BINGO! and Daffodil: Personalized Exploration of Digital Libraries and Web Sources

Martin Theobald
*Max-Planck-Institut für Informatik*

Claus-Peter Klas
*Universität Duisburg-Essen*
Overview

- Challenges in Web Recommendations for Digital Library Users
- DAFFODIL
- BINGO!
- Meta Data Processing
- System Integration via Web Services
- Experiments
1. **Recommendations for Digital Library Users**

- **Current recommendations systems are based on**
  - User collaboration & mining of user behaviour
  - Pre-computed document similarities (clustering, cosine measure, co-citations, etc.) limited to a single DL index

- **Challenges in providing Web recommendations**
  - Well organized and structured Digital Library world vs. heterogeneous and chaotic Web
  - Unified view on related publications *and* topic-specific Web documents (e.g., authors’ homepages)
  - “Virtual links” from meta data through additional Deep Web search
    - Topic-specific hub pages
    - Full text sources for training
  - Filtering, user feedback, and iterative topic refinement
2. **DAFFODIL: Motivation**

- Many access points, e.g., DBLP, CiteSeer, ACM, Springer, Achilles, HCIBIB
- Different query formats & services
- Insufficient coverage
- Insufficient functionality
2.1. Combination of Services & Information

- One access point
- Integration of DLs and individual services
- Meta data extraction
- Synergy effects
  - Strategic information search
  - Information organization in personal folders
2.2. Lifecycle of Scientific Search as an Interactive and Iterative Process

- Create Knowledge
- Information Sources
- Retrieve Information
- Structured Personal Lib
- Interpret Retrieved Information
- Collate
- re-present
- discover
- retrieve

SHARE
2.3. Personal Library

- Publications organized as structured meta data
- Digital Library Objects (DLOs)
  - Authors, title
  - Web links
  - Journals
  - Conferences
  - Keywords, abstracts
  - Queries
- Meta data export in XML or BibTex
- Basis for information exchange with BINGO!
3. **BINGO!**

- **Focused crawler**
  - Combination of traditional crawler with classifier and link analyzer for topic distillation
  - Crawls mainly within the densely connected neighbourhood graph for the topics of interest
  - Filters out irrelevant documents directly at the crawler frontier

- **Information portal generation**
  - Initialized by a set of training documents (e.g., *bookmarks*)
  - Extends and maintains a user-specific topic-directory (Yahoo!)

- **Expert queries**
  - Initialized by a single-topic keyword query
  - Exploits existing index or external search engine for retrieving initial hubs → training & seed
  - Improves recall
3.1. Focused Crawler Architecture

Training documents

Iterative (Re-)Training, Feedback

Crawler

Document analyzer

Feature selection

SVM classifier

HITS link-analyzer

World Wide Web

URL queue

Feature vectors, TF*IDF

Topic-specific features

Taxonomy index

Hubs & Authorities

Theobald, Klas: Personalized Exploration of Digital Libraries and Web Sources
4. Meta Data Processing

- **DLOs do provide**
  - Comprehensive meta data describing the publication

- **DLOs may provide**
  - Links to full text sources (e.g., PS, PDF or DOC - files)
  - Links to homepages, DL portals, or other Web references

- **Focused crawler requires**
  - Full text documents for feature extraction and training
  - Topic-specific hubs as start pages of focused crawl

- **Solution**: Additional Deep Web queries using DLO contents

- **Web Service Framework (WSF)**
  - Replace form submission method of the browser (HTTP post) with new WS
    - Detection of attribute labels/names next to form fields
    - Form types determine attribute types (e.g., String, Date, etc.)
  - Generation of WSDL description for WS parameters
  - Automatic deployment under local Tomcat Web Server
Advanced search form of the DBLP Trier

(HTML <form> and <input> tags)

WSF

WSDL description registered at local UDDI

Theobald, Klas: Personalized Exploration of Digital Libraries and Web Sources
4.1. Ontology-Enabled Form Matching & Portal Selection

- Ontology service
  - Incorporates WordNet, Cyc, and personalized domain ontologies
  - Mapping of attribute names to semantic concepts
  - Quantified relationships between concepts
  - Approximate matches between similar attributes in a DLO and WSDL

- Portal selection
  - Aggregation of attribute similarities between a DLO and WSDL
  - Ranking for the top matching portals to each meta data query

