Rules for Inducing Hierarchies from Social Tagging Data

iConference 2018, Sheffield, UK, March 25-28, 2018

Hang Dong, Wei Wang, Frans Coenen

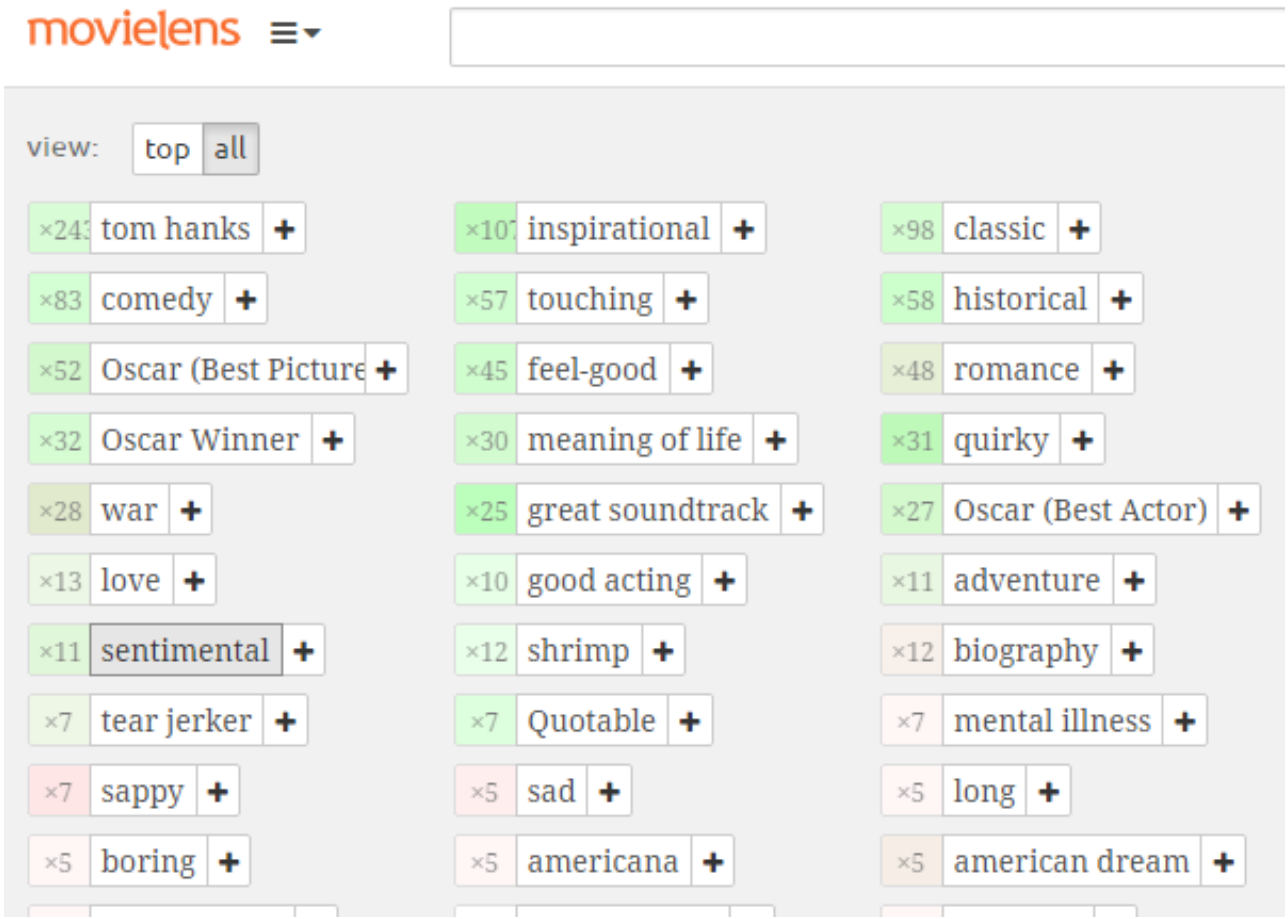
Department of Computer Science,

University of Liverpool

UNIVERSITY OF LIVERPOOL

Social Tagging Data (Folksonomies)

- Users collaboratively generate “key words” for their interests.
- The “key words” form a taxonomy of resources online, called *Folksonomies* (Vander Wal, 2007).



Social tags for movie “Forrest Gump” in MovieLens
https://movielens.org/movies/356

Issues in social tagging data

- (i) Noisy and ambiguous.
  - Data cleaning (Dong, Wang & Coenen, 2017).

- (ii) Plain structure, lack of semantic relations among tags.
  - **This study focuses on hierarchical/subsumption relations between tags.**
  - This is a challenging problem:
    - a cognitive task requiring much human effort (Weller, 2010, p. 139).
    - distinct from mining relations from sentences.

Research Questions

- 1. Which *rule* can effectively capture the hierarchical relations between social tags?

  - - Not systematically discussed in previous studies, although some approaches were proposed & evaluated (Garcia-Silva, 2012; Strohmaier *et al*., 2012).

  - - (Information science & Linguistics) definition of hierarchical relations
  - - Rules in the previous study
  - - Proposed two new rules: Fuzzy set inclusion, Probabilistic Association

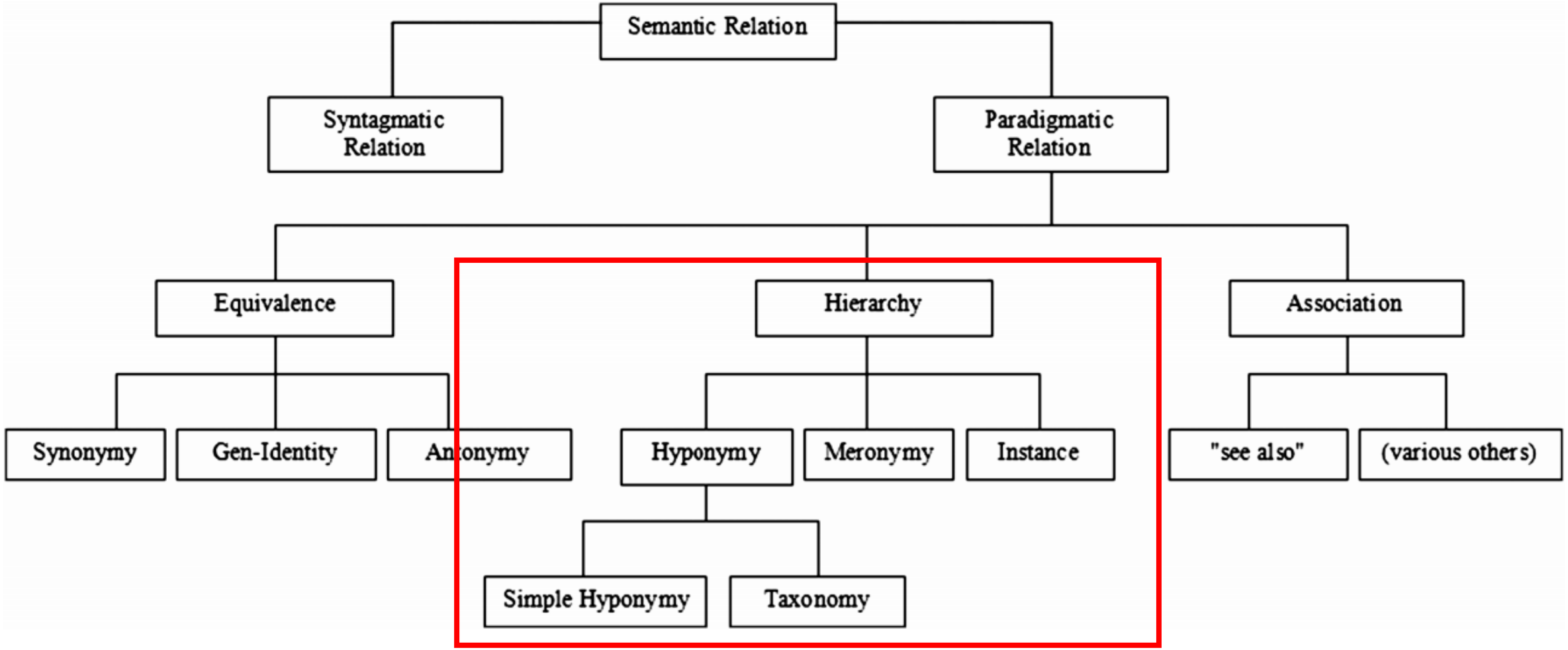
Research Questions (2)

- 2. How do *rules* and *data representations* affect the quality of the induced hierarchies?

