Deriving Dynamic Knowledge From Academic Social Tagging Data: A Novel Research Direction

Hang Dong[1,2], Wei Wang[2], Frans Coenen[1]
1. Department of Computer Science, University of Liverpool; 2. Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University

Introduction & Research Aim

- **Academic social tagging** is the phenomenon that researchers using flexible keywords to annotate papers, websites and other resources online.
- A novel research direction is to investigate **whether the extensive collection of academic tags reflects the evolution of scientific knowledge**.
- The aim of the proposed research is to identify algorithms to model the evolution of knowledge from user-generated tags in academic social media platforms. **In this poster, we present a Data Clean workflow for Academic Social Tagging Data, as our preliminary result.**

Research Challenges & Procedure

Challenges 1: Noise in Social Tags as in Table 1 below (adapted from [1])

	Noise types	Examples
Basic Noise	Singular vs. plural	"analysis" and "analyses"
	Verb forms	"analyzed"
	Spelling variations, errors	"analysed", "analys"
	Multiple words in a tag	"speechanalysis", "Time-series_analysis"
	Multilingual tags	Datenanalyse
	Tag with special characters	"Autoantibodies/*analysis/drug"
	Personalised tags	"mythesis"
	Nonsense tags	"28A75"
Semantic Noise	Synonymy	"Mac", "Macintosh" and "Apple“
	Polysemy	"apple"
	Abstraction	"Programming", "Javascript", "perl"
	Association	"Agriculture", "permaculture"

Challenges 2: Sparsity in data, most tags are used only few times by few users to annotate few resources.

Four stages in the proposed research

Data Cleaning: To develop a workflow for social tag cleaning, during which the twin issues of noise and data sparsity are addressed;

Concept Extraction: To use clustering and topic modeling methods to select and extract concepts from social tags;

Relation Learning: To derive relations among complex terms to build formal concept hierarchies for various research domains;

Knowledge Evolution: To model the evolution of knowledge through chronological analysis.

Data Cleaning & Concept Extraction

The Data Cleaning workflow contains 4 stages:
- (1) Specific character handling, (2) Multiword & single tag group extraction,
- (3) Tag selection using selected metrics, (4) Tag selection by language.

The basic idea: using inter-subjectivity (user frequency) and edited distance to group word forms.

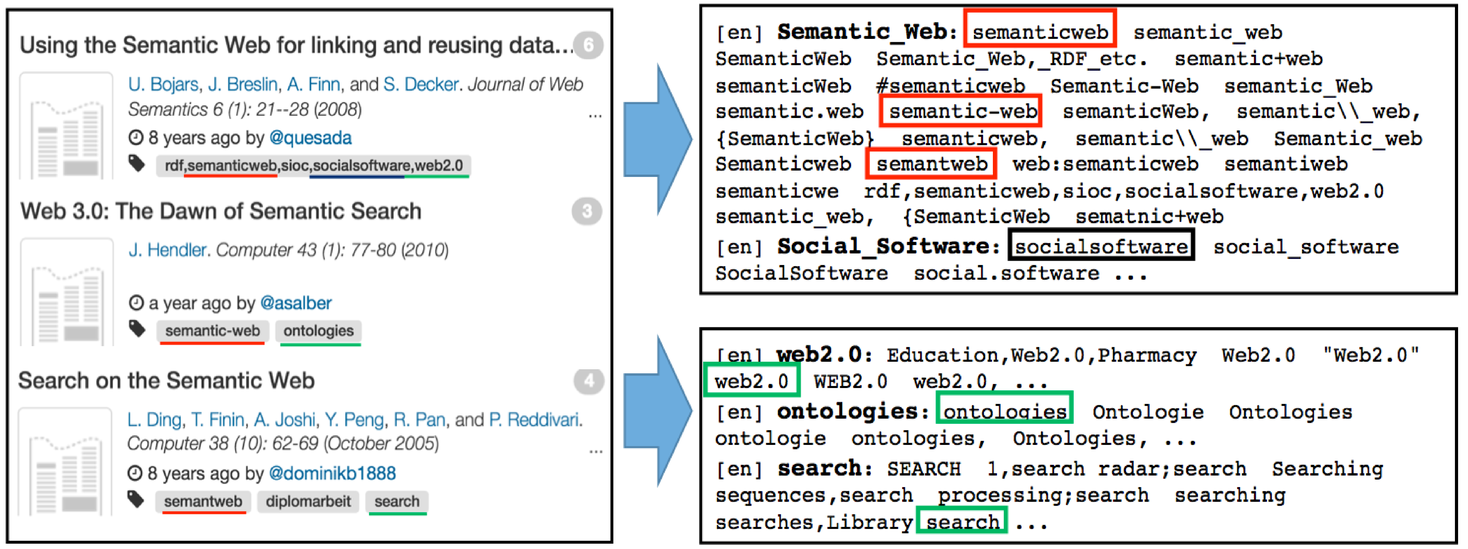

Figure 1 The raw Bibsonomy data (left) were aggregated into groups of standard tags and variants, including multiword tags (right-up) and single word tags (right-down).

	Original dataset	After step 2: Acquisition of multiword and single tag groups	After step 3: Selection by metrics (user frequency >= 4)	After step 4: Selection by language
Multiword tag groups	0	68,168	3,669	2,502
Single tag groups	0	121,902	19,341	14,877
Total tags/tag groups	283,858	190,070	23,010	17,379
Users	11,103	-	-	6,592
Resources	868,015	-	-	663,148

Table 2 Statistics for the Bibsonomy Dataset (2015-07) using Data Cleaning workflow.

Conclusions

- An novel direction is to study the evolution of knowledge from Academic Social Tagging.
- We designed a Data Cleaning workflow for the noisy and sparse academic social tagging data.

Future Work

- **Relation Learning**: supervised learning for tag hierarchy generation with probabilistic topic models.
- **Knowledge Evolution**: chronological analysis, update and visualization of a dynamic tag ontology.

References

[1] Jabeen, F., & Khusro, S. (2015). Quality-protected folksonomy maintenance approaches: a brief survey. *The Knowledge Engineering Review, 30*(05), 521-544.

[2] Garcia-Silva, A., Corcho, O., Alani, H., & Gomez-Perez, A. (2012). Review of the state of the art: Discovering and associating semantics to tags in folksonomies. *The Knowledge Engineering Review, 27*(01), 57-85.

hangdong@liverpool.ac.uk