TaxoLearn: a Semantic Approach to Domain Taxonomy Learning

Emmanuelle Dietz, Damir Vandic, Flavius Frasincar

ERASMUS UNIVERSITEIT ROTTERDAM

Introduction

- Taxonomies important in information science
- Manually construction is time consuming
  - requires expert knowledge
- Solution = taxonomy learning
  - automatically construct taxonomy given a corpus of data

Introduction

Aspects in taxonomy learning

- data sparseness
- syntactical structure vs semantics
- relevance of concepts
- relations between concepts

TaxoLearn

TaxoLearn

- Requires:
  - corpus of documents of interest
  - corpora of documents unrelated to the domain of interest
- Outputs:
  - taxonomy of *concepts*, deduced from the provided documents of interest

TaxoLearn

1. Find (disambiguated) candidate concepts
2. Select relevant concepts
3. Determine concept similarities
4. Construct and label taxonomy

TaxoLearn
1. Find candidate concepts

The stock market was heavily shaken after the European Bank lowered the interest rates.

TaxoLearn

1. Find candidate concepts

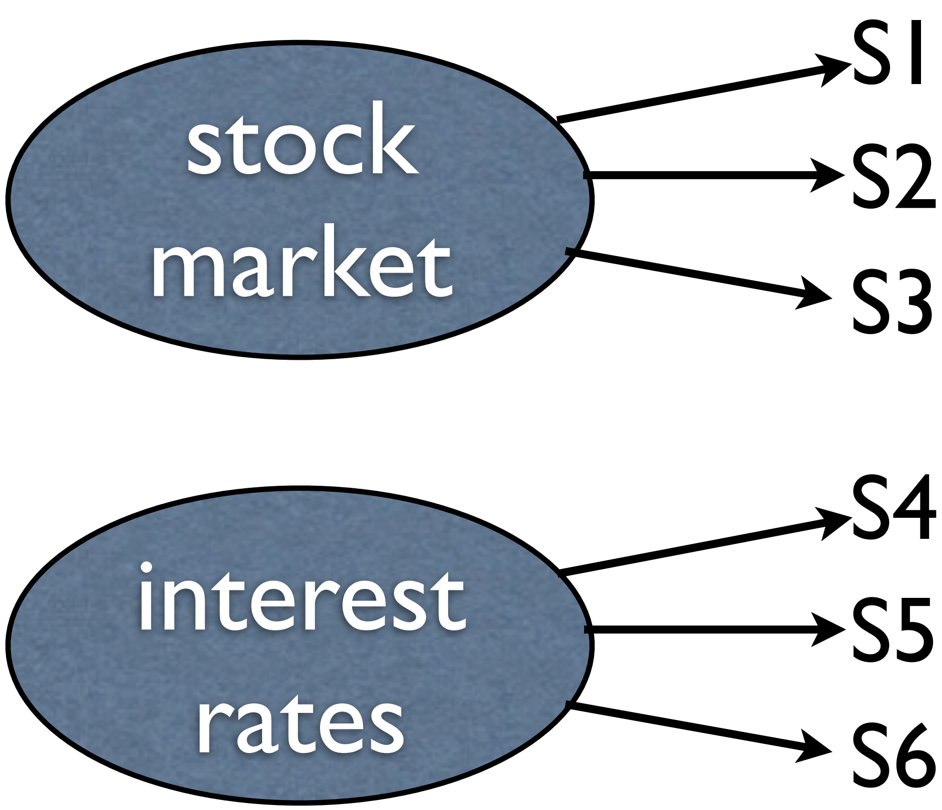
The stock market was heavily shaken after the European Bank lowered the interest rates.