  - Data representation: resource-based representation, probabilistic topic representation

  - Experimental Design:
    - Hierarchical Generation Algorithm
    - Automated evaluation against three gold-standard hierarchies

Hierarchical Relations – information science



Acknowledgement to the image in Stock, W. G. (2010). Concepts and semantic relations in information science. *Journal of the Association for Information Science and Technology*, *61*(10), 1951-1969.

Hierarchical Relations

- Straightforward?
  - (i) Apple *is a [kind of]* fruit. (ii) Library science *is a part of* Information Science.
- Abstraction, Generalisation.

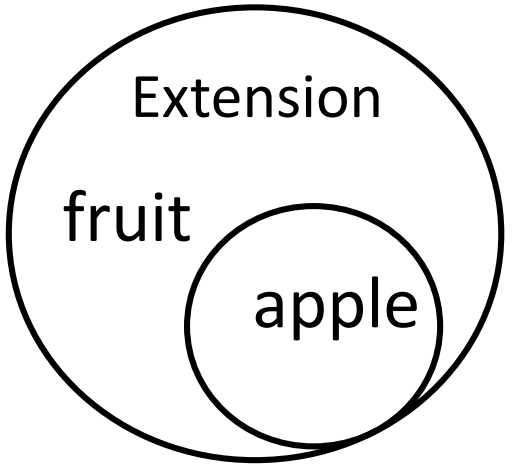
- A type of paradigmatic Relation: fit into the same grammatical slot (Cruse, 2003).

- Tagging data only provide syntagmatic relations, but are a great source for paradigmatic relations (Peters, 2009; Stock, 2010).

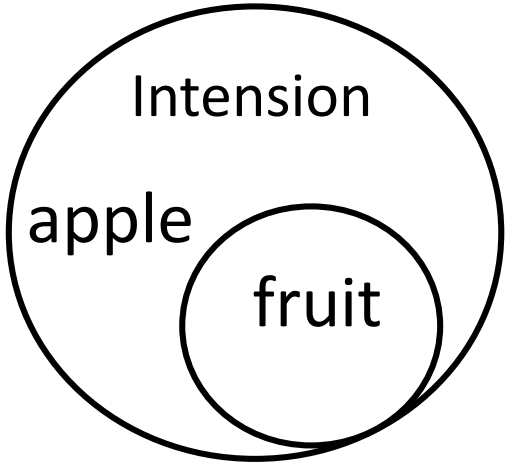
Hierarchical Relations – linguistics (1)

Definitions in Cruse (2003)

- **Logical**: (extensional) X is a hyponym of Y **iff** the extension/objects of *X'* should be included in the extension/objects of *Y'*.
  - Unsymmetrical



(intensional) X is a hyponym of Y **iff** F(X) entails, but is not entailed by F(Y), where F(-) is a sentential function satisfied by X or Y.



Hierarchical Relations – linguistics (2)

Definitions in Cruse (2003)

- **Collocational**: X is a hyponym of Y iff the normal context of X is a subset of the normal context of Y.

  You shall know a word by the company it keeps… - Firth (1957)

- **Componential**: X is a hyponym of Y iff the features defining Y are a proper subset of features defining X.

Hierarchical relations from tags – computational rules

Representing a tag as a vector

- Set Inclusion (Mika, 2007; De Meo, 2009)
- Graph Centrality (Heymann, 2006)

Resource-based (**Res-based**) representation

- Information-Theoretic Condition (Wang, 2010)
- Fuzzy Set Inclusion
- Probabilistic Association

Probabilistic **T**opic **M**odelling (**PTM**) based Representation

Data Representation

- Resource-based Representation:

(Markines et al., 2009)

Vt[*i*] = number of times the tag *t* is annotated to the *i*th resource

resources

tags

	R1	R2	R3
news	1	0	0
Web2.0	1	1	1
knowledge	0	0	1

- Probabilistic Topic Modelling Representation:

Using a probabilistic generative model to infer the **p (tag | topic)** and **p (resource | topic)**

Then calculate p(topic | tag) from p (tag | topic) using Bayesian's Theorem.

topics

tags

	Topic 1	Topic 2	Topic 3
news	0.8	0.1	0.1
Web2.0	0.4	0.3	0.3
knowledge	0.2	0.2	0.6

Each row sums to 1.

Rule 1: Set inclusion (Mika, 2007; De Meo, 2009)

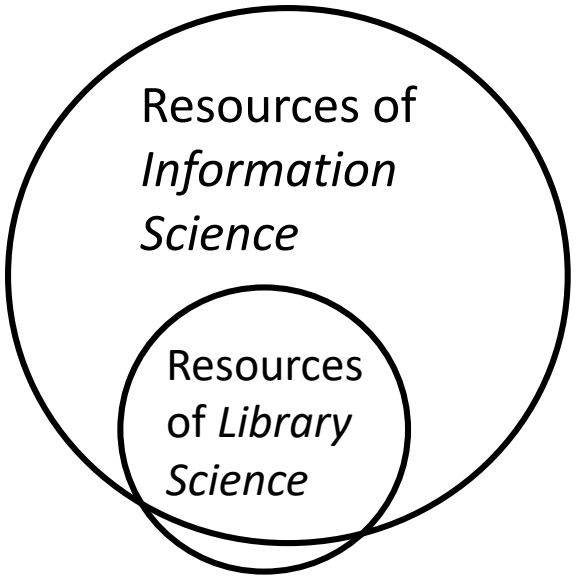
- Tag A is a hyponym of Tag B if set-inc(A, B) >= p ∧ set-inc(B, A) < p ∧ sim(A, B) > s. (p=0.5)

  Sim(A, B) is a similarity measure: cosine similarity.

$$\text{set-inc}(A, B) = \frac{|R_A \cap R_B|}{|R_A|},$$

where R_A means the resource set annotated using the tag A.

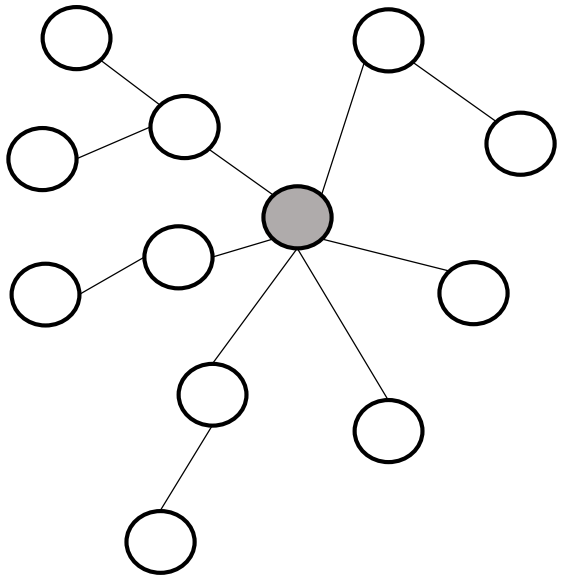
Assumption: The **logical <u>extension</u>** of a tag is measured as its resource context, i.e. the resources that tag is annotated.



Rule 2: Graph Centrality (Heymann, 2006)

- Tag A is a hyponym of Tag B if graph-cent(A) < graph-cent(B) ∧ sim(A,B) > s.

Tag similarity graph, where each node is a tag and edge is established by similarity of tags over a threshold.



graph-cent(A) is a graph centrality measure (centrality, betweenness, etc.) of a tag A in the tag similarity graph.

Assumption: **popularity-generality**
the more popular/influential a tag, the more general it is.
(**collocational**)

Rule 3: Information-Theoretic Condition (Wang, 2010)

- Tag A is a hyponym of Tag B if $\mathbf{D_{KL}(P_B||P_A)-D_{KL}(P_A||P_B) < f}$ Λ sim(A,B) > s. Here P_A and P_B are the probability distributions of A and B over topics. f is a noise factor of a small value (f = 0.05 in this study).

