SOCIAL SEMANTIC WEB AND CROWDSOURCING

ELENA SIMPERL, UNIVERSITY OF SOUTHAMPTON, UK

E.SIMPERL@SOTON.AC.UK, @ESIMPERL

WITH CONTRIBUTIONS FROM “COMBINING THE SOCIAL AND THE SEMANTIC WEB”, TUTORIAL @ESWC2011 BY SHEILA KINSELLA, DERI, IE AND DENNY VRANDECIC, KIT, DE,

HTTP://SEMANTICWEB.ORG/WIKI/COMBINING_THE_SOCIAL_AND_THE_SEMANTIC_WEB
FUNDAMENTALS OF THE SOCIAL WEB
THE SOCIAL WEB

Software deployed over the Web, designed and developed to foster and leverage social interaction

Users are connected via common digital artifacts they share and through social ties

- Sharing platforms such as flickr
- Collaborative platforms such Wikipedia, discussion forums, Quora, Groupon etc
- Social networks such as Facebook combining both aspects
DEVELOPMENT

Early ‘90s

• Read-only, static Web sites
• Interaction via posts on personal Web sites possibly responding to someone else’s statements

Mid '90s

• Read-write applications
• Gradual increase in interaction between site owners and their readers/customers through various communication features

Late ‘90s to present

• First social sites, increasingly ubiquitous deployment of social features in non-social Web sites
CHALLENGES OF SOCIAL SOFTWARE

Many isolated communities

• Data silos

Challenges

• Data cannot be shared and reused easily
• No control over the management and transfer of personal data
• User engagement and incentives engineering
SEMANTIC SOCIAL SOFTWARE
THE ROLE OF SEMANTIC TECHNOLOGIES

SEMANTICS FOR...

Advanced information management
Data interoperability
Data privacy
Data liberation

SCENARIOS

1. Semantic technologies as integral part of the functionality of social software, enabling interoperability

2. Enhanced information management through the usage of machine-processable semantics
ONTLOGIES

FOAF (Friend-of-a-Friend)
SIOC (Semantically-Interlinked Online Communities)
OPO (Online Presence Ontology)

Tagging ontologies
OPEN GRAPH

Represent Web content in a social graph in an interoperable way

Design decisions, see http://www.scribd.com/doc/30715288/The-Open-Graph-Protocol-Design-Decisions

Used by Facebook (‘stories’), Google (snippets), IMDb etc.

Facebook: actors, apps, objects with metadata to create stories

- Example: Elena has finished reading ‘The Economist’, an object of type Newspaper
- Types with attributes, extensions allowed
- Pre-defined and custom actions on objects

Image from https://developers.facebook.com/docs/opengraph/creating-custom-stories
SOFTWARE PLATFORMS

- Drupal (RDFa), https://www.drupal.org/
- Semantic MediaWiki, https://semantic-mediawiki.org/
COLLABORATIVE TECHNOLOGIES FOR THE SEMANTIC WEB
BASICS

Idea: leverage Social Web technologies and platforms to create and manage Semantic Web content

Two classes of approaches

• Mining user contributions, e.g., turning folksonomies into OWL ontologies, not part of this tutorial
• Crowdsourcing Semantic Web tasks, e.g., GWAPs, paid microtasks, campaigns, contents
"Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential .“

[Howe, 2006]
THE MANY FACES OF CROWDSOURCING

Umbrella of Crowdsourcing

- **Crowdsourcing**: Getting a crowd of people to help you with a task that’s typically performed by a single individual or group. Can be divided into 4 groups:
  - **Microtasks**: Breaking a large project into tiny, well-defined tasks for a crowd of workers to complete. Great for: Data validation, research, image tagging, translation. Major platforms: Mturk, Microtask.com, Clickworker, Lingotek.
  - **Macrotasks**: Presenting a project to the crowd & asking them to get involved with the portions they’re knowledgeable in. Participants are empowered to determine the best course of action. Great for: R&D, product innovation. Major platforms: Quirky, Innocentive, Chaordix.
  - **Contests**: Asking a crowd for work and only providing compensation to the chosen entries. Great for: Logo design, business names. Major platforms: 99designs, crowdSPRING, Squadhelp.
  - **Crowdfunding**: Asking a crowd to donate a defined amount of money for a specified cause, project, or other use within a predetermined timeframe. If your goal isn’t met, all donations are refunded. Great for: Project fundraising, disaster relief, artistic support, startups, market research. Major platforms: Kickstarter, crowdrise, Seedrs.