- Unified handling of heterogeneous portals
  - Pre-computed (static) mapping on schema level only
  - Automatic form completion and submission
DLO extracted from Achilles

```xml
<metadata docid="d738000632"
  folder="/webdavroot/Uni+Saarbruecken/Workflow">
  <authors>
    <author id="hung#patrickck">
      <firstname>Patrick C.K.</firstname>
      <lastname>Hung</lastname>
    </author>
    <author id="karlapalem#kamalakar">
      <firstname>Kamalakar</firstname>
      <lastname>Karlapalem</lastname>
    </author>
  </authors>
  <booktitle>Australasian Information Security Workshop</booktitle>
  <pages>33-41</pages>
  <publisher>ACS</publisher>
  <publisheraddress>Adelaide, Australia</publisheraddress>
  <series>Conferences in Research and Practice in Information Technology</series>
  <title>A Secure Workflow Model</title>
  <url>Achilles</url>
  <volume>21</volume>
  <year>2003</year>
</metadata>
```

WSDL for DBLP Trier

```xml
<message name="WSF_Form0_InputMessage">
  <documentation>http://www.informatik.uni-trier.de/ley/db/dbin/dblpquery.cgi</documentation>
  <part name="Authors[author]WSF_Form0Text0_Part" type="xsd:string" />
  <part name="[author]WSF_Form0Text1_Part" type="xsd:string" />
  <part name="[author]WSF_Form0Text2_Part" type="xsd:string" />
  <part name="[author]WSF_Form0Text3_Part" type="xsd:string" />
  <part name="Title[title]WSF_Form0Text4_Part" type="xsd:string" />
  <part name="Year[year]WSF_Form0Text5_Part" type="xsd:string" />
  <part name="Page[pages]WSF_Form0Text6_Part" type="xsd:string" />
  <part name="Conference[booktitle]WSF_Form0Text" type="xsd:string" />
  <part name="ID[cite]WSF_Form0Text8_Part" type="xsd:string" />
  <part name="Journal[journal]WSF_Form0Text9_Part" type="xsd:string" />
  <part name="Volume[volume]WSF_Form0Text10_Part" type="xsd:string" />
</message>
```
5. System Integration

- **Web Service Infrastructure**
  - Two standalone Tomcat Web Servers for WS routing
    - Daffodil at Universität Duisburg
    - Bingo! + WSF at MPII Saarbrücken

- **Asynchronous coupling**
  - Distributed multi-user requests
  - Queuing of recommendation requests at Daffodil (FIFO)
  - Integration of Bingo! into Daffodil’s agent-based structure

- **Simple communication protocol (via RPCs)**
  - 1) Daffodil: **Init** request with basic crawler parameters
  - 2) Bingo!: **Fetch** folders along with user id & timestamp
  - 3) Bingo!: **Return** top recommendations
Theobald, Klas: Personalized Exploration of Digital Libraries and Web Sources

**DAFFODIL**
Distributed Agents for User-Friendly Access of Digital Libraries

**BINGO!**
Bookmark-Induced Gathering of Information

1) **init**
- Core DB
- Web IR
- Workflow

2) **fetch**
- Full text sources
- Core DB
- Web IR
- Workflow

3) **return**
- Web Portal 1
- Web Portal m

**User Folders**
- DLOs, Metadata
- Author = ...
- Title = ...

**Training Docs**
- Ontology-Enabled Portal Selector
- Web Service Framework

**BINGO! Agent for Recommendations**

**Focused Crawler**

<table>
<thead>
<tr>
<th>No.</th>
<th>Author</th>
<th>Title</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>W8.</td>
<td>P.Grefen</td>
<td>Transactional Workflows or Workflow Transactions ?</td>
<td>2002</td>
</tr>
</tbody>
</table>

