Analyzing OpenCourseWare usage by means of social tagging

Julià Minguillón\textsuperscript{1,2}, Jordi Conesa\textsuperscript{2}

\textsuperscript{1}UOC UNESCO Chair in e-Learning
\textsuperscript{2}Universitat Oberta de Catalunya
Barcelona, Spain

CAL 2011, 13-15 April, Manchester, UK
Table of contents

• Open Educational Resources
• Social tagging
• Exploratory analysis
• Preliminary results
• Discussion
• Conclusions
• Future research
Open Educational Resources

• Everybody is creating and publishing OERs

• Available through repositories, websites, ...
  – MERLOT, ...
  – OpenCourseWare

• But...
  – Poor metadata descriptions, flat structure
  – Wide degree of granularity
  – Who is (re)using them?

CAL 2011, 13-15 April, Manchester, UK
Searching at OCW

• Limited possibilities:
  – Filter by keyword
  – Language
  – Source (OCW site)

• No shared taxonomies between OCW sites

• **Goal**: improve users’ support when browsing and searching for OERs

*CAL 2011, 13-15 April, Manchester, UK*
Social tagging

- Web 2.0 service (a.k.a. social bookmarking)
- Organize, describe and share URLs
- Triplets: URL, user, {tags}
- Delicious:
  - Largest service of its kind
  - Sustainability: Yahoo wants to close/sell delicious

- Simple poll: how many of you use delicious?
open courseware

Show: Everybody's bookmarks (1000+)

Sponsored Results

Open Courseware
Move your career forward with an accredited online degree!
www.CourseAdvisor.com

Everybody's bookmarks

OpenCourseWare Consortium
open-courseware.org/

Free Online Course Materials | MIT OpenCourseWare
ocw.mit.edu/Ocwlweb/home/home/index.htm

Academic Earth - Video lectures from the world's top scholars
academicearth.org/

OpenCourseWare Finder
opencourse.org/ocwfinder/

Sign in to search your own bookmarks

edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning opensource free opencourserware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learning open courseware
edication learni
OCW Consortium, eta bere plataforma mundoko hozkuntza-edukiak eskuragarriak eta berrirabilizteko prest eskaintzen du.

A free and open digital publication of high quality university-level educational materials. Open Access uniberbitate-mailako irakaskuntzarako publikazio digitalak.

Ikaskuntzarako biltzeki digitala: OCW

mundo osoko hozkuntza-edukiak eskuragarriak eta berrirabilizteko prest eskaintzen du.
What / How do users tag?

- Top level web sites
- Specific links (i.e. courses)

- Web site: mit, ocw
- Content: math, course, pdf
- Usage: phd, ds106, toread

- If delicious is used for content organization, then tags become summarized descriptions

CAL 2011, 13-15 April, Manchester, UK
Basic idea

• What if we obtain “all” tags from “all” users sharing “all” the URLs for a given concept?

• Tags could reveal a hidden semantic structure used for organizing and describing content

• What does “OCW” mean for OCW users?
Methodology

• Retrieve “all” triplets for a given search term (opencourseware + ocw)
• Do some data cleansing
• Select the most important triplets
• Build a binary data set (N x T):
  – URL X tagged by user Y with tags $T_1,...,T_T$
• Perform some exploratory data analysis
• Interpret results

CAL 2011, 13-15 April, Manchester, UK
Retrieving data from delicious

- scripting + wget + HTML parsing
- Problem: IP throttling
- Slow, “risky”, tricky process
- Search term: “opencourseware+OR+ocw”
- Before data cleansing and selection:
  - 2479 URLs
  - 50 pages of users’ tags per URL
  - 3 GB of information
Data cleansing

- Plurals, capital letters
- Misspelling: course, coruse
- Multilingualism: curso, curs, kurs, ...
- Equivalent tags: math = mathematics
- Abbreviations: cs = compsci = computer science
- Compound terms: web + 2.0 = web20
- Spammers
- URLs with no tags
Triplet selection

• Remove those users tagging less than M resources (spammers, occasional users)
• Remove those URLs tagged by less than N users (spammers, broken URLs)
• Keep only a fraction of tags (P) which are the most used (Pareto-like)