Image from [http://dailycrowdsourcing.com/](http://dailycrowdsourcing.com/)
RELATED: HUMAN COMPUTATION

Outsourcing tasks that machines find difficult to solve to humans
IN THIS TALK:
CROWDSOURCING DATA

‘The USEWOD experiment’

- Goal: collect information about the usage of Linked Data sets in research papers
- Explore different crowdsourcing methods
- Online tool to link publications to data sets (and their versions)
- 1st feasibility study with 10 researchers in May 2014

http://prov.usewod.org/

9650 publications
HOW TO CROWDSOURCE

What: Goal
Who: Contributors
How: Process
Why: Motivation and incentives

http://sloanreview.mit.edu/article/the-collective-intelligence-genome/
EXAMPLE: CITIZEN SCIENCE

WHAT IS OUTSOURCED

• Object recognition, labeling, categorization in media content

WHO IS THE CROWD

• Anyone

HOW IS THE TASK OUTSOURCED

• Highly parallelizable tasks
• Every item is handled by multiple annotators
• Every annotator provides an answer
• Consolidated answers solve scientific problems

WHY DO PEOPLE CONTRIBUTE

• Fun, interest in science, desire to contribute to science
Intrinsic vs extrinsic motivation

Rewards/incentives influence motivation.
- They are granted by an external ‘judge’

Successful volunteer crowdsourcing is difficult to predict or replicate
- Highly context-specific
- Not applicable to arbitrary tasks

Reward models often easier to study and control (if performance can be reliably measured)
- Not always easy to abstract from social aspects (free-riding, social pressure)
- May undermine intrinsic motivation
WHY (2)

In our example:
Who benefits from the results
Who owns the results

Money
• Different models: pay-per-time, pay-per-unit, winner-takes-it-all
• Define the rewards, analyze trade-offs accuracy vs costs, avoid spam

Love
• Authors, researchers, Dbpedia community, open access advocates, publishers, casual gamers etc.

Glory
• Competitions, mentions, awards
WHAT

In general: something you cannot do (fully) automatically

In our example: annotating research papers with data set information

• Alternative representations of the domain
• What if the paper is not available?
• What if the domain is not known in advance or is infinite?
• Do we know the list of potential answers?
• Is there only one correct solution to each atomic task?
• How many people would solve the same task?
WHO

In general
- An open (‘unknown’) crowd
- Scale helps
- Qualifications might be a pre-requisite

In our example
- People who know the papers or the data sets
- Experts in the (broader) field
- Casual gamers
- Librarians
- Anyone (knowledgeable of English, with a computer/cell phone etc)
- Combinations thereof
HOW: PROCESS

See also [Quinn & Bederson, 2012]

In general:

• Explicit vs. implicit participation
  • Affects motivation
• Tasks broken down into smaller units undertaken in parallel by different people
  • Does not apply to all forms of crowdsourcing
• Coordination required to handle cases with more complex workflows
  • Sometimes using different crowds for workflow parts
  • Example: text summarization
• Task assignment to match skills, preferences, and contribution history
  • Example: random assignment vs meritocracy vs full autonomy
• Partial or independent answers consolidated and aggregated into complete solution
  • Example: challenges (e.g., Netflix) vs aggregation (e.g., Wikipedia)
HOW: PROCESS (2)

In our example:

• Use the data collected here to train a IE algorithm
• Use paid microtask workers to go a first screening, then expert crowd to sort out challenging cases
• What if you have very long documents potentially mentioning different/unknown data sets?
• Competition via Twitter
  • ‘Which version of DBpedia does this paper use?’
  • One question a day, prizes
  • Needs golden standard to bootstrap and redundancy
• Involve the authors
  • Use crowdsourcing to find out Twitter accounts, then launch campaign on Twitter
  • Write an email to the authors…
• Change the task
  • Which papers use Dbpedia 3.X?
    • Competition to find all papers
HOW: VALIDATION

