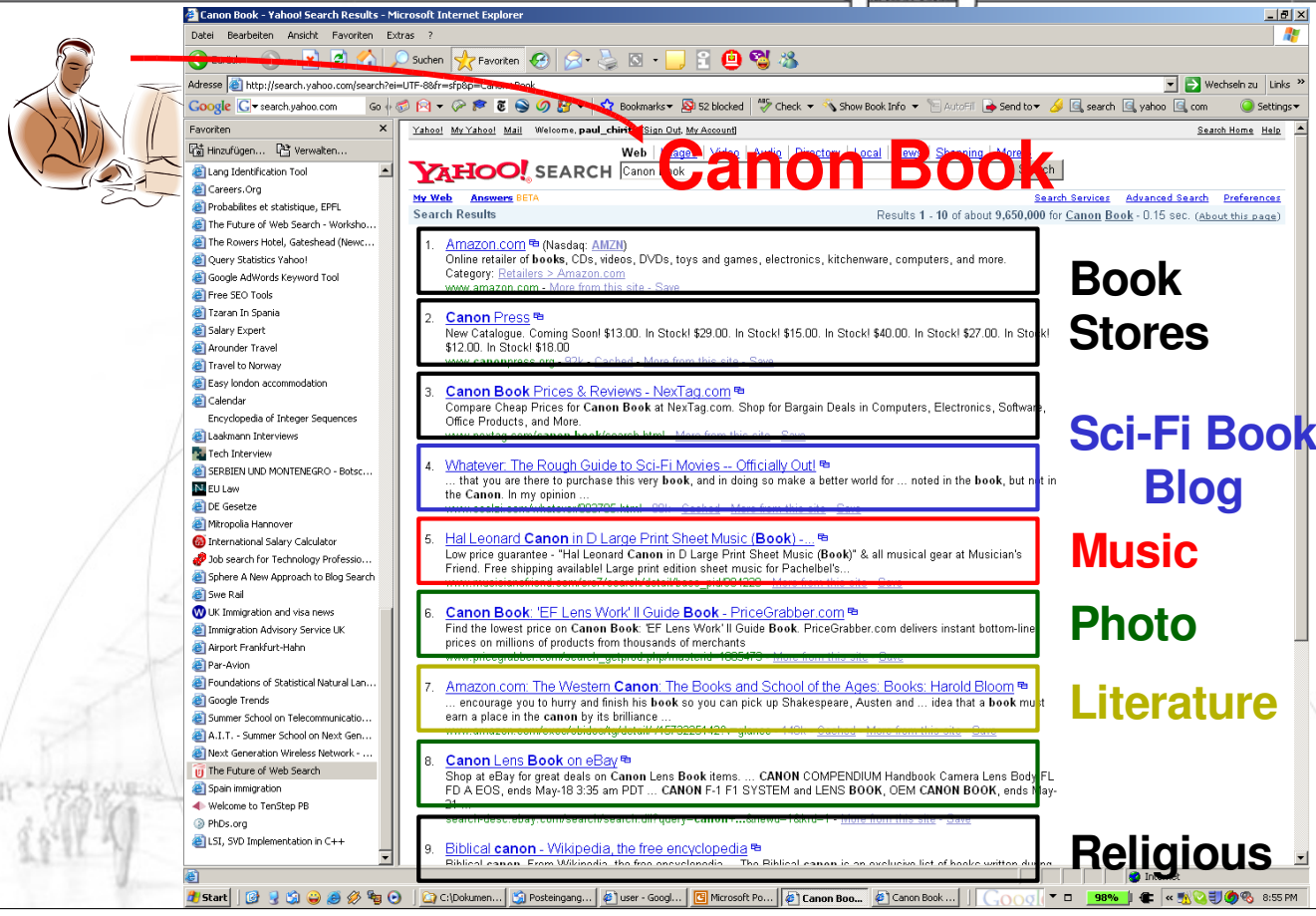
Forschungszentrum L3S

Current Approaches to Personalize Web Search

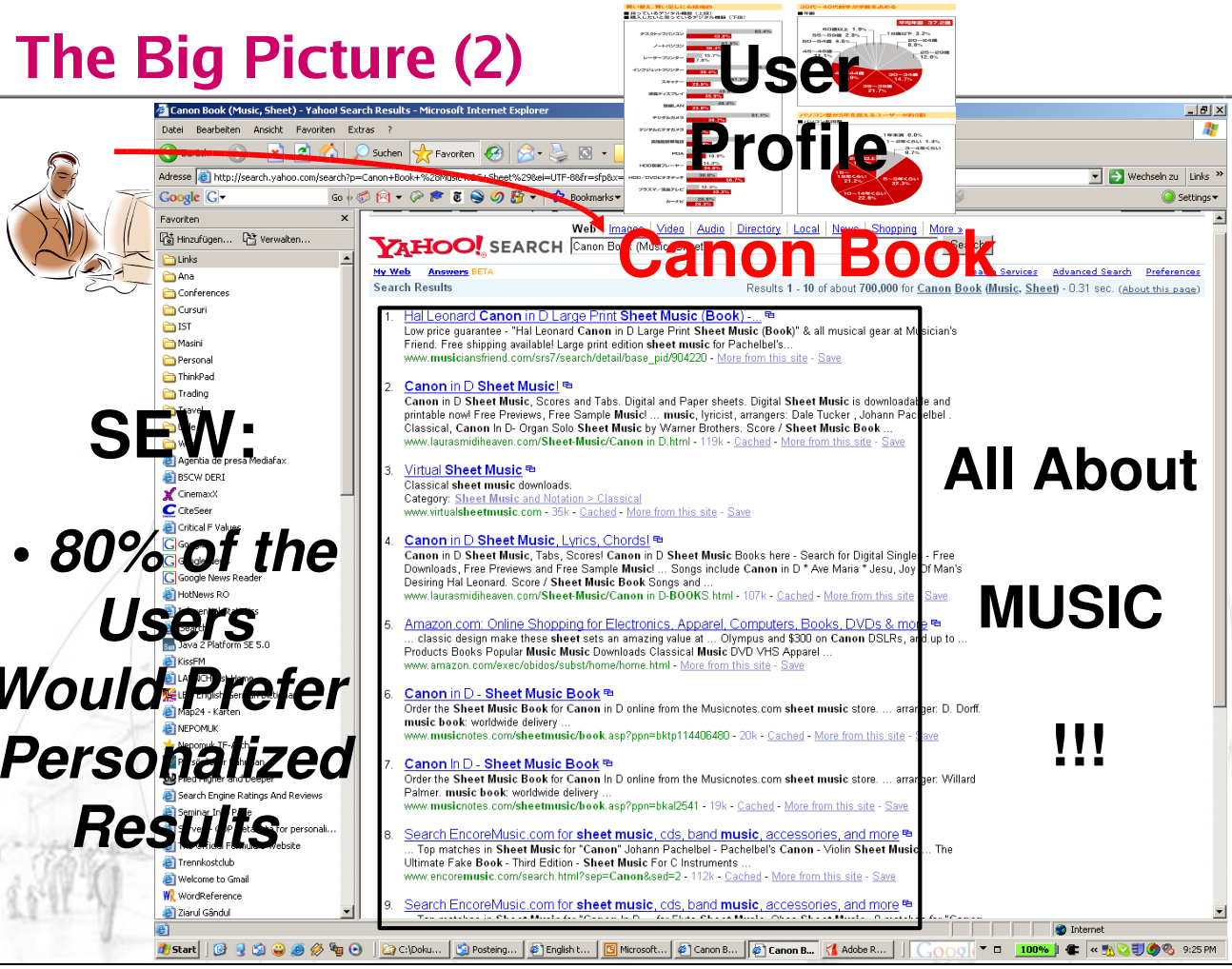
Paul – Alexandru Chirita

Workshop on the Future of Web Search
Barcelona, Spain, May 20th, 2006

The Big Picture



The Big Picture (2)



Forschungszentrum L3S

Outline

- Introduction to User Profiling
- Approaches to Web Search Personalization
  - Personalization Built in PageRank
  - Filtering Uninteresting Search Results
  - Personalization as a Separate Ranking Factor
- Recent Innovations and Challenges

Outline

- Introduction to User Profiling
- Approaches to Web Search Personalization
  - Personalization Built in PageRank
  - Filtering Uninteresting Search Results
  - Personalization as a Separate Ranking Factor
- Recent Innovations and Challenges

User Profile Elicitation

Sources of Input Data

- Previously visited pages
  - Total visit time
  - Last visit / No. of visits
  - Expiration time
  - Page title / text contents
  - No. of out-links followed
- Bookmarks
- Client Data
  - IP address (location)
  - Access method
  - Browser / OS
- Previous search queries / Output URLs clicked
- External / Client side data

Inferring Profiles

- Preprocessing
  - Data cleaning
  - User identification
  - Session identific.
- Profiling (ML & Statistics)
  - Clustering text / Browsing patterns
  - Classification / Decision Trees
  - Discovery of Association Rules
  - Temporal Pattern Discovery

Profile Repres.

