30 Years of DL Research
@ UNIPD

Maristella Agosti, Giorgio Maria Di Nunzio, Nicola Ferro, Maria Maistro, Stefano Marchesin, Nicola Orio*, Chiara Ponchia*, Gianmaria Silvello

Department of Information Engineering
*Department of Cultural Heritage
University of Padua, Italy

14TH ITALIAN RESEARCH CONFERENCE ON DIGITAL LIBRARIES
University of Udine
26 gennaio 2018
Outline

- DUO and the birth of the DL area
- Annotations
- CULTURA and Digital Archives
- Log Analysis
- Personalisation
- User-System Interaction
- Data Driven Digital Libraries
DUO: An Innovative OPAC
The Italian Library Automation Project and the OPAC

- The Italian national project of library automation, called SBN - Servizio Bibliotecario Nazionale, is an advanced library automation project started in 1970s.

- Different library automation systems at national/regional/local level cooperating in a networked/hierarchical organisation.

- Until late in the 1980s

- The public online access to bibliographic data was not available, only traditional card catalogues were in use.
The University of Padua became a node of the SBN project in the late 1980s.

At that time, there was much interest in OPAC:

- A first indication that information retrieval might start to interest the general public of libraries.
- We launched a project for a third generation OPAC with advanced library catalogue and IR functions.
The OPAC DUO Interface (in Italian)

Maristella Agosti – University of Padova, Italy
The time was not ripe for Web applications: the IR functions were lost
The Birth of the Digital Library Area

- The Library Automation community realises the lack of computer science and engineering knowledge

- The area of Digital Library starts in those years as a new scientific area
  - In USA - Digital Libraries Initiatives (DLI-1 and DLI-2) of the National Science Foundation (from late 1993)

- It is an area of confluence: library automation, database management, information retrieval, the Web, …
Evolution of Digital Libraries

DL a meeting point of disciplines and research fields:

- DB management
- Information retrieval
- Library and information systems
- Document and information systems
- The Web
- Information visualization
- Artificial intelligence
- Human-computer interaction
References


- S. Walker. Improving subject access painlessly: recent work on the Okapi online catalogue projects. Program, 1988, 22(1), 21-31


Annotations
Historical Annotations: Padua University

- Italia, Padova, Archivio dell’Università di Padova, Archivio antico, Matricula Nationis Germanicæ artistarum, reg. 465, c. 69v

a₁

Godefridus Woyssel, filius Sigismundi Woysselli reipublicae Vratislaviensis archiatrii, incipitæ Germanorvm nationi nomen suum dedit et solvit mense Augusto. Patavi anno 1605 die 29 Augustii

[Godfrey Woyssel, Sigmund’s son, from Bratislava, enrolled and paid 6 liras]

a₂

Dedit libræ 6
[He payed 6 liras]

a₃

Propter ignominias litteras Nationi nostræ scriptas ignominose ex Nationis alio in publico conventu, omnium consensu, extirpatus est

[He was ignominiously expelled by the Council of Association, because he wrote ignominious letter to the Association]
Annotation Model

Legend
- annotate link
- relate-to link

http://dbpedia.org/resource/Carrarese

http://ipsa.ipsa-project.org/ipsa-web/r/illustration/1213

 Meaning
dc:contributor

Sign
Story of the Carrarese family

Meaning
ipsa:is-copied-from

<p>The <i>Racconcella</i> Herbal illumination is clearly copied from the <i>Carrarese Herbal</i> one, as it ...</p>

Sign
This illumination presents an extraordinary search for realism

http://ipsa.ipsa-project.org/ipsa-web/r/illustration/136

Meaning
rdfs:comment
I would like to find information about *illuminated manuscripts* and their usage with *annotations* as discussed by ferro.
Search by Using Annotations

I would like to find information about illuminated manuscripts and their usage with annotations as discussed by ferro.