- Kullback-Leibler divergence as a measure of "surprise" of receiving P_B when P_A is expected.

$$D_{\mathrm{KL}}(P_A||P_B) = \sum_i P_{A\,i} \log \frac{P_{A\,i}}{P_{B\,i}},$$

Rule 4: Fuzzy set inclusion

- An extension of Set Inclusion, based on probabilistic topic representation:
- Tag A is a hyponym of tag B if **fuzzy-set-inc(S_A,S_B) >= p** ∧ **fuzzy-set-inc(S_B,S_A)< p** ∧ sim(A,B)>s, where p is set as 0.5.

$$\text{fuzzy-set-inc}(S_A, S_B) = \frac{\sum_i \min(S_{A\,i}, S_{B\,i})}{\sum_i S_{B\,i}}$$

, where S_A is a fuzzy set for tag A as a pair (U, m), U is the set of topics for tag and m:U → [0, 1] is a membership function: for each topic $z \in U$, $m(z) = p(A|z)$.

Set inclusion vs Fuzzy set inclusion

- Resource-based Representation:

Vt[*i*] = number of times the tag *t* is annotated to the *i*th resource

resources

tags

	R1	R2	R3
news	1	0	0
Web2.0	1	1	1
knowledge	0	0	1

- Probabilistic Topic Modelling Representation:

Using a probabilistic generative model to infer the **p (topic | tag)** and **p (resource | topic)**

Use p (topic | tag)

Note: this is different from the previous p (tag | topic)

topics

tags

	Topic 1	Topic 2	Topic 3
news	0.57	0.17	0.1
Web2.0	0.29	0.5	0.3
knowledge	0.14	0.33	0.6

Each column sums to 1.

Rule 5: Probabilistic Association

- Based on PTM representation
- Tag A is a hyponym of Tag B if **p(A|B) < p(B|A)** ∧ sim(A, B) > s,

$$\sum_z p(A|z)p(z|B), \text{ where } p(z|B) \propto p(B|z)p(z).$$

- p(A|B) =

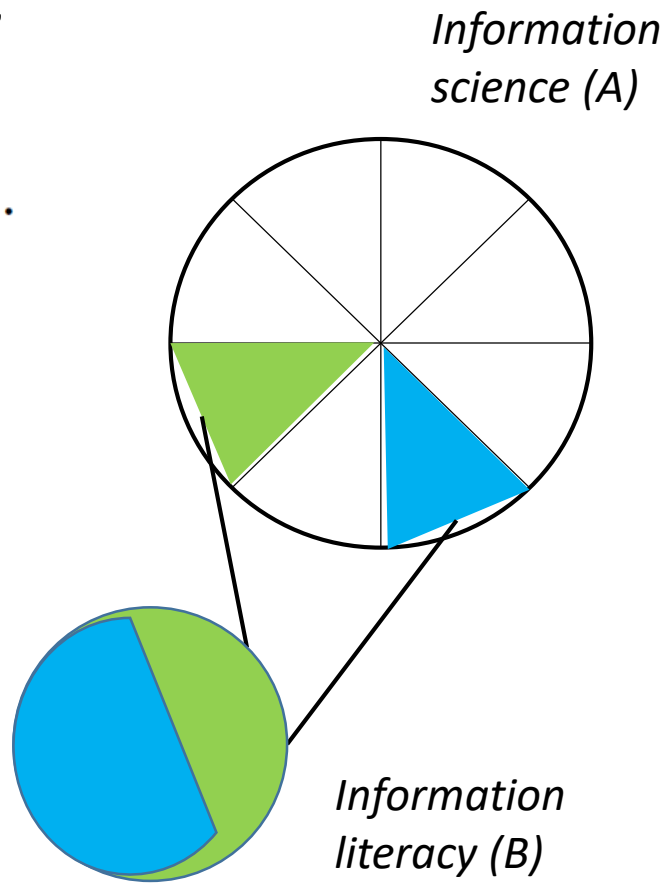
  (Griffith & Steyvers, 2002)
- *z is a member in the set of topics.*

- Assumption: **componential** measure of hierarchical relation.

An example

p (A | B) = 1

p (B| A) = 0.25



Methodology: Algorithm to Hierarchy Generation

Algorithm 1. Hierarchy generation algorithm using *any* rules.

Input: $L_{PairSim}$ is a pre-computed list of tag pairs $< t_i, t_j >$ ranked in descending order by similarity. s is a similarity threshold for a tag pair. $\text{sim}(t_i, t_j)$ computes the cosine similarity between t_i and t_j; $hasParent(t, G)$ returns a boolean indicating whether tag t has a parent node in G; $isHypo(t_i, t_j)$ determines whether t_i is a hyponym of t_j.

Output: G, an induced hierarchy as a directed graph.

```
1  Initialise G;
2  for i ← 1 to |L_PairSim| do
3  |   < t_i, t_j > ← L_PairSim[i];
4  |   if sim(t_i, t_j) < s then
5  |   |   continue to the next i;
6  |   end
7  |   if NOT hasParent(t_i, G) then
8  |   |   if isHypo(t_i, t_j) then
9  |   |   |   G ← G ∪ < t_i, t_j >;
10 |   |   end
11 |   end
12 end
```

- For RQ1 about rules:
  Replacing the isHypo() function to one of the five rules each time, and compare the results.

- For RQ2 about data representations:
  - using different representation to calculate sim(t*i*,t*j*).
  - using the compatible data representation for each rule.

Experiments

- Data Collection and Processing

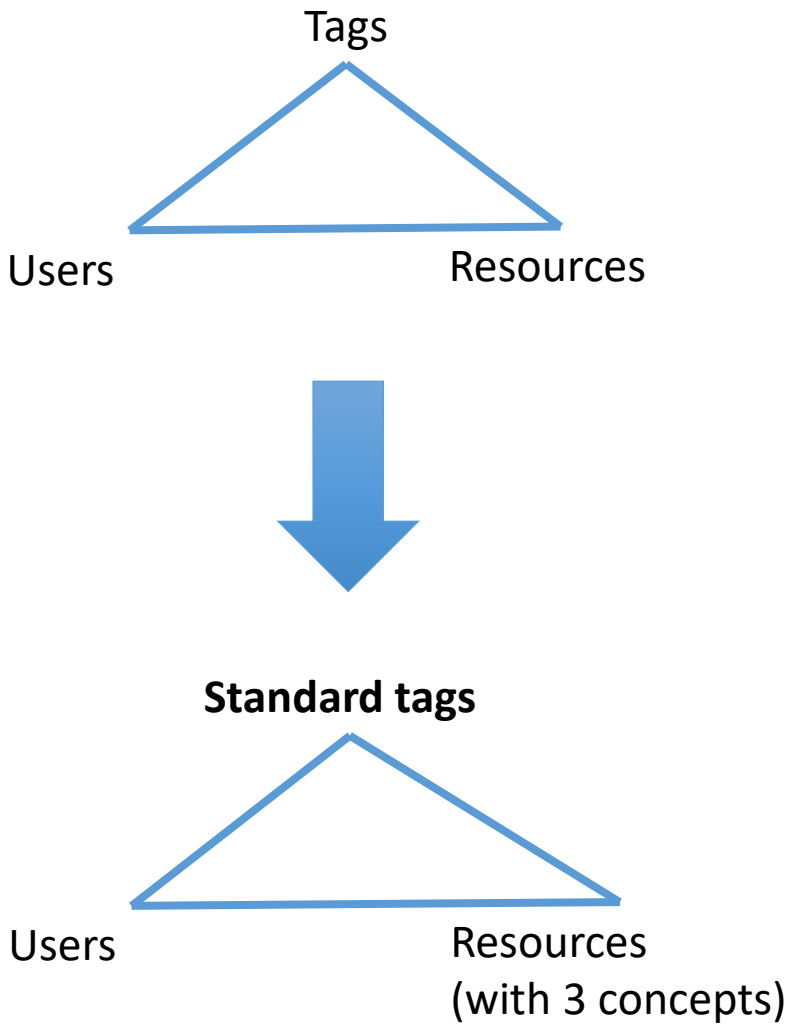
Bibsonomy dataset 2003-2015: 3,794,882 annotations, 868,015 resources, 283,858 tags, 11,103 users.

We used a streamline to clean academic social tagging data (Dong, Wang & Coenen, 2017):

- Unified different variants of tags.
- Selected tags having user frequency >= 4.