TaxoLearn

1. Find candidate concepts

The stock market was heavily shaken after the European Bank lowered the interest rates.

link to WordNet synsets

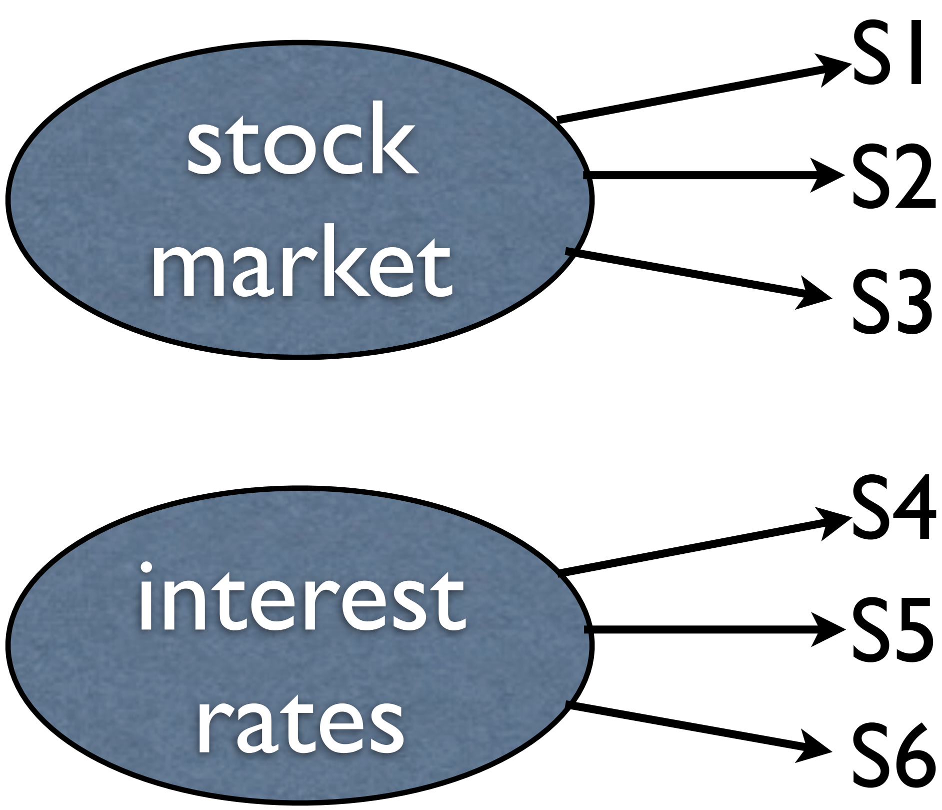


TaxoLearn

1. Find candidate concepts

The stock market was heavily shaken after the European Bank lowered the interest rates.

link to WordNet synsets



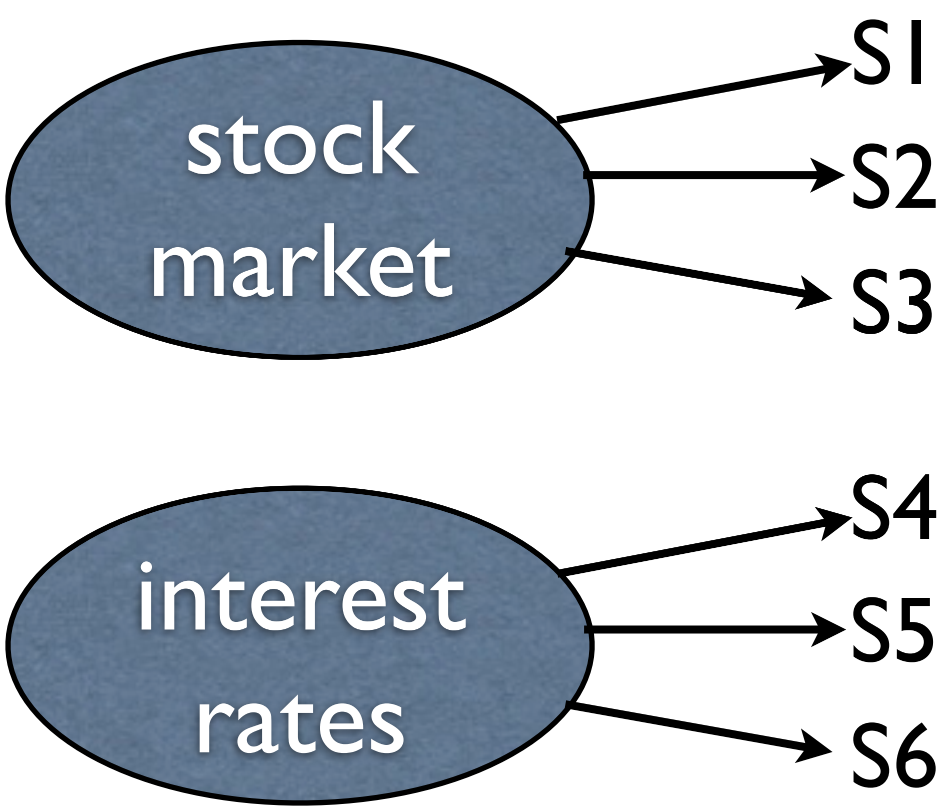
$$\mathrm{KPP}(v, V) = \frac{\sum_{u \in V : u \neq v} \frac{1}{d(u,v)}}{|V| - 1}$$

