A Semantic Tool to Support Navigation in a Folksonomy

David Laniado

August, 23 2007
HP Research Labs, Palo Alto
Collaborative Systems

- **The cathedral**
  - "experts"
  - central authority
  - isolation

- **The bazaar**
  - "users-developers"
  - decentralization
  - information sharing
  - "organized chaos"
Folksonomies

- *folk* + *taxonomy* (Vander Wal, 2004)
- users tag resources with freely chosen keywords
- two kinds of folksonomies:
  - *broad*: each resource can be tagged by any user
  - *narrow*: each resource can be tagged only by some users
- a way to “remember in public” (J. Schachter, 2004)
- different contexts:
  - social bookmarking (del.icio.us)
  - photographs (Flickr)
  - scientific papers (Bibsonomy)
  - blog posts (Technorati)
Taxonomies vs Folksonomies

- top-down
- experts do the job
- central authority
- hierarchies of concepts
- controlled vocabularies
- *searching* and *browsing*

- bottom-up
- users do the job
- “democracy”
- flat space
- freely chosen keywords
- *finding* and *serendipity*
Folksonomies: advantages

- inclusive
- democratic
- actual
- desire lines
- reflecting the users point of view
- overlapping
- low cost
Folksonomies: limitations

- no synonym control
- lack of precision
- lack of recall
- lack of hierarchy
- possibility of *gaming*
- lack of standards
A Semantic Tool to Support Navigation in a Folksonomy

David Laniado
A Semantic Tool to Support Navigation in a Folksonomy

David Laniado
Context choices

• a popular broad folksonomy -> del.icio.us
• ontology -> hypernymy relation on WordNet nouns: general categories implicitly contained in natural language
• *active navigation* -> Firefox + Greasemonkey: browser extension to dynamically enrich the pages with additional semantic information
• programming language -> Perl
System architecture

[Diagram showing the system architecture with components such as Browser, del.icio.us, Server, Crawler, Tree builder, Database, WordNet, Disambiguator, and Grease-monkey.]
Tags contained in WordNet

- Data relative to 30,000 del.icio.us users:
  - 480,000 different tags
    - only 8% of tags are contained in WordNet
  - over 20 million tagging actions
    - 68% of the times the tag used is contained in WordNet
Tag disambiguation

- polisemy: words can have more than one meaning -> need to solve ambiguities

- disambiguation of each tag in relation to the resource it is referred to
  - context: the other tags referring to the same resource
  - correlation metrics: adapted lesk, Hirst & St. Onge

- algorithm:
  - for each resource, take the first N most used tags, and compare them to each other to find the most likely meaning
  - disambiguate each of the remaining tags by comparing it to the first N

- advantages:
  - less noise sensitive: based only on the most used tags
  - efficient: the number of comparisons to be effected doesn't grow exponentially with the number of tags attributed to the same resource
The tag semantic tree

- the tree is a subset of WordNet concept hierarchy, containing all the tags related to the chosen one
- each node represents a (disambiguated) tag or a concept, that is hypernym of at least one tag
- each node has two attributes:
  - the node's weight (number of resources tagged with that word in that acceptation)
  - the subtree weight (the sum of the weights of all nodes beneath)
- steps:
  - tree building
  - compression
  - branches ordering, according to their weight
  - result output, in HTML or XML format
**Tree building**

- for each tag in a particular acceptation:
  - build the hypernym chain from the tag to WordNet roots of concepts
  - *merge* the new path with the existing tree, starting by the root
- as the tree has limited dimensions, the algorithm complexity is linear with the number of tags received in input and depends on:
  - the medium branching factor \( b \) of the tree
  - the tree depth \( d \)
Tree compression

- problem: WordNet's fine granularity
- need: compressing the tree without making it too flat or losing any interesting concept

- algorithm:
  - delete nodes corresponding to high level categories, fixed in a list (such as "physical entity" or "causal agent")
  - delete internal nodes that:
    - do not correspond to any tag and
    - have less than K children or have no siblings
- default value for K is 2, so the hierarchy structure is preserved, but more specific concepts can rise in the tree
A Semantic Tool to Support Navigation in a Folksonomy

David Laniado
A Semantic Tool to Support Navigation in a Folksonomy

- food (102)
  - pasta (283)
    - macaroni (9)
    - noodle (9)
    - spaghetti (6)
    - ravioli (3)
    - gnocchi (2)
    - penne (2)
    - vermicelli (1)
    - tortellini (1)
    - lasagna (1)
    - ravioli small ring-shaped stuffed pasta
    - lasagne (1)
    - fettuccine (1)
  - produce
    - vegetable (12)
    - edible_fruit
    - veggie (2)
  - meat (5)
    - poultry
      - chicken (13)
      - turkey (1)
  - related tags
    - cooking
    - recipes
    - vegetarian
    - italian
    - healthy
    - main
    - parmesan
    - food
    - sauce
    - sausage
    - gnocchi
A Semantic Tool to Support Navigation in a Folksonomy

sites tagged with pasta and broccoli on del.icio.us
A Semantic Tool to Support Navigation in a Folksonomy

David Laniado
A Semantic Tool to Support Navigation in a Folksonomy

David Laniado
A Semantic Tool to Support Navigation in a Folksonomy

David Laniado
Multiple Personality Disorder-DID  save this
by ademajo to MentalHealth mental health medicine interesting
dissociative mind&body ... 19 mins ago

RIDE Project - Semantic technology for eHealth  save this
by sionbach to snomed health semanticweb ... saved by 1 other person ... 19 mins ago

Natural Medicine, Herbal Remedies–Treatment of ADD ADHD Anxiety UTI and more  save this
Great Natural Remedy site
by dluntz to herbal natural treatment remedies health ... saved by 6 other people ... 20 mins ago

MediCorp Health System | Community Calendar  save this
Classes at Mary Washington Hospital
by endswell to health scouts ... 20 mins ago

Navy SEALs.com - US Navy SEAL Workout  save this
by dajamerson to Fitness workout health ... saved by 392 other people ... 20 mins ago

'Cannabis' may help mentally ill  save this
by deanjb to health bipolar cannabis medical ... 22 mins ago

www.whonamedit.com  save this
by chimpsky to reference science search tools health history ... saved by 89 other people ... 22 mins ago

blog.myspace.com/independz  save this
by ralphjarmon to health blog ... 25 mins ago

Aging - State of Aging and Health Report  save this
by afeeney to aging health demographics healthcare ... saved by 1 other

David Laniado  A Semantic Tool to Support Navigation in a Folksonomy
Conclusions

- Filtering tags through a taxonomy of concepts, implicitly contained in a language lexicon, may improve the possibilities of exploring a folksonomy.

- Some of the main limitations of folksonomies can be partially solved:
  - lack of hierarchy
  - synonym control
  - gaming

- Future work:
  - Improving the tree compression algorithm
  - Integrating WordNet with domain ontologies and multilingual wordnets
  - Extracting data from different folksonomies
  - Using tag disambiguation to filter the resources