Beyond Document Similarity: Understanding Value-Based Search and Browsing Technologies

Andreas Paepcke
Hector Garcia-Molina
Gerard Rodriguez-Mula (Spain)
Junghoo Cho

Outline

- Introduction
- Conceptual Architecture
- Content-based Value Filtering
- Action-based Value Filtering
- Summary

<table>
<thead>
<tr>
<th>Date</th>
<th>Submitted to</th>
</tr>
</thead>
<tbody>
<tr>
<td>5/18/1999</td>
<td>ACM Computing Surveys</td>
</tr>
<tr>
<td>2/4/1999</td>
<td>DL99</td>
</tr>
</tbody>
</table>
Introduction

Problem

- Current IR systems: searching and ranking
  - one or two words per query ⇒ high volumes of documents on the Web
  - multimedia data ⇒ new techniques beyond similarity measure

Solution

- Value filtering approaches
  - indicators of information value
    - independent of similarity with any given query
  - help users throttle the flow of information
  - attach searchable index to non-textual data
Introduction

Example

- Query
  - 購買中古汽車

- Conditions of similarity measure
  - 跑車、超強馬力
  - 深色系、流線型

- Indicators of information value
  - 經銷商、原車主
  - 可議售價、出廠年份、使用情況
  - 市場評價、個人偏好
Introduction

- Possible Techniques
  - Explicit user participation
  - Automatic extraction from documents
  - Observation of user accesses

- Applications
  - Search: cull documents among query results
  - Rank: help user digest a number of results
  - Browse: guide users with valuable links
  - Filter: selective dissemination of information
  - Cache: indexing of highly valuable information
Conceptual Architecture

User interface:
- Query/refinement
- Browsing
- Filter control
- Feedback

User:
- Filter engine
  - Filtering/SDI
  - Ranking
  - Browsing view generation

Filter engine:
- Access accounting
  - The Web
  - A website
  - Databases
  - Email folders
  - News feed

Collections:
- Metadata
  - Indexes
  - Links
  - Categories
  - Genres

Metadata Extractor

Information flow

Access accounting

Filter engine

Collection Metadata
- Access statistics
- Profiles
- Bookmarks
- Votes

Judgment Metadata
Content-based Value Filtering

■ Definition
  ‣ By static clues from documents or collections

■ Categories
  ‣ Document analysis
    • analyze individual documents
  ‣ Collection analysis
    • analyze entire collections
  ‣ Information context
    • determine the context of documents
  ‣ Document-internal content tags
    • manually place tags within documents
Content-based Value Filtering

- **Document Analysis**
  - PHOAKS finds URLs from Usenet messages
    - words surrounding URLs
    - URLs’ positions
  - TileBar provides visual clues about locations
    - support users in manually filtering query results
  - Vocabulary complexity
    - rate the reading level of documents for each user
  - Genre of documents
    - newspapers, journals, advertisements, interviews
    - certificated samples ⇒ patterns ⇒ predict genres
Content-based Value Filtering

- Collection Analysis
  - Google crawls the Web for indexing
    - prefer a document with more links pointing to it
    - by the authors’ opinions
  - SCAM finds mirrored documents of a website
    - prefer such documents with survivability precautions or performance enhancements
  - PHOAKS excludes the URLs in the messages posted to multiple news groups
    - hint of advertisements
Content-based Value Filtering

- Information Context
  - Publisher of documents
  - Time at which the document was published
    - individual preferences $\Rightarrow$ customized services
  - ReferralWeb finds experts for consultations
    - registrant $\Rightarrow$ related individuals $\Rightarrow$ a community
    - prefers documents that are connected with anyone in the user’s context
  - Scatter/Gather and SONIA create contexts by interactively clustering documents
    - manually control the filtering activities
Content-based Value Filtering

- Information Context (continued)
  - SenseMaker combines controlled clustering with automated filtering
    - criteria: author, publication date, website, ...
  - COATER determines the semantic contexts
    - WordNet is a list of concepts with related words

- Document-internal Content Tags
  - PICS has publishers add tags to documents
    - prevent minors from inappropriate materials
  - RDF allows complex schema to be built for websites
    - a framework for using metatags
Action-based Value Filtering

■ Definition
  ‣ By dynamic clues from human actions

■ Categories
  ‣ Explicit judgment
    • relevance feedback for filtering
    • data-triggered filters
    • synthesized filters
  ‣ Implicit judgment
    • conjecture from collective user behavior
    • conjecture from individual user behavior
Action-based Value Filtering

Relevance Feedback for Filtering

- Tapestry allows users to annotate documents
  - a collaborative filtering system
  - explicit judgments by more than a single user
  - the feedback itself is the grist for filtering
    - ignore the contents (not content-based filtering)
- Fab and GroupLens find out which users are best suited as sources of recommendation
  - feedback $\Rightarrow$ interest profiles $\Rightarrow$ colleagues $\Rightarrow$ recommendation by voting
Action-based Value Filtering

■ Data-triggered Filters
  ‣ Mail filters allow users to discard messages
    • manually construct filter expressions
  ‣ NetNanny eliminates undesirable data based on a list of words or phrases
    • websites, news groups, chat rooms, …
  ‣ SIFT allows users to enter interest profiles
    • Selective Dissemination of Information (SDI)

■ Synthesized Filters
  ‣ LyricTime picks and plays songs for users
    • mood indicators: cheerful, romantic, calm, sad, …
    • one profile per listener, per mood
Action-based Value Filtering

- Conjecture from Collective User Behavior
  - WebWatcher supports guided tours
    - correlations between links and user interests
  - Path clustering $\Rightarrow$ user/page clustering
    - a path matched $\Rightarrow$ hyperlink suggestions
  - KSS annotates links by access frequencies
    - served by a proxy
  - HotBot and DirectHit collect the return rates of query results
    - improve their ranking algorithms
Action-based Value Filtering

- Conjecture from Individual User Behavior
  - HotBot and DirectHit
    - record the keywords and collections associated with the returned results
  - Predictors of user interests
    - reading time
    - operations: save, bookmark, follow links, reply, …
  - Siteseer aggregates personal bookmarks to support collective filtering
  - WebWatcher and Letizia evaluate the merits of links by matching keyword vectors
Summary

- Linkage
- Mirror sites

Collection Analysis

Document Analysis

Content-based Value Filtering

- URL scavenging
- Location within documents
- Vocabulary complexity
- Genres of documents

Information Context

- Authorship
- Temporal context
- Personal context
- Human filter control

Document-internal Content Tags

- Metatags
- Schemas
Summary

Action-based Value Filtering

Collective User Behavior

Data-triggered Filters
- Mail filters
- Content blocks
- SDI profiles

Mood indicators

Synthesized Filters
- Reading time
- Operations
- Bookmarks

Explicit Judgment

Implicit Judgment

Relevance Feedback

Individual User Behavior

Operations

Bookmarks

Voting

Annotations

Matched paths

Access frequencies

Return rates
Summary

- Contributions
  - A conceptual architecture of value filtering
  - A survey and a categorization of techniques

- Research Issues
  - Side-effects of continuous positive feedback
  - Comparative filtering effectiveness
    - logging, sampling, value decay, ...

- Research Directions
  - New types of collection and judgment metadata
  - User queries for information values