- Removed resources with tags < 3

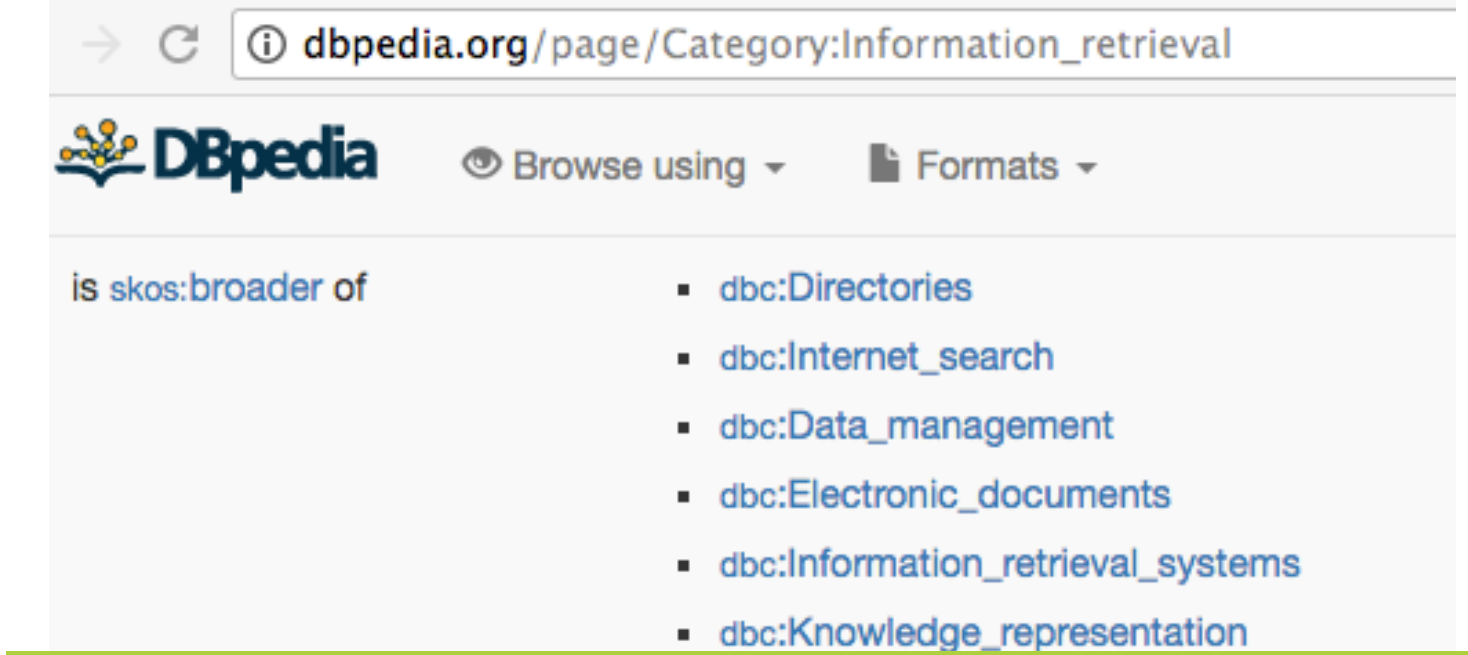
The cleaned dataset contains 7,846 tag concepts and 128,782 resources.

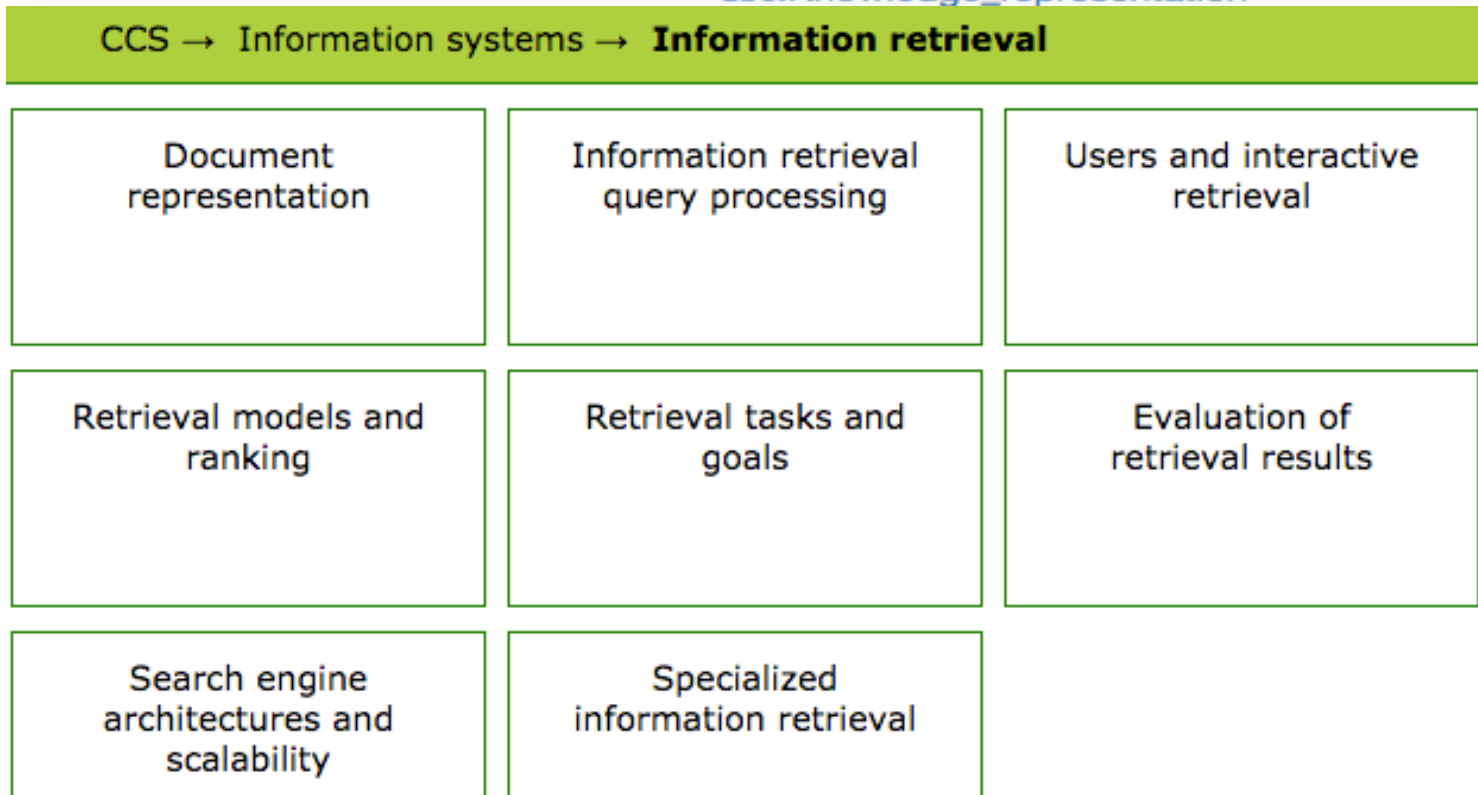


Reference-based evaluation

Measuring the similarity of a learned hierarchy, *L*, to gold-standard hierarchies.

- Gold-standard, denoted as *G*:
  - DBpedia (6616 concepts overlap)
  - Microsoft Concept Graph (6029 concepts overlap)
  - ACM computing classification system (691 concepts overlap)





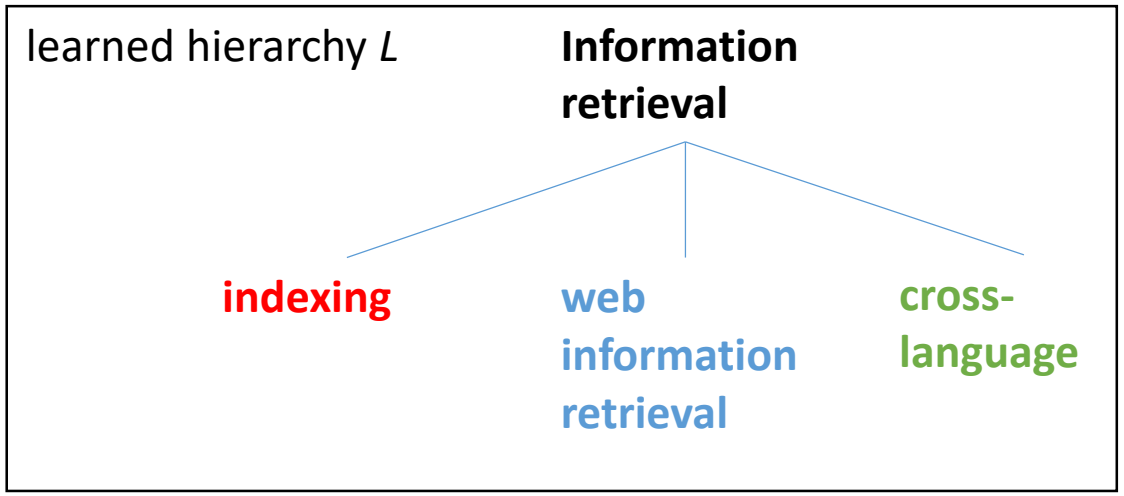
Acknowledgment to Images in
http://dbpedia.org/page/Category:Information_retrieval and
https://dl.acm.org/ccs/ccs.cfm?id=10003317&lid=0.10002951.10003317