- Categories (or Facets)
  - Yahoo! Dir. / Open Directory
  - Selected from a run-time hierarchy
  - Self-defined
- List of URLs
- Bag of words
- Tensors (e.g., user, queries, clicked pgs.)

- Pierrakos et Al.: Web Usage Mining as a Tool for Personalization: A Survey, 2003.
- P. Chan: Constructing Web User Profiles: A Non-Invasive Learning Approach, 2000.
- P. Pu: Survey of Preference Elicitation Methods, 2004.

Outline

- Introduction to User Profiling
- Approaches to Web Search Personalization
  - Personalization Built in PageRank
  - Filtering Uninteresting Search Results
  - Personalization as a Separate Ranking Factor
- Recent Innovations and Challenges

Biased PageRank

- Proposed since 1998 by Page and Brin:

$$\vec{x}^{(n+1)} \leftarrow (1-c) \bullet A \bullet \vec{x}^{(n)} + c \bullet \vec{e}$$

- Profile: List of URLs to bias on
  - Same for all PageRank based approaches
- Advantages:
  - Enhanced quality of rankings
- Disadvantages:
  - Very time consuming when computed separately for each user
  - Not clear how to automatically construct highly qualitative user profiles (promotes pages in the "vicinity" of what we know already ?...)

Direct Extensions to Biased PageRank

- Bias on Topic Specific Sets of URLs (Topic-Sensitive PageRank, WWW02):
  1. Computes 16 PageRank vectors biased on 16 different top-level categories taken from the Open Directory
  2. A similarity score with these topics is computed for each user query
  3. The 16 vectors are then combined using the previous scores as weights
  4. The resulting vector is used to rank the query answers
- Advantages:
  - Ranking adaptive to query topics, as well as to the current user activity context
- Disadvantages:
  - Works only with a limited number of categories
  - Not really a personalization method
- Bias on Preferred URL Domains (Aktas et Al., WebKDD 2004)

→ Quite Limited.

Personalized PageRank

- Biased PageRank used to personalize on user interests
- Compute one Personalized PageRank vector for each user
  - Challenges
    - Reduce storage required
    - Reduce time for computation
- Starts from the idea that Personalized PageRank (PPR) can be expressed as a linear combination of Basis Hub Vectors:

$$PPR_S = \sum_{v \in S} PPR_{\chi_v}$$

- Decomposes each Basis Hub Vector in two parts:
  - *Hub skeleton* vector (common interrelationships, and pre-computed)
  - *Partial* vector (unique values, and computed at construction-time)
- Jeh & Widom: Scaling Personalized Web Search. WWW 2002.

Personalized PageRank (2)

- Monte Carlo method for scaling (Fogaras & Racz, Internet Mathematics)
  - A fingerprint for vertex u is a random walk starting from u
  - The length of the walk is of geometric distribution of parameter c, i.e., after every step the walk ends with probability c, and takes a further step with probability 1 - c
  - The ending vertex of a fingerprint has the distribution of PPR_{Xu}
  - Disadvantage: Approximate values are returned
- Rounding (Sarlos et Al., WWW 2006)
  - All partial values are rounded to a multiple of a prescribed ε

  $\rightarrow$ Little space needed, as we will have maximum $(1/\varepsilon)$ non-zero entries
- Count-Min Sketching
  - Replace $PPR_u^{(k)}$ with its Count-Min Sketch (a randomized approximate representation) within Jeh's algorithm

Merging Personalized and Topic-Sensitive PageRank

- Learn the user profile as a topic preference vector for Topic-Sensitive PR (Qiu & Cho: Automatic Identif. of User Interest for Pers. Search, WWW 2006)
  - User profile initially as a click history
  - Linear Regression
    - If *V* represents the visit probability vector (based on the click history), and *T* the topic preference vector, then minimize over all *M* topics:

$$[V - (\sum_{i=1}^{M} T(i) * TSPR_i^{9/4})]^2$$

  - Maximum Likelihood Estimator over the *k* visited / clicked pages:

$$T = \arg\max_T [\prod_{j=1}^{k} (\sum_{i=1}^{M} T(i) * TSPR_i^{9/4}(p_j))]$$

Outline

- Introduction to User Profiling
- Approaches to Web Search Personalization
  - Personalization Built in PageRank
  - Filtering Uninteresting Search Results
  - Personalization as a Separate Ranking Factor
- Recent Innovations and Challenges