In general:

• Solutions space closed vs. open
  • Redundancy vs. iterations/peer-review
• Performance measurements/ground truth
  • Available and accepted
• Automated techniques employed to predict accurate solutions

In our example:

• Domain is fairly restricted
  • Spam and obvious wrong answers can be detected easily
  • When are two answers the same?
  • Can there be more than one correct answer per question?
• Redundancy may not be the final answer
  • Most people will be able to identify the data set, but sometimes the actual version is not trivial to reproduce
• Make educated version guess based on time intervals and other features
Ontology for microtask crowdsourcing V0.1

Task
- creationDate
- expirationDate
- hasTemplate
- hasInputFile

TaskCardinality
- One-to-one
- Many-to-one
- Many-to-many

Who
- Service
- Requestor
- Contributor
- BlockedContributor

Task
- TaskDescription
- TaskRequirement
- hasDescription

QualityAssessmentMethod
- prov:wasDerivedFrom
- hasQualityAssessmentMethod

Why
- Reward
- Bonus
- hasReason

How
- ValidationQuestion
- Question
- hasRightAnswer
- prov:wasDerivedFrom
- hasQuestion

Result
- hasResult
- hasValue
- hasRightAnswer

AggregationMethod
- Average
- Voting
- Consensus
- hasThreshold
- OtherAggregationMethod
- hasSpecification

AnswerInstance
- hasValue
- prov:hadMember
- prov:wasGeneratedBy

Provenance
- hasStatus
- hasLocation
- hasValue
- ReputationSpecification
- type

Collection
- Available
- InProgress
- UnderReview
- Finished
- Disabled

Image courtesy of M. Acosta

EXAMPLES

03.09.14
TASTE IT! TRY IT!

Restaurant review Android app developed in the Insemtives project
Uses Dbpedia concepts to generate structured reviews
Uses mechanism design/gamification to configure incentives
User study

- 2274 reviews by 180 reviewers referring to 900 restaurants, using 5667 DPpedia concepts


03.09.14
ONTOPRONTO

GWAP ~ Game with a purpose
Players’ contributions are used to improve the results of technically challenging tasks

OntoPronto

- Task: map Wikipedia topics to classes of an ontology
- Selection-agreement game: players have to agree on answers
- Step 1: decide whether a topic is a class or an instance
- Step 2: classify topic in the ontology

03.09.14
URBANOPOLY

CROWDMAP

Experiments using MTurk, CrowdFlower and established benchmarks

Enhancing the results of automatic techniques

Fast, accurate, cost-effective

See also [Sarasua, Simperl, Noy, ISWC2012]

<table>
<thead>
<tr>
<th>Concept A: Misc</th>
<th>Concept B: Misc</th>
<th>Is Concept A the same as Concept B? (required)</th>
<th>Select the name of Concept A (required)</th>
<th>How many distinct words are in the name of Concept A? (required)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Concept A: Misc</th>
<th>Misc is a kind of: Reference</th>
<th>Other elements that are of kind Reference: &quot;Academic&quot; &quot;Informal&quot; &quot;MotionPicture&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept B: Misc</td>
<td>Misc. is a kind of: REFERENCE</td>
<td>Other elements that are of kind REFERENCE: &quot;Book&quot; &quot;Academic&quot; &quot;Motion_picture&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept A: Misc</th>
<th>Misc is a kind of: Reference</th>
<th>Other elements that are of kind Reference: &quot;Academic&quot; &quot;Informal&quot; &quot;MotionPicture&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept B: Misc</td>
<td>Misc. is a kind of: REFERENCE</td>
<td>Other elements that are of kind REFERENCE: &quot;Book&quot; &quot;Academic&quot; &quot;Motion_picture&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How many distinct words are in the name of Concept A? (required)</th>
<th>Please write the number in the text box</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CartP 301-304</th>
<th>100R50P Edas-lasted</th>
<th>100R50P Ekaw-lasted</th>
<th>100R50P Cmt-Ekaw</th>
<th>100R50P ConfOf-Ekaw</th>
<th>Imp 301-304</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRECISION</td>
<td>0.53</td>
<td>0.8</td>
<td>1.0</td>
<td>1.0</td>
<td>0.93</td>
<td>0.73</td>
</tr>
<tr>
<td>RECALL</td>
<td>1.0</td>
<td>0.42</td>
<td>0.7</td>
<td>0.75</td>
<td>0.65</td>
<td>1.0</td>
</tr>
</tbody>
</table>
DBPEDIA CURATION