**Best Match**

**Exact Match**

Hypertext

imDL: Illuminated Manuscripts Digital Library

csDL: Computer Science Digital Library
FAST (Flexible Annotation Semantic Tool): a Tool to Innovate
Annotations in the CULTURA Project

○ The CULTURA project
  ○ innovative environment for users with a range of different expertise
  ○ users can collaboratively explore, interrogate and interpret complex and diverse digital cultural heritage collections

○ Use cases
  ○ IPSA: a digital archive of illuminated manuscripts produced in northern Italy during the 14th and 15th centuries
  ○ The 1641 Depositions: the documents contain witness testimonies from men and women from all over Ireland and report on the rebellion of October 1641
References

Digital Archives, IPSA and Dante
- Digital archive of illuminated manuscripts

- Result of a joint research project
  - Ex-Department of History of Visual Arts and Music
  - Department of Information Engineering

- Initially designed for a specialist public of scholars and researchers

- 56 manuscripts and more than 3000 images
DANTE.
A web-application for the History of Art
References


CULTURA
- STREP project active from February 2011 to January 2014

- Engaging new user categories with digital cultural heritage collections
  - User centered-approach

- Exploring new ways to experience digital cultural heritage
  - Adaptivity
  - Interactive environment
Collections

- IPSA: a digital archive of illuminated manuscripts produced in northern Italy during the 14th and 15th centuries

- The 1641 Depositions: the documents contain witness testimonies from men and women from all over Ireland and report on the rebellion of October 1641
Continuous Evaluation to collect New User Categories Requirements

- Involving different user categories from the very beginning of the project

Continuous Evaluation:

- November and December 2011: baseline evaluations with professional researchers and students
- April 2012: IPSA trials with postgraduate students
- May 2012: interaction with professional researchers
- September 2012: IPSA@CULTURA
- December 2012 and January 2013: evaluation of IPSA@CULTURA with a new cohort of students and a new user category (Salvalarte Group)
- October 2013 DILL Master evaluation
- October 2013 High school students evaluation
References


Personalisation
Motivation

- Lack of a seamless browsing experience for users across (cultural) websites

- Most of personalisation approaches unable to assist users in information needs that span different (cultural) websites

- Need for a consistent Cross-Site Personalisation (CSP) support mechanism that provides:
  - Valuable exchange of information between a target website and the user modelling service
  - Users’ privacy needs

- Use case:
  - The CULTURA project
Cross-Site Approach

- Complementary approach to the CULTURA Virtual Research Environment’s demonstrator

- Designed as an open domain approach

Start

1. CULTURAL HERITAGE ONE
   - Track User Activities
   - Trigger
   - Provide data

2. Authentication/Sign-In
   - Provide tracked behaviours & terms

3. User Profile Component

4. User Model Provider
   - API

5. Term Identification Component
   - Extract Text Entities

6. CULTURAL HERITAGE TWO
   - Track User Activities
   - Trigger
   - Provide data

IMS Research Group

30 Years of Digital Library Research at the University of Padua
Architectural Design

Cross-Site Browsing Space
Web space introduced by websites that integrate the cross-site service through module extensions

Interface Layer
Facilitate communication between websites and the service

User Profile
Store and aggregate users' activities and content related text entities

Term Identification
Extract relevant text entities from browsed content within websites

Cross-Site Browsing Space

Website
Extension

Interface Layer

User Profile Component
User Model Provider
Term Identification Component

IMS Research Group
30 Years of Digital Library Research at the University of Padua
References


Log Analysis
- Case study: A portal born to offer access to the resources of 45 European National libraries.

- Analyse the data contained in the logs of their Web servers.

- Find a general methodology of HTTP Web Log analysis in order to
  - give advice about the security of the portal,
  - discover and study personalization issues.
15/10/2008
User Questionnaire

Search Log
HTTP Log

15/10/2008
User session

14.07 - 14.11
Search "divina commedia"

14.04 - 14.07
Search actions

14.04 - 14.07
HTTP requests

14.04 - 14.07
GET /index.html
GET /image.jpg

14.04 - 14.11
Search actions

14.07 - 14.11
HTTP requests

14.12 - 14.15
Search actions

14.10 - 14.12
HTTP requests

15/10/2008
User session

14.04 Start

Liked Divina Commedia

Search Log
HTTP Log

13/10/2008
19/10/2008
14/10/2008
15/10/2008
16/10/2008
17/10/2008
18/10/2008
19/10/2008

14.07 - 14.11
Search "divina commedia"