Evaluation metrics (Dellschaft, Staab, 2006)

(i) Find *common concepts* between the learned hierarchy *L* and the gold-standard hierarchy *G*,

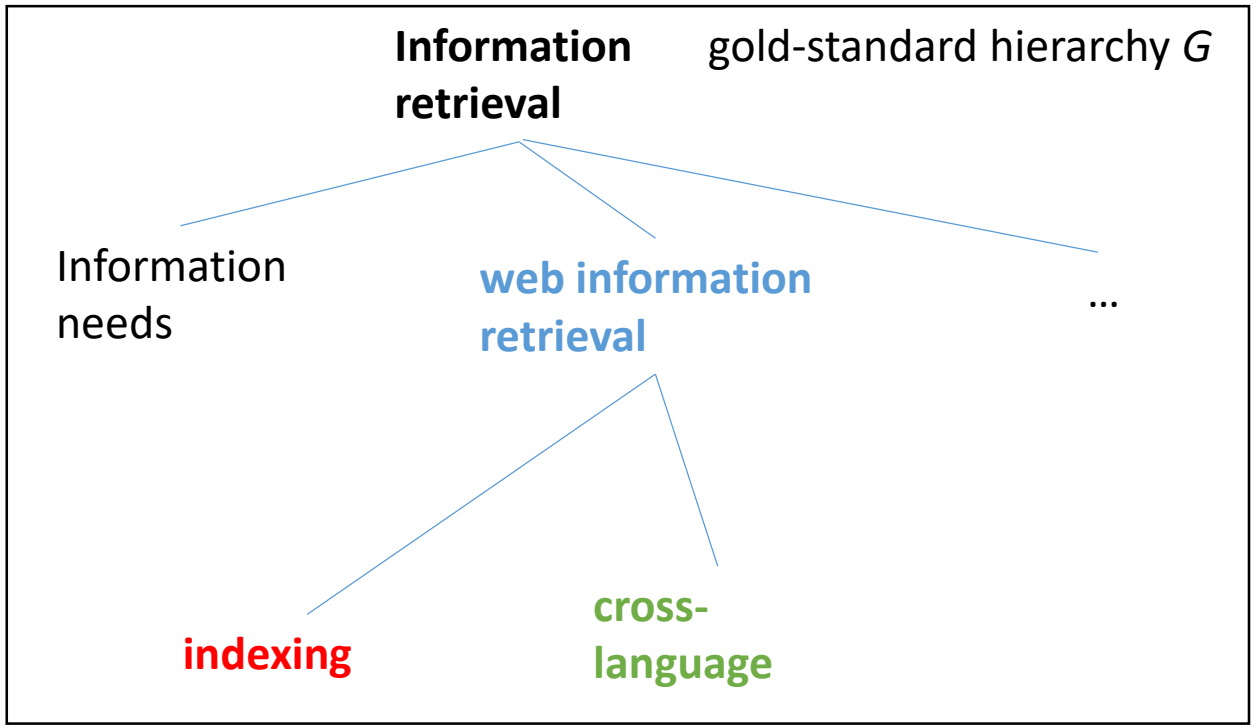
(ii) Extract a ***characteristic excerpt*** for each concept. We use *common direct subsumption* (cdsub) as the characteristic excerpt.

(iii) The similarity of hierarchies is defined based on the characteristic excerpts.



cdsub(Information retrieval, *L, G*) = {indexing, web information retrieval, cross-language}

cdsub(Information retrieval, *G, L*) = {web information retrieval}



- *Taxonomic* Precision (*TP*), *Taxonomic* Recall (*TR*) and *Taxonomic* F-measure (*TF*)
- *Taxonomic* Overlap (*TO*)

- *Taxonomic F'-measure* (*TF'*)

$$tp_{\text{cdsub}}(c, L, G) = \frac{|\text{cdsub}(c, L, G) \cap \text{cdsub}(c, G, L)|}{|\text{cdsub}(c, L, G)|} \quad (1)$$

$$TP(L, G) = \frac{1}{|L \cap G|} \sum_{c \in L \cap G} tp_{\text{cdsub}}(c, L, G) \quad (2)$$

$$to_{\text{cdsub}}(c, L, G) = \frac{|\text{cdsub}(c, L, G) \cap \text{cdsub}(c, G, L)|}{|\text{cdsub}(c, L, G)| \cup \text{cdsub}(c, G, L)|} \quad (3)$$

$$TO(L, G) = \frac{1}{|L \cap G|} \sum_{c \in L \cap G} to_{\text{cdsub}}(c, L, G) \quad (4)$$