### 6.1. First Iteration & Feedback

<table>
<thead>
<tr>
<th>No.</th>
<th>URL</th>
<th>BINGO! Rating</th>
<th>Key Resource</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td><a href="http://www.e-workflow.org">http://www.e-workflow.org</a></td>
<td>3.86</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td><a href="http://www.waria.com">http://www.waria.com</a></td>
<td>3.76</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td><a href="http://www.wfmc.org/information/info.htm">http://www.wfmc.org/information/info.htm</a></td>
<td>2.91</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td><a href="http://www.usc.edu/dept/ATRIUM/Papers/PDI.pdf">http://www.usc.edu/dept/ATRIUM/Papers/PDI.pdf</a></td>
<td>2.89</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td><a href="http://www.informatik.uni-stuttgart.de/ipvr/as/">http://www.informatik.uni-stuttgart.de/ipvr/as/</a></td>
<td>2.68</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>projekte/poliflow/IPVR/adaptive_workflows.html</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td><a href="http://www.wfmc.org/information/awards.htm">http://www.wfmc.org/information/awards.htm</a></td>
<td>2.68</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td><a href="http://www.dfgi.uni-kl.de/~aab">http://www.dfgi.uni-kl.de/~aab</a> Becker/Freiburg/Final/Wargitsch/Wargitsch.pdf</td>
<td>2.63</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td><a href="http://www-dbs.cs.uni-sb.de/projekte/workflow_de.htm">http://www-dbs.cs.uni-sb.de/projekte/workflow_de.htm</a></td>
<td>2.37</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Precision**: 0.45
### 6.2. Second Iteration

<table>
<thead>
<tr>
<th>No.</th>
<th>URL</th>
<th>BINCO! Rating</th>
<th>Key Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.</td>
<td><a href="http://www-dbs.cs.uni-koeln.de/~gillmann/Publications/ConfigTool-EDB.pdf">http://www-dbs.cs.uni-koeln.de/~gillmann/Publications/ConfigTool-EDB.pdf</a></td>
<td>1.79</td>
<td>1</td>
</tr>
<tr>
<td>5.</td>
<td><a href="http://www.informatik.uni-stuttgart.de/ipvr/as/projekte/apricots/atma96.pdf">http://www.informatik.uni-stuttgart.de/ipvr/as/projekte/apricots/atma96.pdf</a></td>
<td>1.78</td>
<td>1</td>
</tr>
<tr>
<td>7.</td>
<td><a href="http://www.cs.colorado.edu/~skip/proclcts.pdf">http://www.cs.colorado.edu/~skip/proclcts.pdf</a></td>
<td>1.70</td>
<td>1</td>
</tr>
<tr>
<td>8.</td>
<td><a href="http://www.ifi.unizh.ch/dbtg/Projects/TRAMs/trams.html">http://www.ifi.unizh.ch/dbtg/Projects/TRAMs/trams.html</a></td>
<td>1.64</td>
<td>1</td>
</tr>
<tr>
<td>12.</td>
<td><a href="http://ccs.mit.edu/klein/cscw98/paper07/">http://ccs.mit.edu/klein/cscw98/paper07/</a></td>
<td>1.57</td>
<td>1</td>
</tr>
<tr>
<td>13.</td>
<td><a href="http://ccs.mit.edu/klein/cscw98/paper08/">http://ccs.mit.edu/klein/cscw98/paper08/</a></td>
<td>1.54</td>
<td>1</td>
</tr>
<tr>
<td>14.</td>
<td><a href="http://www.computer.org/proceedings/Hiccs2/0001/00010198Babs.htm">http://www.computer.org/proceedings/Hiccs2/0001/00010198Babs.htm</a></td>
<td>1.51</td>
<td>0</td>
</tr>
<tr>
<td>17.</td>
<td><a href="http://osm7.cs.byu.edu/ER97/workshop4/ls.html">http://osm7.cs.byu.edu/ER97/workshop4/ls.html</a></td>
<td>1.33</td>
<td>1</td>
</tr>
<tr>
<td>18.</td>
<td><a href="http://www.informatik.uni-ulm.de/dbis/f&amp;t/forschung/flow/workflow/ftext-adept_e.html">http://www.informatik.uni-ulm.de/dbis/f&amp;t/forschung/flow/workflow/ftext-adept_e.html</a></td>
<td>1.33</td>
<td>1</td>
</tr>
<tr>
<td>20.</td>
<td><a href="http://www.ifi.unizh.ch/groups/dbtg/Projects/SEAMAN/seaman1.html">http://www.ifi.unizh.ch/groups/dbtg/Projects/SEAMAN/seaman1.html</a></td>
<td>1.22</td>
<td>0</td>
</tr>
</tbody>
</table>

**Precision:** 0.9
7. Conclusions & Outlook

- Coupling of Bingo! and Daffodil provides significantly added value for advanced users
  - For Daffodil: Context-based search over the Web
  - For Bingo!: Access to semi-structured metadata and the Deep Web

- Future work:
  - More exhaustive experimental evaluation
  - Extensive study of relevance feedback
  - DLO extraction from recommended URLs
  - Personalized crawling as a publicly available service for Daffodil users (dedicated server)
Try it out!

www.daffodil.de