TaxoLearn

1. Find candidate concepts

The stock market was heavily shaken after the European Bank lowered the interest rates.

link to WordNet synsets



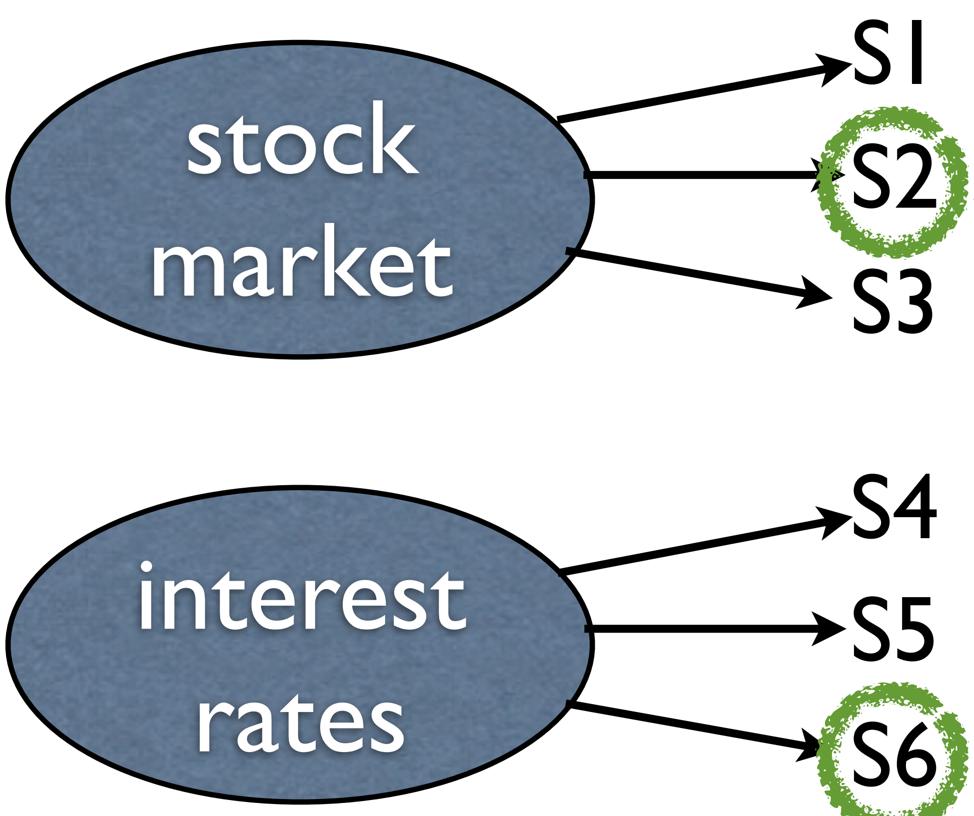
$$\mathrm{KPP}(v, V) = \frac{\sum_{u \in V : u \neq v} \frac{1}{d(u,v)}}{|V| - 1}$$

{S1, S2, S3, S4, S5, S6}

TaxoLearn
1. Find candidate concepts

The stock market was heavily shaken after the European Bank lowered the interest rates.

link to WordNet synsets



$$\mathrm{KPP}(v, V) = \frac{\sum_{u \in V : u \neq v} \frac{1}{d(u,v)}}{|V| - 1}$$