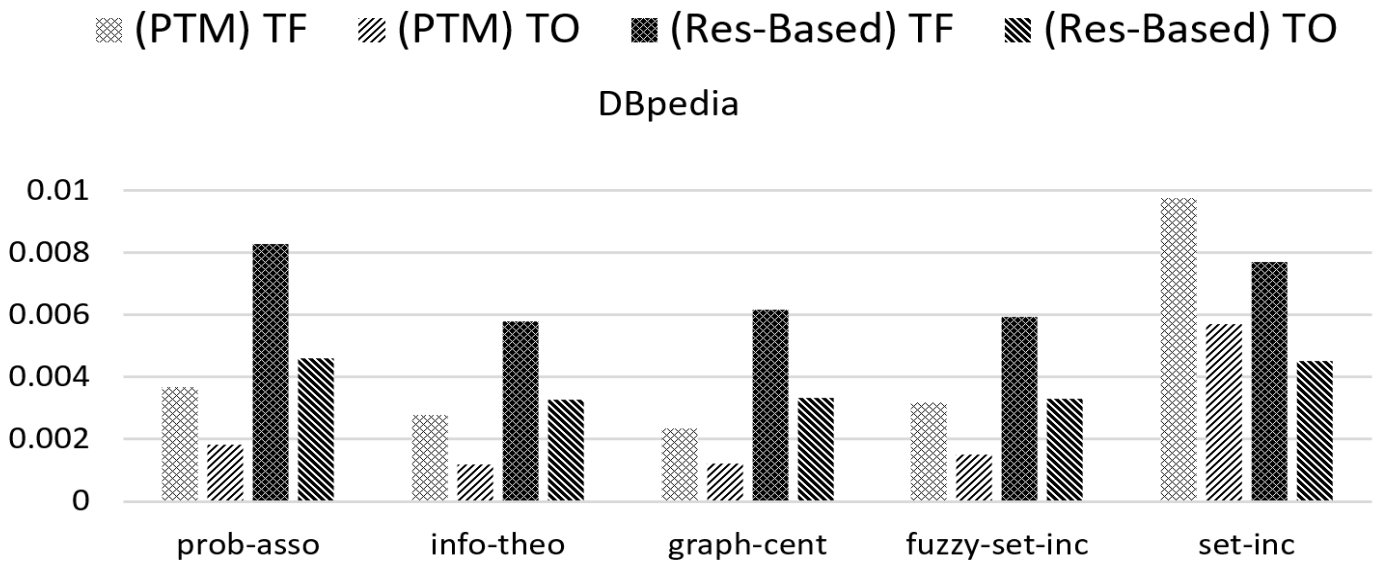
$$TR(L, G) = TP(G, L)$$

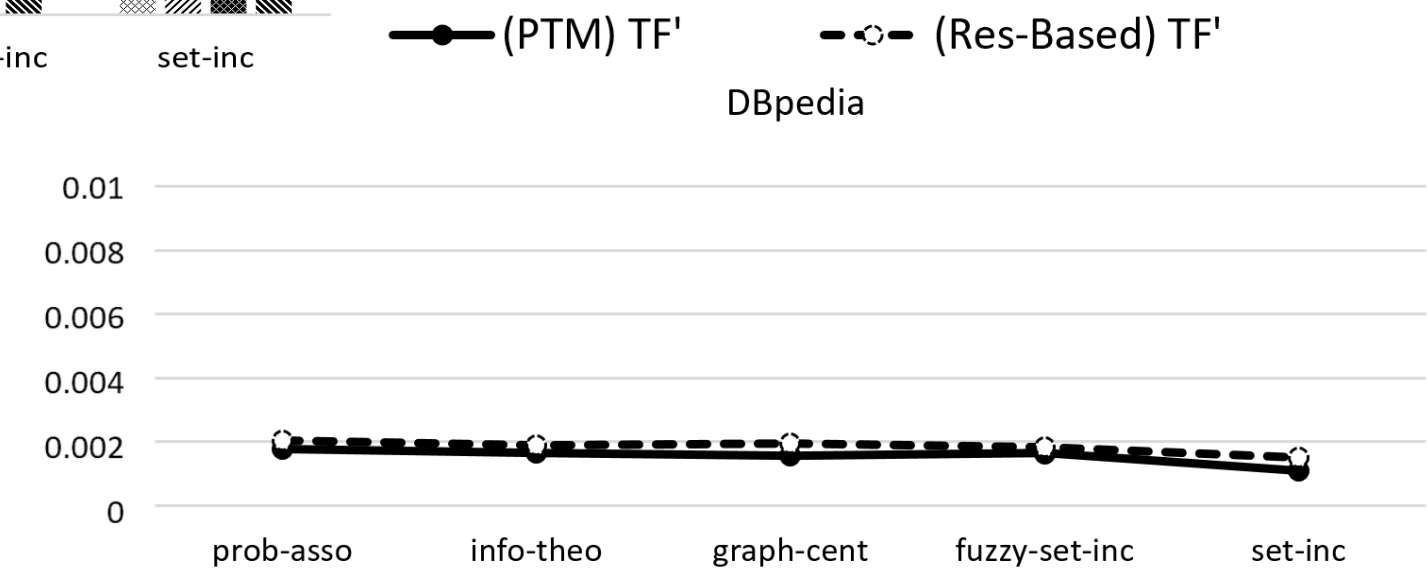
$$TF(L, G) = \frac{2 * TP(L,G) * TR(L,G)}{TP(L,G) + TR(L,G)}$$

$$LR(L, G) = \frac{|V_L \cap V_G|}{|V_G|}$$

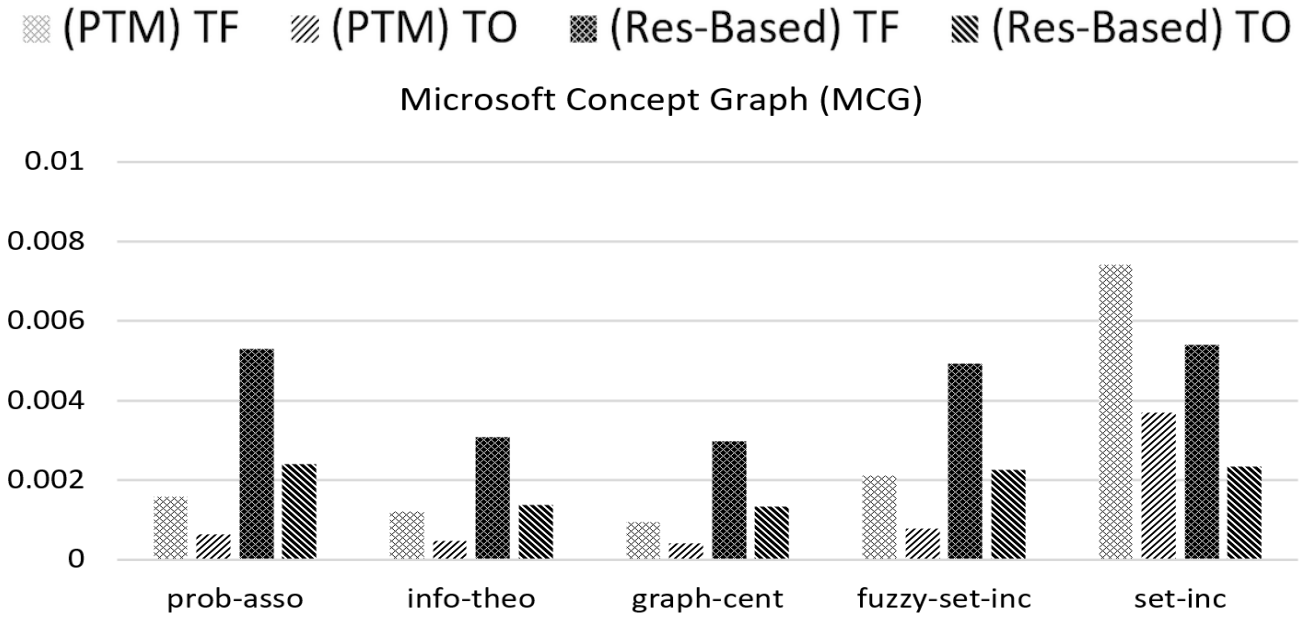
$$TF'(L, G) = \frac{2 * LR(L,G) * TF(L,G)}{LR(L,G) + TF(L,G)}$$

