

# **UNIVERSITY OF** CAMBRIDGE **Computer Laboratory**



### I. THE SITUATION

### On a typical **tagging website**

(e.g. Delicious, LastFM, Bibsonomy, LibraryThing etc.):

- Multiple users can assign tags (keywords) to the same document
- Each document forms a *tag cloud*, that visualises tag popularity within the document
- ► The entire collection of documents forms a *folksonomy*, i.e. a "folk" (crowd-sourced, emerging) "taxonomy" of documents

Here is a folksonomy of pictures:

- Users assign tags to images
- Some users have the same 'opinion'
- Clouds form for each image:
- ▶ e.g. "landscape" is large (popular) in the first tag cloud

mountains cezanne trees landscape green nature countryside

### **II. THE IDEA**



Look at a tag cloud! It resembles a paragraph summarising how the picture is perceived by the general public. This paragraph is very fragmented. Can we fill in the gaps? Can we re-create (parts of) the underlying paragraph?

**HOW?**  $\rightarrow$  Starting with simple triples like Noun1 – relation – Noun2

- Noun1 and Noun2 are tags usually found in corpora as nouns
- relation is whatever stretch of language can connect the nouns and make a 'statement' about the picture
- Nouns can be enriched with adjectives etc.

**WHY?**  $\rightarrow$  It can help in:

- automatic caption generation
- more accurate search

# **Uncovering Implicit Relations in Folksonomy Theodosia Togia** Natural Language and Information Processing Group, Computer Laboratory University of Cambridge





brushstroke bluemountains summer analytical angular annegreenleaf architecture autumn block blue geometric calm Cezanne coolcolors cottage country countryhouse deepgreentrees dynamic earthtones english farmhouse farmland fields france geometricmasses green hills hillside horizon house housenestledinalandscape impressionistic and scape layers linear modernism montsaintvictoire montstevictoire mountain mountains nature ochre orange orchard painting pastoral peace placid postimpressionist postimpsressionistcomposition provence quaint quite rectangularbrushstrokes

scenery scottish seagreen shapes sky slategray southeuropeanstyle Southoffrance strong countryside reduced trees village postimpressionism french irish rugged rural

house * countryside
painting * landscape> painting of a
postimpressionist painting * Cezanne -> post-impression
house * blue mountains> house surro
trees * house> trees aroun
landscape * southoffrance> landscape in

For questions ana bibliography please contact me on tt309@cam.ac.uk

countryside

landscape

sionist painting by Cezanne

unded by blue mountains

the house

South of France



### **III. THE PROCESS**



the task:

**STEPS:** 

- "house nestled in a landscape")

- painting by Cezanne")

# IV. THE METHOD



**DATASETS**  $\rightarrow$  folksonomies & supporting corpora Steve Musuem image folksonomy Wikiwoods corpus, BNC (British National Corpus)

# MAIN TECHNIQUES

- Distributional Semantics

- using corpora
- V. PAST, CURRENT & FUTURE WORK



**DONE**  $\rightarrow$  Steps 1, 2, 3 and (partly) Steps 4 and 5 **IN PROGRESS**  $\rightarrow$  analysing recently collected human data (208 participants providing both paragraphs and *tags* for images). We compare text vs. tags in order to: see what kind of text is underlying tag clouds perform some initial relation extraction **TO BE DONE**  $\rightarrow$  corpus- and search-engine-based relation extraction & (human) evaluation



**FOCUS**  $\rightarrow$  on image folksonomies because this makes

more useful (generating text for non-textual data) more interesting (no supporting text to help the task)

1. Split multi-word tags (e.g. "housenestledinalandscape"  $\rightarrow$ 2. Find tags that are likely to act as nouns (e.g. "mountains") 3. Find pairs of related noun-tags, i.e. ones that it is worth extracting relations for (e.g. "painting" and "cezanne"). 4. Extract possible natural language 'relations' between each pair (e.g. "painting by Cezanne", "painting composed by Cezanne", "Cezanne *is the artist of this* painting" etc.) 5. Identify possible collocations (e.g. "post-impressionist" + "painting") and expand the triples (e.g. "post-impressionist

to find 'related' pairs of tags in the folksonomy (Step 3 above) to find collocations from corpora (Step 5 above) Paraphrase-type noun-noun compound *Relation Extraction* 

using wildcard search engine queries (e.g. "trees \* house")