• M=5-10, N=5-10, P=80%-90%
Resulting data sets

- 389866 triplets (N):
- Before: 1456 URLs, 26753 users, 21384 tags
- After: 1297 URLs, 24981 users, 434 tags (T)
- Course subset: course URLs at MIT OCW
  - 17024 triplets
  - 351 URLs, 5689 users, 362 tags (120 if P=80%)
- Data set – subset (Non-course):
  - 946 URLs, 12290 users, 412 tags (116 if P=80%)

CAL 2011, 13-15 April, Manchester, UK
# Most used tags

<table>
<thead>
<tr>
<th>Non-course data set</th>
<th>Course data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>- education</td>
<td>- mit</td>
</tr>
<tr>
<td>- resource</td>
<td>- education</td>
</tr>
<tr>
<td>- free</td>
<td>- opencourseware / ocw</td>
</tr>
<tr>
<td>- learning</td>
<td>- course</td>
</tr>
<tr>
<td>- elearning</td>
<td>- programming</td>
</tr>
<tr>
<td>- video</td>
<td>- video</td>
</tr>
<tr>
<td>- teaching</td>
<td>- learning</td>
</tr>
<tr>
<td>- science</td>
<td>- lecture</td>
</tr>
<tr>
<td>- lessonplan</td>
<td>- math</td>
</tr>
<tr>
<td>- web20</td>
<td>- science</td>
</tr>
<tr>
<td>- reference</td>
<td>- python</td>
</tr>
<tr>
<td>- technology</td>
<td>- algorithm</td>
</tr>
<tr>
<td>- online</td>
<td>- free</td>
</tr>
<tr>
<td>- course</td>
<td>- reference</td>
</tr>
<tr>
<td>- lecture</td>
<td>- resource</td>
</tr>
</tbody>
</table>

*CAL 2011, 13-15 April, Manchester, UK*
Exploratory data analysis

- **Goal**: analyze how tags are used
- **R 2.12.1 + psych package**
- **Factor analysis**: 
  - F factors (eigenvalue ≥ 1, F << T)
  - Maximum likelihood
  - Varimax rotation
  - Only factors with weight ≥ 0.1 are considered (**0.3**)
- **Problems**: instability, too many parameters

*CAL 2011, 13-15 April, Manchester, UK*
Preliminary results

• 20.1% of variance explained (20 factors, 16.7%)
• Interesting factors (out of 50):
  – ML11: programming, algorithm, tutorial, howto
  – ML17: algebra, linear, math, linearalgebra
  – ML21: scheme, lisp, sicp, programming, book
  – ML29: psychology, neuroscience, brain, cognitive
• Using P=90% (361 tags) also generates:
  – ML31: politics, society
  – ML39: design, usability, ui, hci, interface

CAL 2011, 13-15 April, Manchester, UK
Discussion

• Tags related to web sites are too general...
• ...but both data sets share \( \approx 50\% \) of tags
• General (i.e. most used) tags are not relevant for factoring purposes (mit, ocw, course, ...)  
• Factors group tags semantically related, mostly describing the domain knowledge
• \( P=80\% \) seems to produce more stable results, but interesting factors are missed (i.e. HCI)

CAL 2011, 13-15 April, Manchester, UK
Conclusions

• OCW could take advantage of “advanced” users tagging its courses
• Non “tech” domains are underrepresented
• Data cleansing is critical
• Proposal: describe domain (keywords), type of resource (book, course) and format (pdf, video)
• Some parameters need fine-tuning
• More experimentation is needed!

CAL 2011, 13-15 April, Manchester, UK
Future research

• Automate and improve data cleansing
  – Use of Wordnet / DBpedia / ontologies

• Non-Negative Matrix Factorization + clustering

• Build recommendation systems:
  – Tags / Related resources

• Build reputation schemes:
  – Most expert users

• Pilot at http://ocw.uoc.edu
Thank you!

Julià Minguillón
jminguillona@uoc.edu

CAL 2011, 13-15 April, Manchester, UK