Contest
Linked Data experts
Difficult task
Final prize

Microtasks
Workers
Easy task
Micropayments

Find
Verify

See also [Acosta et al., ISWC 2013]
CrowdQ

Understand the meaning of a keyword query
Build a structured (SPARQL) query template
Answer the query over Linked Open Data

Q = birthdate of actors of Forrest Gump

Query annotation
- Noun
- Noun
- Named entity

Verification
- Is Forrest Gump this entity in the query?

Entity Relations
- Which is the relation between: actors and Forrest Gump → starring

Schema element
- Starring → dbpedia-owl:starring

Verification
- Is the relation between:
  - Indiana Jones – Harrison Ford
  - Back to the Future – Michael J. Fox
  - of the same type as
  - Forrest Gump – actors

Gianluca Demartini, Beth Trushkowsky, Tim Kraska, and Michael Franklin. **CrowdQ: Crowdsourced Query Understanding.** In: 6th Biennial Conference on Innovative Data Systems Research (CIDR 2013)
PROBLEMS AND CHALLENGES

• What is feasible and how can tasks be optimally crowdsourced, e.g., translated into microtasks?
  • Examples: data quality assessment for technical and contextual features; subjective vs objective tasks (also in modeling); open-ended questions

• What to show to users
  • Natural language descriptions of Linked Data/SPARQL
  • How much context
  • What form of rendering
  • How about links?

• How to combine with automatic tools
  • Which results to validate
    • Low precision (no fun for gamers...)
    • Low recall (vs all possible questions)

• How to embed crowdsourcing into an existing application
  • Tasks are fine granular, perceived as additional burden to the actual functionality

• What to do with the resulting data?
  • Integration into existing practices
  • Vocabularies
FURTHER READING

The Collective Intelligence Genome

A user's guide to the building blocks of collective intelligence: By recombining CI “genes” according to the work required.

SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND PRACTICE

Series Editors: James Hendler, Rensselaer Polytechnic Institute
Ying Ding, Indiana University

Incentive-Centric Semantic Web Application Engineering
Elena Simperl, University of Southampton, United Kingdom
Roberta Cueli, University of Trento, Italy
Martin Stein, University of Siegen, Germany

While many Web 2.0-inspired approaches to semantic content authoring do acknowledge monetary incentives as the main drivers of user involvement, the amount of useful human contributions actually made will always remain a scarce resource. In this context, there are aspects of semantic content development which automatic techniques have proven to perform reliably, and the added value of human (and curation) intelligence is often a question of cost and timing. The challenge that this book attempts to tackle is to reconcile these two approaches (machine- and human-driven computation) in order to improve the cost/performance ratio of creating, managing, and meaningfully using semantic content.

To do so, we need to first understand how theories and practices from social sciences and economics of user behavior and incentives could be applied to semantic content authoring. We will introduce a series of studies to help software designers to embed incentive-motivated functionalities into semantic applications, and practice guidelines. We will present several examples of such applications, addressing knowledge management, media annotation, and information extraction, which have been built with consideration for user incentives in mind. These examples illustrate key design issues of incentivized Semantic Web applications that might have a significant impact on the success and sustainable development of the applications.

ABOUT SYNTHESIS
This volume is a selected version of a work that appears in the Synthesis Digital Library of Engineering and Computer Science. Synthesis Lectures provide concise, original presentations of important research and development topics, published quickly, in digital and print formats. For more information, visit www.morganclaypool.com