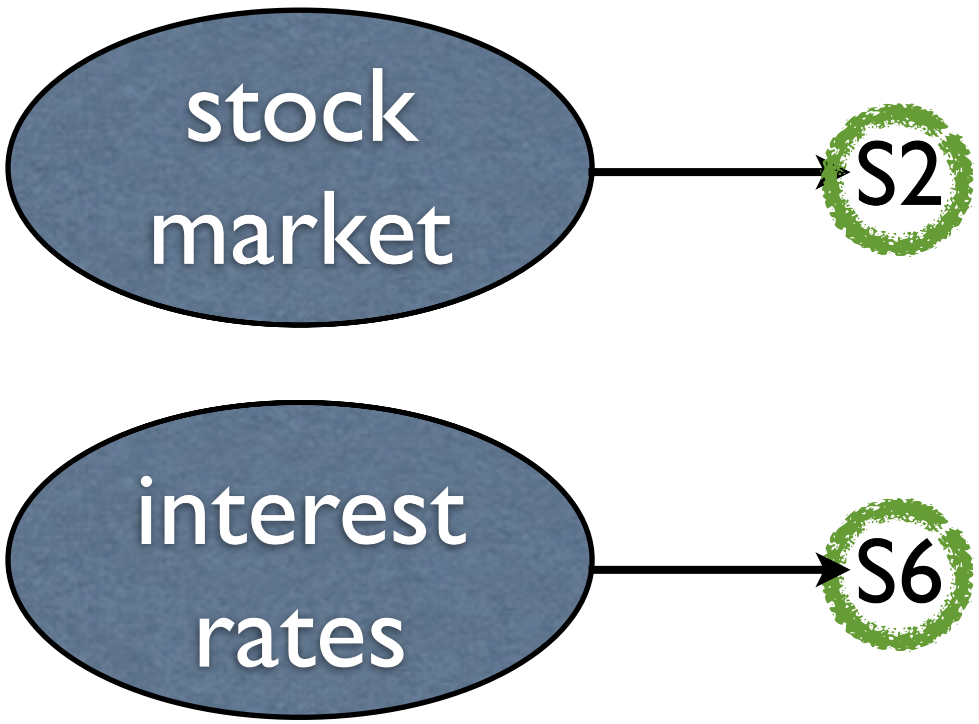
{S1, S2, S3, S4, S5, S6}

TaxoLearn

2. Select relevant concepts

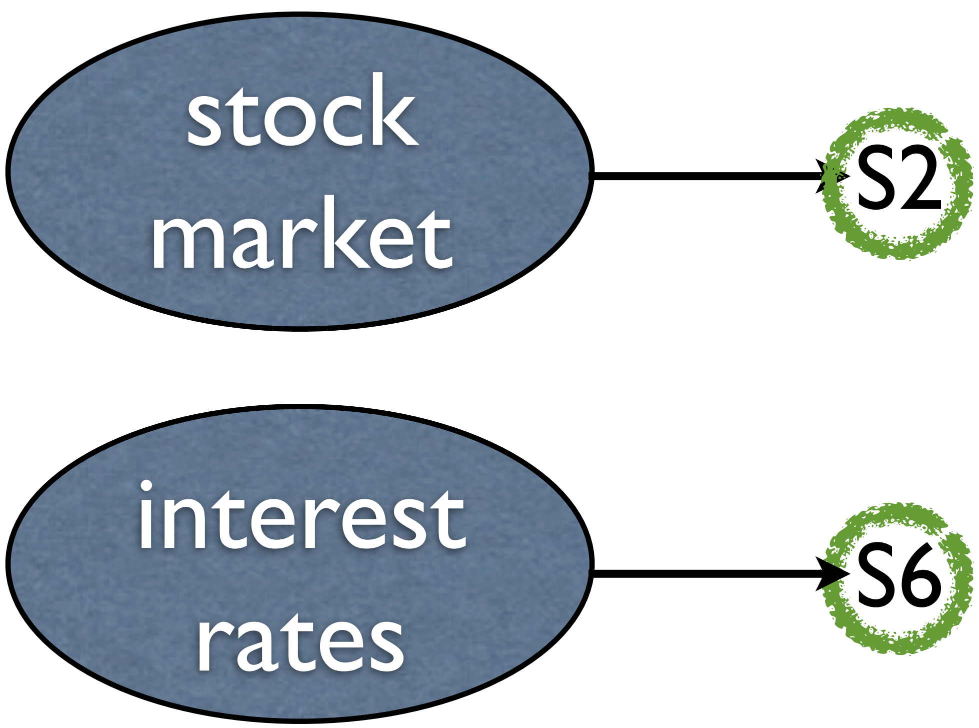
TaxoLearn

2. *Select relevant concepts*



TaxoLearn

2. Select relevant concepts



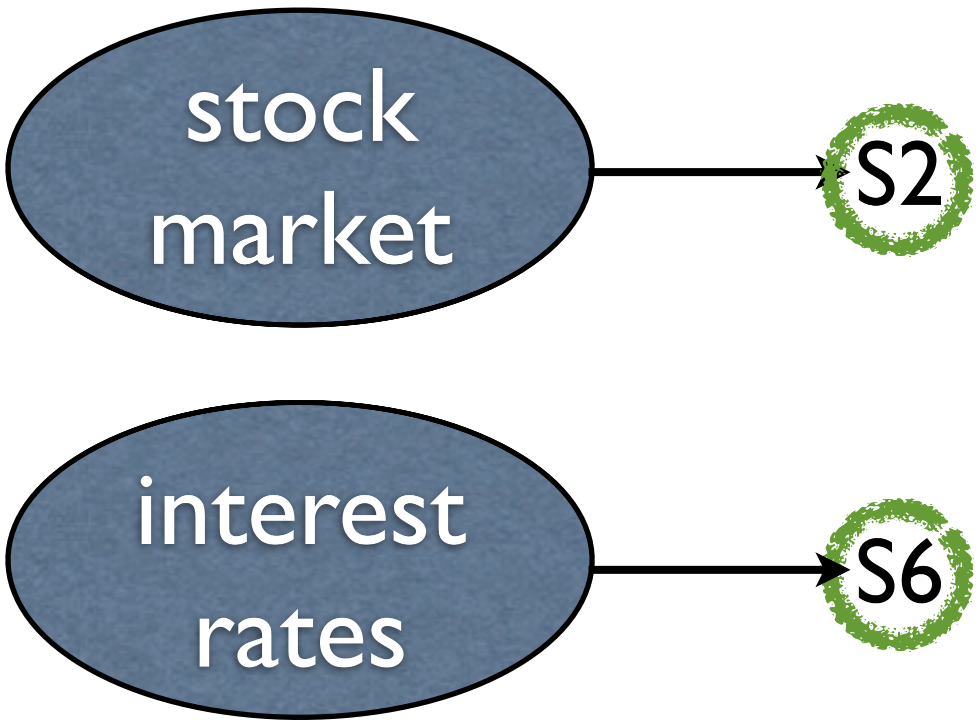
Use two filters:

- Domain Pertinence

$$\mathrm{DP}(c, D^*) = \frac{\mathrm{freq}(c, D^*)}{\max_{j, D_j \neq D^*}\left(\mathrm{freq}(c, D_j)\right)}$$

TaxoLearn

2. *Select relevant concepts*



Use two filters:

- Domain Pertinence

$$\mathrm{DP}(c, D^*) = \frac{\mathrm{freq}(c, D^*)}{\max_{j, D_j \neq D^*} \left(\mathrm{freq}(c, D_j)\right)}$$

- Domain Consensus

$$\mathrm{DC}(c, D^*) = -\sum_{d_k \in D^*} \mathrm{norm\_freq}(c, d_k) \times \log\left(\mathrm{norm\_freq}(c, d_k)\right)$$

with

$$\mathrm{norm\_freq}(c, d_k) = \frac{\mathrm{freq}(c, d_k)}{\max(\mathrm{freq}(c, D))}$$

TaxoLearn

3. Determine concept similarities

Three methods for computing similarity:

- The WordNet method
- The PMI method
  (Pointwise Mutual Information)
- The Web method

TaxoLearn

3. Determine concept similarities

The WordNet method

$$\mathrm{sim}_{\mathrm{WN}}(c_i, c_j) = \frac{1}{d(c_i, c_j)}$$

TaxoLearn

3. Determine concept similarities

The PMI method

$$\mathrm{sim}_{\mathrm{PMI}}(c_i, c_j) = \log \frac{F_{c_i \cap c_j} / F_{all}}{\left(F_{c_i} / F_{all}\right) \times \left(F_{c_j} / F_{all}\right)}$$

TaxoLearn

3. Determine concept similarities

The Web method

$$\mathrm{sim}_{\mathrm{WEB}}(c_i, c_j) = \log \frac{H_{c_i \cap c_j} / H_{all}}{\left(H_{c_i} / H_{all}\right) \times \left(H_{c_j} / H_{all}\right)}$$

TaxoLearn

4. Construct and label taxonomy

Constructing the taxonomy

- Hierarchical clustering is used for the WordNet, PMI, and Web method
- Advantages:
  - Able to inspect dendogram
  - Average linkage is used

TaxoLearn

4. Construct and label taxonomy

TaxoLearn

4. Construct and label taxonomy

Labeling the taxonomy

TaxoLearn

4. Construct and label taxonomy

Labeling the taxonomy

- Two approaches from literature:

TaxoLearn

4. Construct and label taxonomy

Labeling the taxonomy

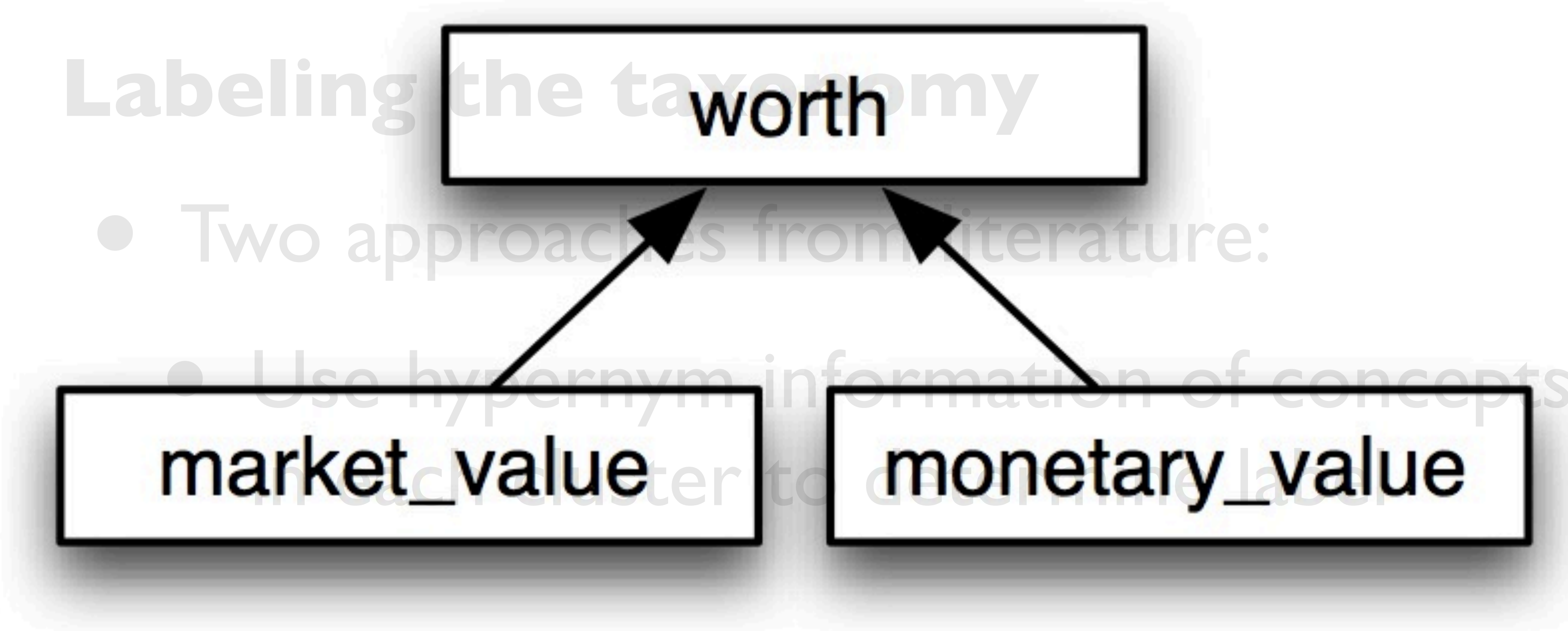
- Two approaches from literature:
  - Use hypernym information of concepts in each cluster to determine label