Output Filtering

- Restrict to results from a humanly edited taxonomy / Open Directory
  (F. Liu et Al.: Pers. Search for Improving Retrieval Effectiveness. TKDE 2004)
  - Profile: Bag of words, learned from user's browsing behavior
  - Upon issuing a query, the relevant categories are identified using the profile
  - Output: Original list, merged with category restricted list (only for taxonomies!)
- Cluster filtering (Ferragina & Gulli, WWW 2005)
  - Snippets for clustering search engine results
  - User may select labels of interest, whose clusters are then filtered / promoted
- Classification (Pahlevi: Taxonomy Based Adaptive Web Search, ITCC 2002)
  - Users select some topics from a large taxonomy (e.g., Open Directory)
  - At run-time, the classifier would pick / promote the search output URLs classified into at least one of the categories of interest

Outline

- Introduction to User Profiling
- Approaches to Web Search Personalization
  - Personalization Built in PageRank
  - Filtering Uninteresting Search Results
  - Personalization as a Separate Ranking Factor
- Recent Innovations and Challenges

Re-Ranking Techniques

- Main idea:
  - Use the *distance* between the user profile and each output URL as a separate ranking factor
- Re-Ranking with the profile as bag of words
  (Sugiyama et Al.: Adaptive Web Search ,.. WWW 2004)
  - Profile: Bag of words, learned from user's previously visited pages
  - Interesting separation of *permanent profile* and *recent profile* (for current session)
  - Upon issuing a query, the search results are re-ordered based on their similarity to the user profile
- Classification (Gauch et Al.: Ontology Based Pers. Search and Browsing, 2003)
  - User's browsing data classified into topics of interest
  - Search output re-ranked according to the classifier scores for (1) topic of each URL, and (2) user interest in that topic

Re-Ranking Techniques (2)

- Problems with previous approaches:
  - Time consuming (for computing similarity scores at the word level, for classification, etc.)
  - Privacy: A lot of data / browsing behavior collected from the user
- Re-Ranking exploiting ODP

  (Chirita et Al.: Using ODP Metadata to Personalize Web Search, SIGIR 2005)
  - Profiles as topics from ODP
    - Could be learned *only from previous search queries* by analyzing the ODP topics associated to their results
  - Re-Ranking according to the conceptual similarity between the topics within the user profile and the topics associated to each output URL (very fast)
  - About 40 of the Top-100 results are either in ODP, or in the Yahoo! Directory

Outline

- Introduction to User Profiling
- Approaches to Web Search Personalization
  - Personalization Built in PageRank
  - Filtering Uninteresting Search Results
  - Personalization as a Separate Ranking Factor
- Recent Innovations and Challenges

Newer & Different Approaches

- Personalized Search as an Application of LSA
  (Sun et Al.: CubeSVD, WWW 2005)
  - Three dimensional LSA on User ID / Query / Clicked Page
  - Catches both similar users and similar queries (similar to Collaborative Filtering)
    → May recommend queries, pages, etc.
  - Disadvantage: Seems to be too resource consuming for the Web scale
- Exploiting Personal Data
  (Teevan et Al., Personalizing Search via Automated Analysis ,… SIGIR 2005)
  - Involves desktop data for user profiling
  - Modified BM25 to incorporate external sources (e.g., various types of personal information, etc.)
  - Takes a Relevance Feedback approach in which the desktop documents are automatically considered relevant
    → Search results are Re-Ranked
  - Interesting, though only minimal precision improvements were reported

The *Future* of Web Search Personalization

- More complex profiles
  - Tighter connection to the Search Algorithm
  - Less information collected / Less user interaction required / Increased accuracy of interest prediction
- More adaptivity
  - Some queries are more ambiguous than others
  - Amount of personalization as a function of query ambiguity
- More social
  - Profiles enriched from friends and neighbors in the web

Thank You!

Presentation by Paul – Alexandru Chirita, chirita [at] l3s [dot] de.