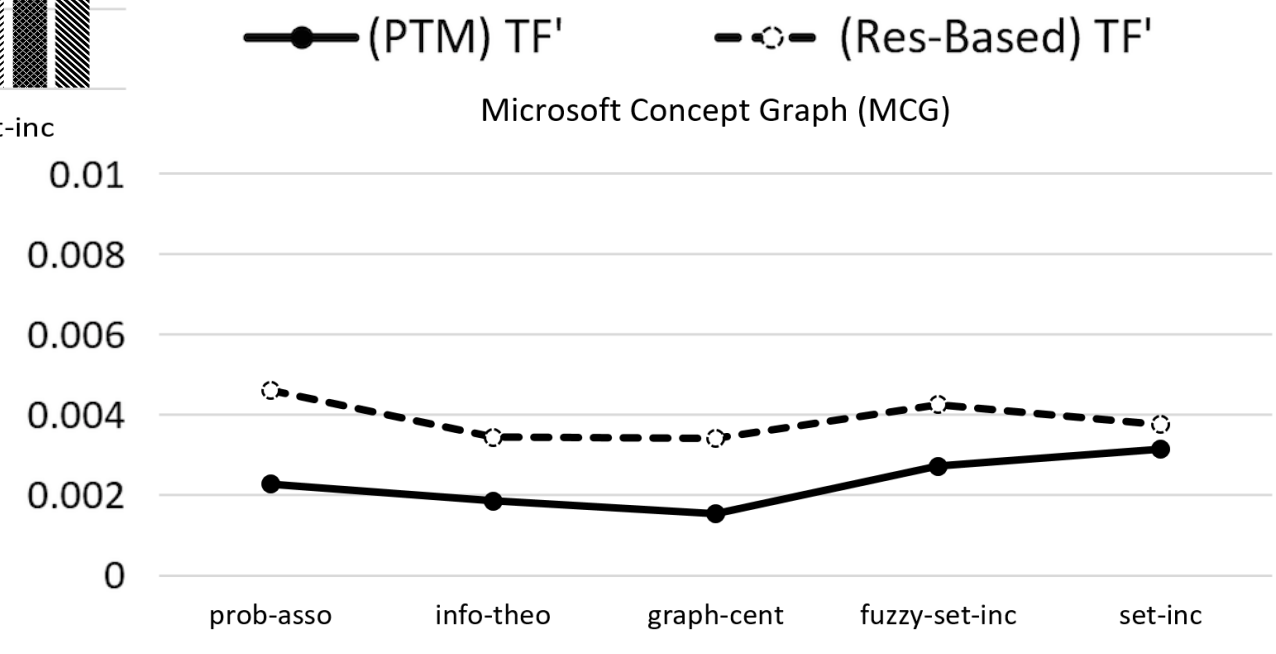
Results - DBpedia



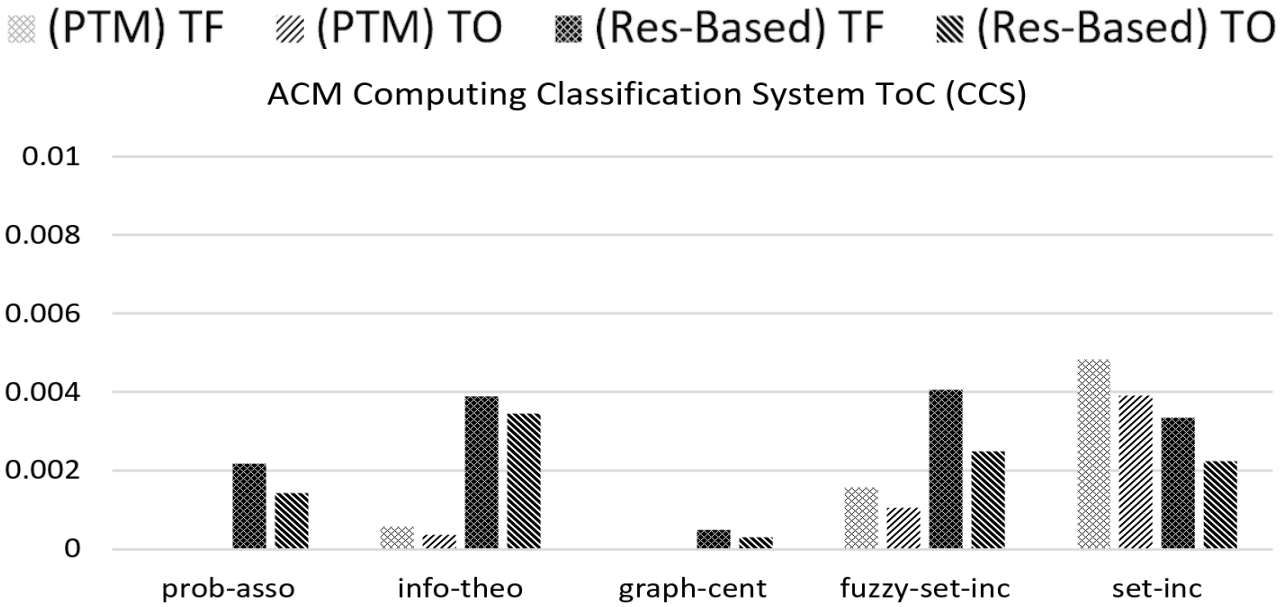


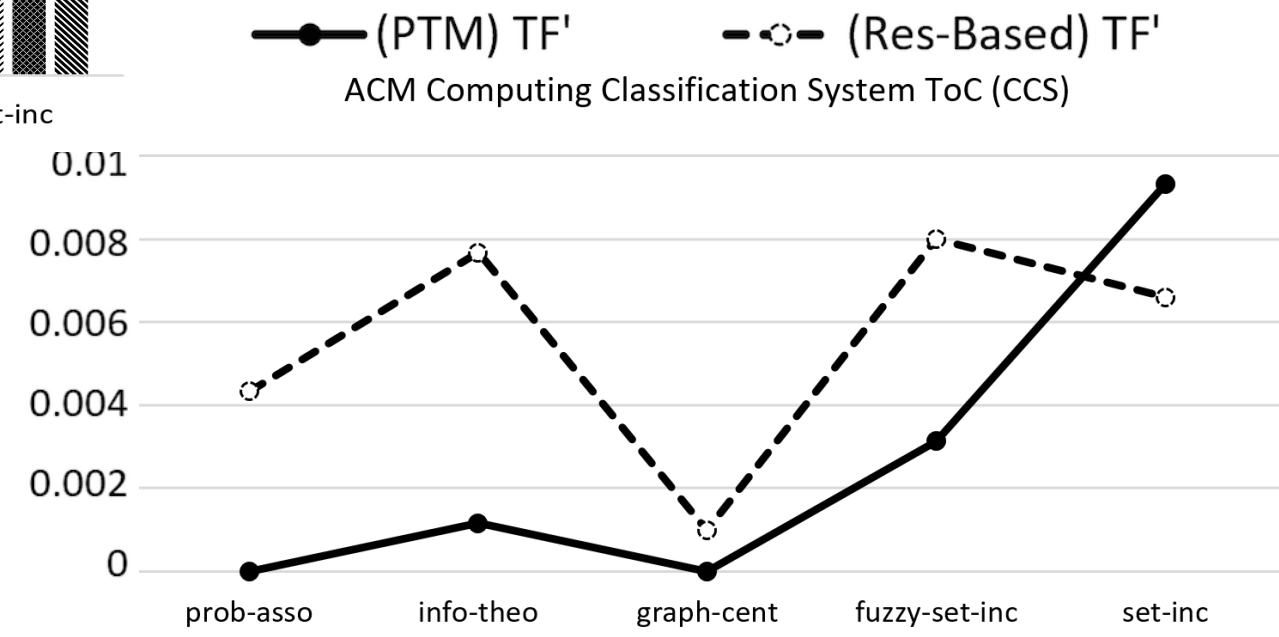
Results – Microsoft Concept Graph





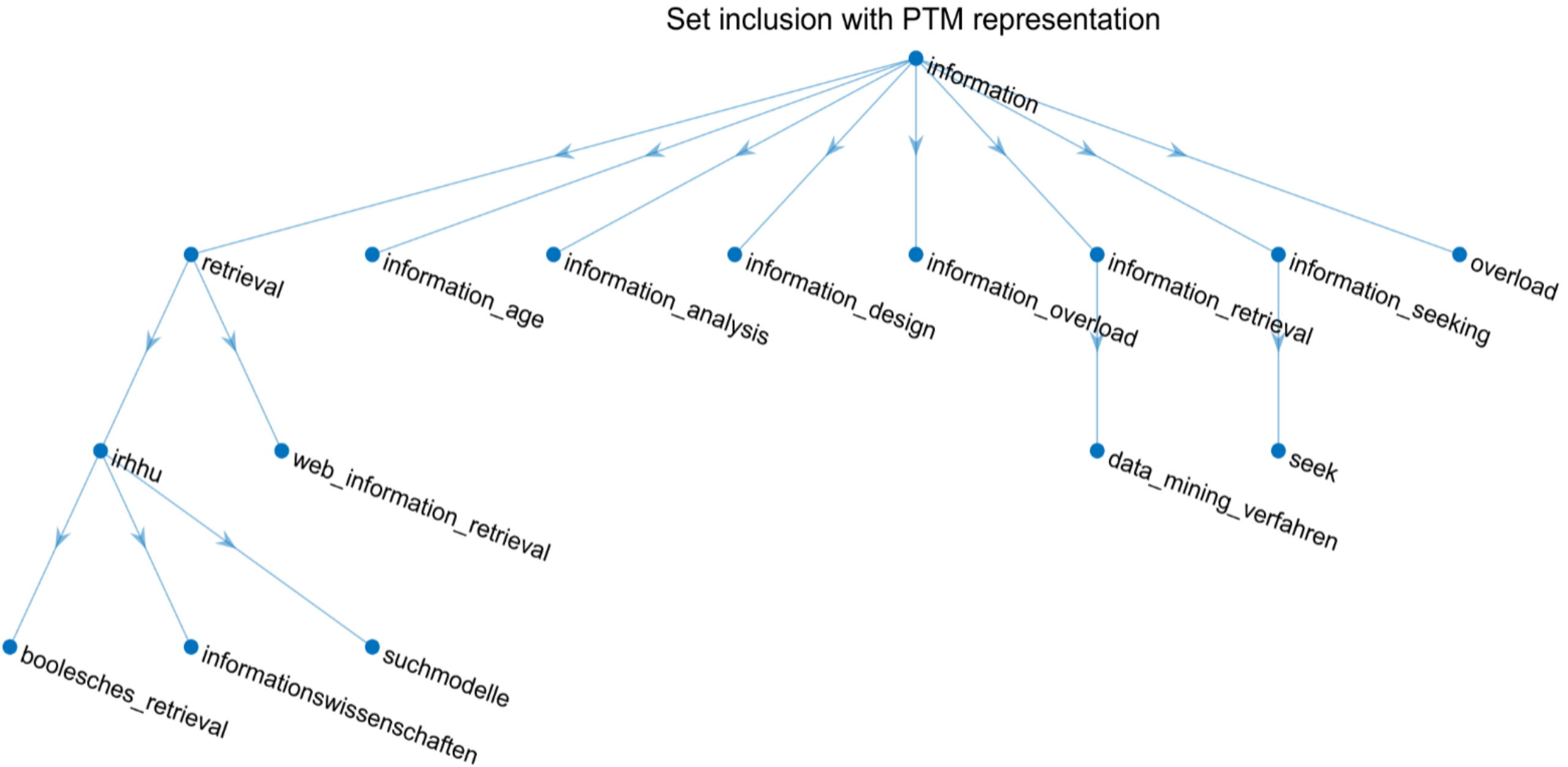
Results – ACM Computing Classification System