TaxoLearn

4. Construct and label taxonomy

Labeling the taxonomy

- Two approaches from literature:
  - Use hypernym information of concepts in each cluster to determine label



TaxoLearn

4. Construct and label taxonomy

Labeling the taxonomy

- Two approaches from literature:
  - Use hypernym information of concepts in each cluster to determine label

TaxoLearn

4. Construct and label taxonomy

Labeling the taxonomy

- Two approaches from literature:
  - Use hypernym information of concepts in each cluster to determine label
  - Use centroid of each cluster as label

TaxoLearn

4. Construct and label taxonomy

Labeling the taxonomy

- Two approaches from literature:
  - Use hypernym information of concepts in each cluster to determine label
  - Use centroid of each cluster as label
- We employ a hybrid approach

TaxoLearn

4. Construct and label taxonomy

Labeling the taxonomy (hybrid)

- Our hybrid approach:
  - first checks whether there is a concept that is a hypernym of *x* other concepts
  - for clusters of size 2, we use the common hypernym
  - otherwise: we use a modified version of the centroid approach

Evaluation

Several measures used to obtain the precision, recall, and F1-measure

- Lexical recall
- Taxonomy Overlap

Evaluation

Several measures used to obtain the precision, recall, and F1-measure

- Lexical recall

$$LR(O_1, O_2) := \frac{|O_1 \cap O_2|}{|O_2|}$$

- Taxonomy Overlap

Evaluation

Taxonomy Overlap

$$\overline{TO}(O_1, O_2) := \frac{1}{|O_1|} \times \sum_{c \in O_1} TO(c, O_1, O_2)$$

$$TO(c, O_1, O_2) := \begin{cases} TO^{'}(c, O_1, O_2), & c \in O_2 \\ TO^{''}(c, O_1, O_2), & c \notin O_2 \end{cases}$$

$$TO^{'}(c, O_1, O_2) := \frac{|SC(c, O_1) \cap SC(c, O_2)|}{|SC(c, O_1) \cup SC(c, O_2)|}$$

$$TO^{''}(c, O_1, O_2) := max_{c' \in O_2} \frac{|SC(c, O_1) \cap SC(c', O_2)|}{|SC(c, O_1) \cup SC(c', O_2)|}$$

Evaluation

$$Precision : P(O_1, O_2) := \overline{TO}(O_1, O_2)$$

$$Recall : R(O_1, O_2) := \overline{TO}(O_2, O_1)$$

$$F{-}Measure :$$

$$F(O_1, O_2) := \frac{2 \times P(O_1, O_2) \times R(O_1, O_2)}{P(O_1, O_2) + R(O_1, O_2)}$$

$$F'(O_1, O_2) := \frac{2 \times LR(O_1, O_2) \times F(O_1, O_2)}{LR(O_1, O_2) + F(O_1, O_2)}$$

Evaluation

- Data set from Erasmus RePub repository
  - consists of 236 papers in the domain of Financial Economics
  - abstracts of papers in medicine & health, law, and culture & society were also used
- Manually constructed golden taxonomy
  - using WordNet synsets
  - only utilizing knowledge from the data set

Evaluation

Measure	**WordNet**	**Web**	**PMI**
Lexical recall	0.42	0.43	0.44
Precision	0.50	0.99	0.69
Recall	0.27	0.19	0.21
F-measure	0.35	0.32	0.32
F’-measure	0.38	0.37	0.37

Questions?