14.04 - 14.07
Search actions

14.04 - 14.07
HTTP requests

14.04 - 14.07
GET /index.html
GET /image.jpg

14.04 - 14.11
Search actions

14.07 - 14.11
HTTP requests

14.12 - 14.15
Search actions

14.10 - 14.12
HTTP requests

15/10/2008
User Questionnaire

Search Log
HTTP Log

13/10/2008
19/10/2008
14/10/2008
15/10/2008
16/10/2008
17/10/2008
18/10/2008
19/10/2008

14.07 - 14.11
Search "divina commedia"

14.04 - 14.07
Search actions

14.04 - 14.07
HTTP requests

14.04 - 14.07
GET /index.html
GET /image.jpg

14.04 - 14.11
Search actions

14.07 - 14.11
HTTP requests

14.12 - 14.15
Search actions

14.10 - 14.12
HTTP requests
Multilingual LogCLEF

- Verifiability and repeatability of experiments
  - using the same period of time across different studies.

- From 2009 to 2011 LogCLEF Task
  - Language identification task
  - Query classification
  - Success of a query
  - Query refinement
References


- Maristella Agosti, Franco Crivellari, Giorgio Maria Di Nunzio, Yannis E. Ioannidis, Elefterios Stamatogiannakis, Mei Li Triantafyllidi, Maria Vayanou: Searching and Browsing Digital Library Catalogues: A Combined Log Analysis for The European Library. IRCDL 2009: 120-135

- Maristella Agosti, Franco Crivellari, Giorgio Maria Di Nunzio: Evaluation of Digital Library Services Using Complementary Logs. UIRIR@SIGIR 2009


- Giorgio Maria Di Nunzio, Johannes Leveling, Thomas Mandl: Multilingual Log Analysis: LogCLEF. ECIR 2011: 675-678

- Thomas Mandl, Giorgio Maria Di Nunzio, Julia Maria Schulz: LogCLEF 2010: the CLEF 2010 Multilingual Logfile Analysis Track Overview. CLEF (Notebook Papers/LABs/Workshops) 2010
User-System Interaction
Query logs have proven to be a valuable and informative source of implicit user feedback:

- they can be **easily collected** by IR systems;
- they are available in **real time**;
- they represent **personalized** user preferences.
Query logs have proven to be a valuable and informative source of implicit user feedback:

- they can be easily collected by IR systems;
- they are available in real time;
- they represent personalized user preferences.
User Models

Different Types of Users

Different Search

Different Evaluation
Markovian User Model

We assume that each user decides, independently from the random time spent in the first document, to move forward or backward to another document in the list.
We assume that each user decides, independently from the random time spent in the first document, to move forward or backward to another document in the list.
We assume that each user decides, independently from the random time spent in the first document, to move forward or backward to another document in the list.
We assume that each user decides, independently from the random time spent in the first document, to move forward or backward to another document in the list.

\[ X_1, X_2, X_3, \ldots \in \mathcal{R} = \{1, 2, \ldots, R\} \]

random sequence of document ranks visited by
Markovian User Model

We assume that each user decides, independently from the random time spent in the first document, to move forward or backward to another document in the list.

22.6% Backward Transitions
Application of the Markovian User Model

- New family of evaluation measures, called Markov Precision, which injects user models into precision

- Describe the user dynamic, which can be integrated in the Learning to Rank algorithm LambdaMart.
References


Data Driven DL
Data-Driven Digital Libraries

Data-intensive Research

Acquisition & modelling

Archiving and preserving

Collaboration and visualisation

Dissemination & sharing

Analysis & data mining

fourthparadigm.org

Science@Microsoft
Reproducibility

No research paper can ever be considered to be the final word, and the replication and corroboration of research results is key to the scientific process


An experimental result is not fully established unless it can be independently reproduced. Additional benefits ensue if the research artifacts are themselves made publically available so that any interested party may audit them. This also enables replication experiments to be performed.

Reproducibility is connected to data citation

- Principles and standards for data citation are unlikely to be used unless the process of extracting information is coupled with that of providing a citation for it.

- We need to automatically generate citations as the data is extracted.

- Data citation is a computational problem.

Buneman, Davidson, Frew: Why data citation is a computational problem. 
Commun. ACM 59(9): 50-57 (2016)
References


- G. Silvello and N. Ferro (2016). Data Citation is Coming. Introduction to the Special Issue on Data Citation, Bulletin of IEEE Technical Committee on Digital Libraries, Volume 12 Issue 1, pp. 1–5, May 2016.


Thanks

Any Questions?