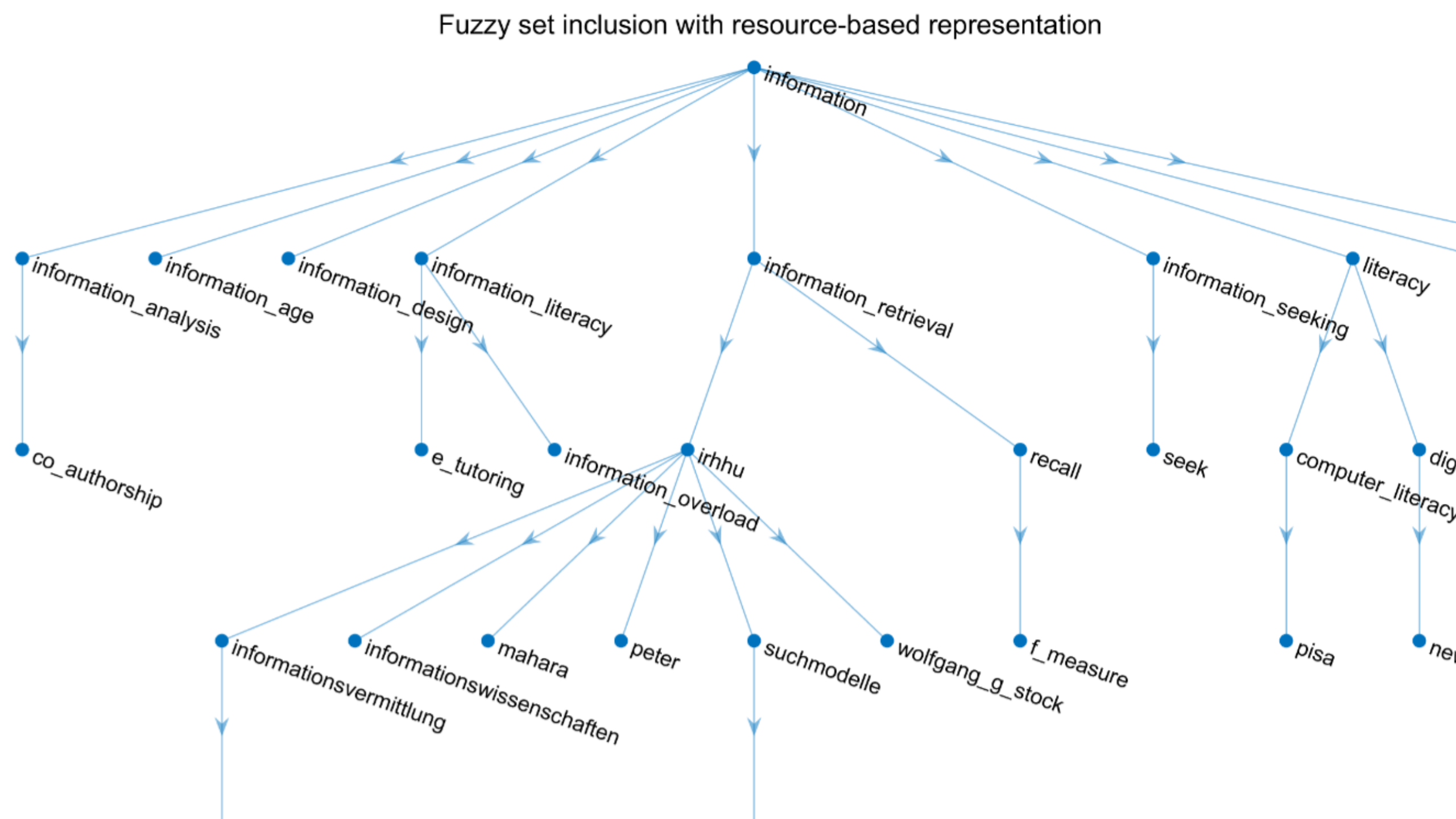




Learned Hierarchies

Probabilistic Association Rule, with resource-based representation.


Set inclusion with PTM representation
information
retrieval
information_age
information_analysis
information_design
information_overload
information_retrieval
information_seeking
overload
irhhu
web_information_retrieval
data_mining_verfahren
seek
boolesches_retrieval
informationswissenschaften
suchmodelle


Fuzzy set inclusion with resource-based representation
information
information_analysis
information_age
information_design
information_literacy
information_retrieval
information_seeking
literacy
co_authorship
e_tutoring
information_overload
irhhu
recall
seek
computer_literacy
dig
informationsvermittlung
informationswissenschaften
mahara
peter
suchmodelle
wolfgang_g_stock
f_measure
pisa
ne

Discussions

- Q1: regarding the rules
  - Set Inclusion Rule results overall best & stable hierarchies in most experimental settings.
  - Fuzzy Set Inclusion and Probabilistic Association rules have competitive results.

- Q2: regarding the data representation techniques
  - The Res-based representation performs best in most experimental settings.
  - Except the PTM representation with Set Inclusion rule had overall best results (TF and TF’).

- Issue:
  - Not consistent among three gold-standard hierarchies, demonstration the distinction of the nature of the chosen gold-standard hierarchies.

Future Studies

- Evaluation: Not just automated evaluation.
- Higher quality hierarchies through machine learning:
  - Use the rules altogether to induce hierarchies: features in supervised learning
  - Add further information/context: resource contents, external lexical resources, transfer learning, etc.
- Use deep learning approaches:
  - Forget about the rules?
  - Using very rich data representations: word embedding

References

- Benz, D., Hotho, A., Stumme, G., Stutzer, S.: Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In: Proceedings of the 2nd Web Science Conference (WebSci10) (2010)
- Cruse, D.A.: Hyponymy and its varieties. In: Green, R., Bean, C.A., Myaeng, S.H. (eds.) The Semantics of Relationships: An Interdisciplinary Perspective, pp. 3–21. Springer, Dordrecht (2002). https://doi.org/10.1007/978-94-017-0073-3 1
- Dellschaft, K., Staab, S.: On how to perform a gold standard based evaluation of ontology learning. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 228–241. Springer, Heidelberg (2006). https://doi.org/10.1007/11926078 17
- Dong, H., Wang, W., Frans, C.: Deriving dynamic knowledge from academic social tagging data: a novel research direction. In: iConference 2017 Proceedings (2017)
- Griffiths, T.L., Steyvers, M.: Prediction and semantic association. In: Proceedings of the 15th International Conference on Neural Information Processing Systems, pp. 11–18. MIT Press (2002)
- Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford University (2006)
- Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., & Stumme, G. Evaluating similarity measures for emergent semantics of social tagging. In Proceedings of the 18th international conference on World wide web, 641-650. ACM (2009, April).
- Meo, P. D., Quattrone, G., Ursino, D.: Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies. Inf. Syst. 34(6), 511–535 (2009)
- Mika, P.: Ontologies are us: a unified model of social networks and semantics. Web Semant.: Sci. Serv. Agents World Wide Web 5(1), 5–15 (2007)
- Peters, I., Becker, P.: Folksonomies: Indexing and Retrieval in Web 2.0. De Gruyter/Saur, Berlin (2009)
- Strohmaier, M., Helic, D., Benz, D., Korner, C., Kern, R.: Evaluation of folksonomy induction algorithms. ACM Trans. Intell. Syst. Technol. 3(4), 1–22 (2012)
- Vander Wal: Folksonomy Coinage and Definition. http://www.vanderwal.net/folksonomy.html (2007)
- Wang, W., Barnaghi, P.M., Bargiela, A.: Probabilistic topic models for learning terminological ontologies. IEEE Trans. Knowl. Data Eng. 22(7), 1028–1040 (2010)
- Weller, K.: Knowledge Representation in the Social Semantic Web. De Gruyter Saur, Berlin/New York (2010).

Thank you for your attention.

Hang Dong's Home page: http://cgi.csc.liv.ac.uk/~hang/
Contact: hangdong@liverpool.ac.uk