CHECKSEM
SEMANTICINTELLIGENCERESEARCH

An unsupervised classification process for large datasets based on web reasoning

Rafael PEIXOTO, Thomas HASSAN, Christophe CRUZ, Aurelie BERTAUX, Nuno SILVA
thomas.hassan@u-bourgogne.fr

Laboratoire LE2I – UMR CNRS 6306 – Université de Bourgogne

Outline

Context

- Global problem
- The Semantic HMC

Specific Problem

- Proposed Solution

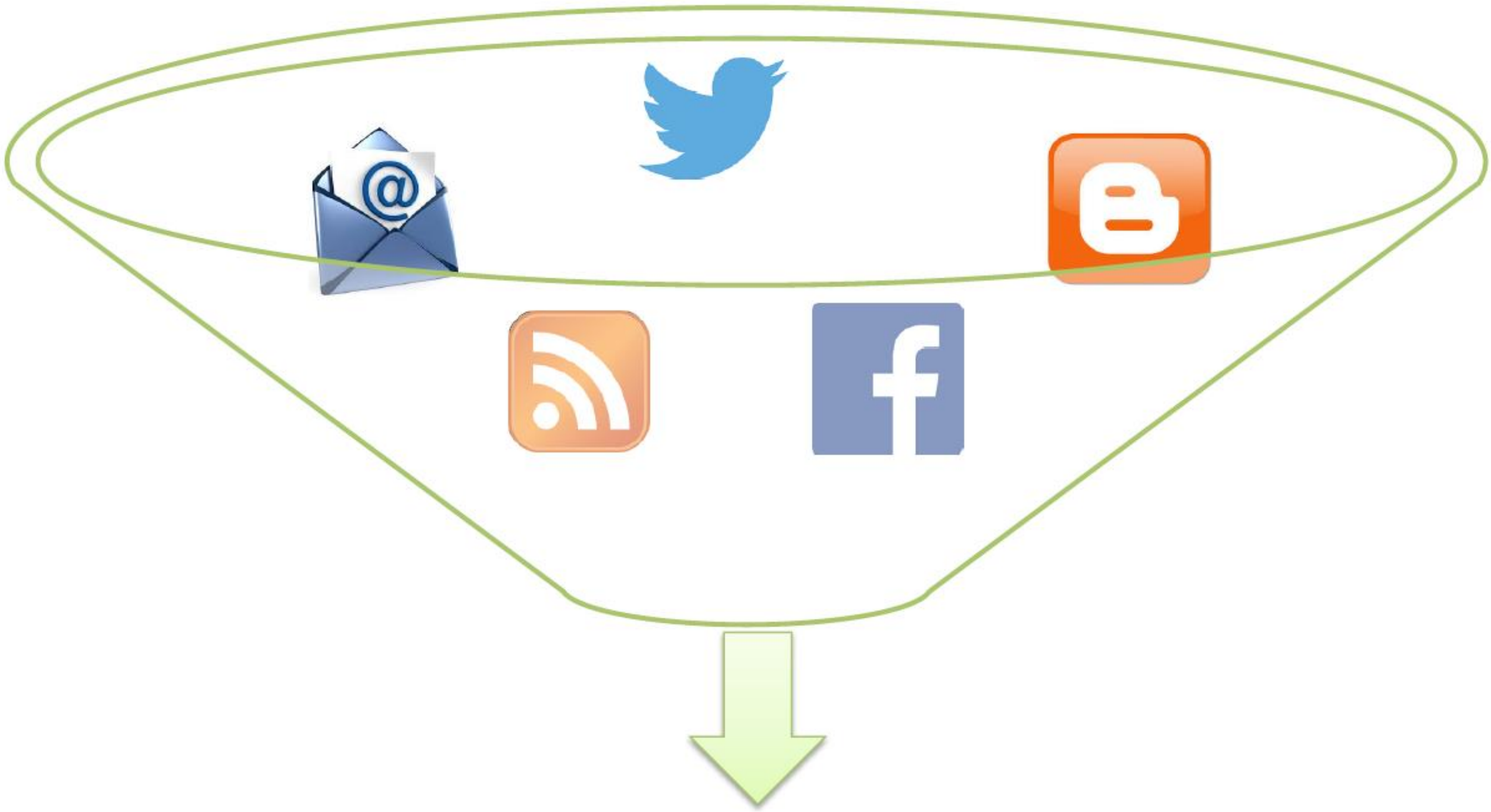
Implementation

- Setup
- Results

Conclusion and future work

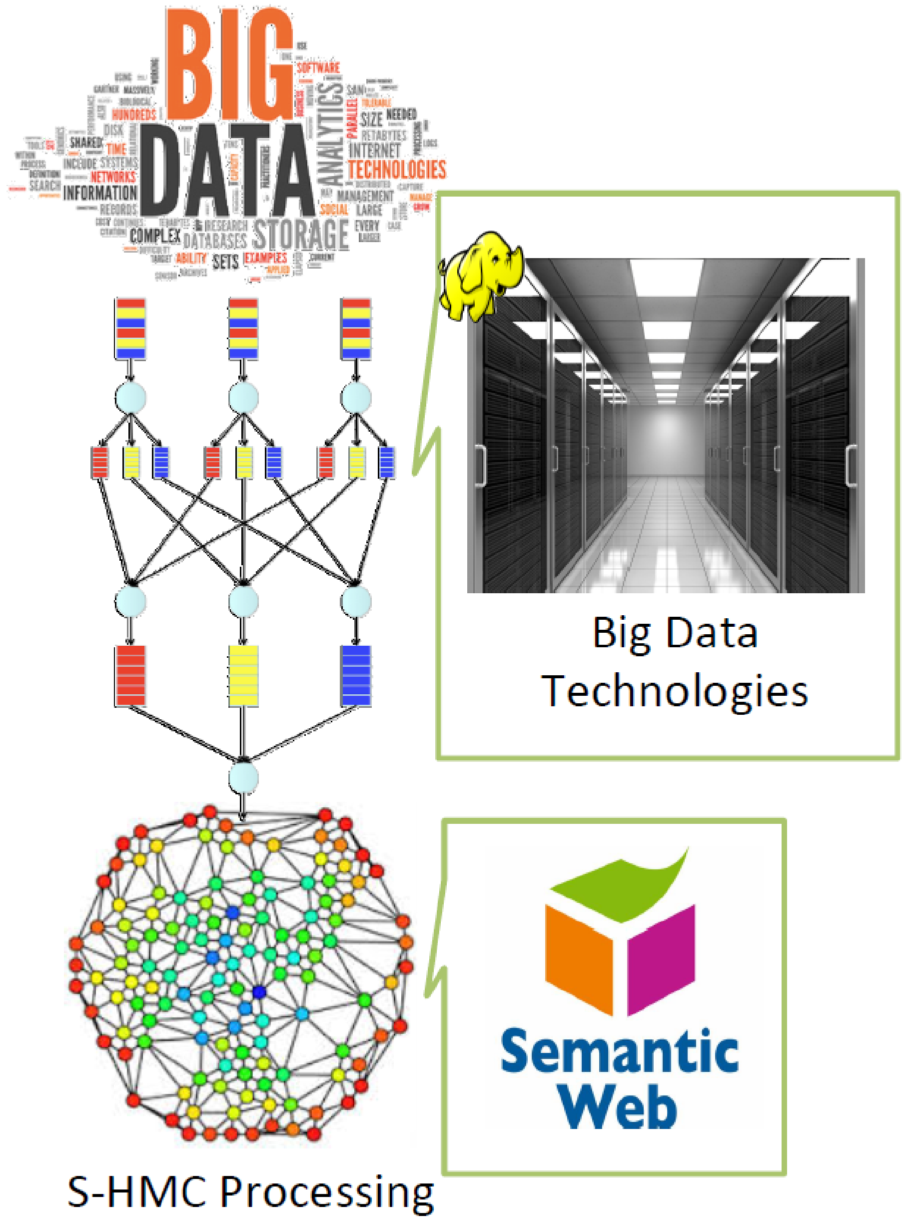
Global Problem

Value extraction from **Big Data** sources

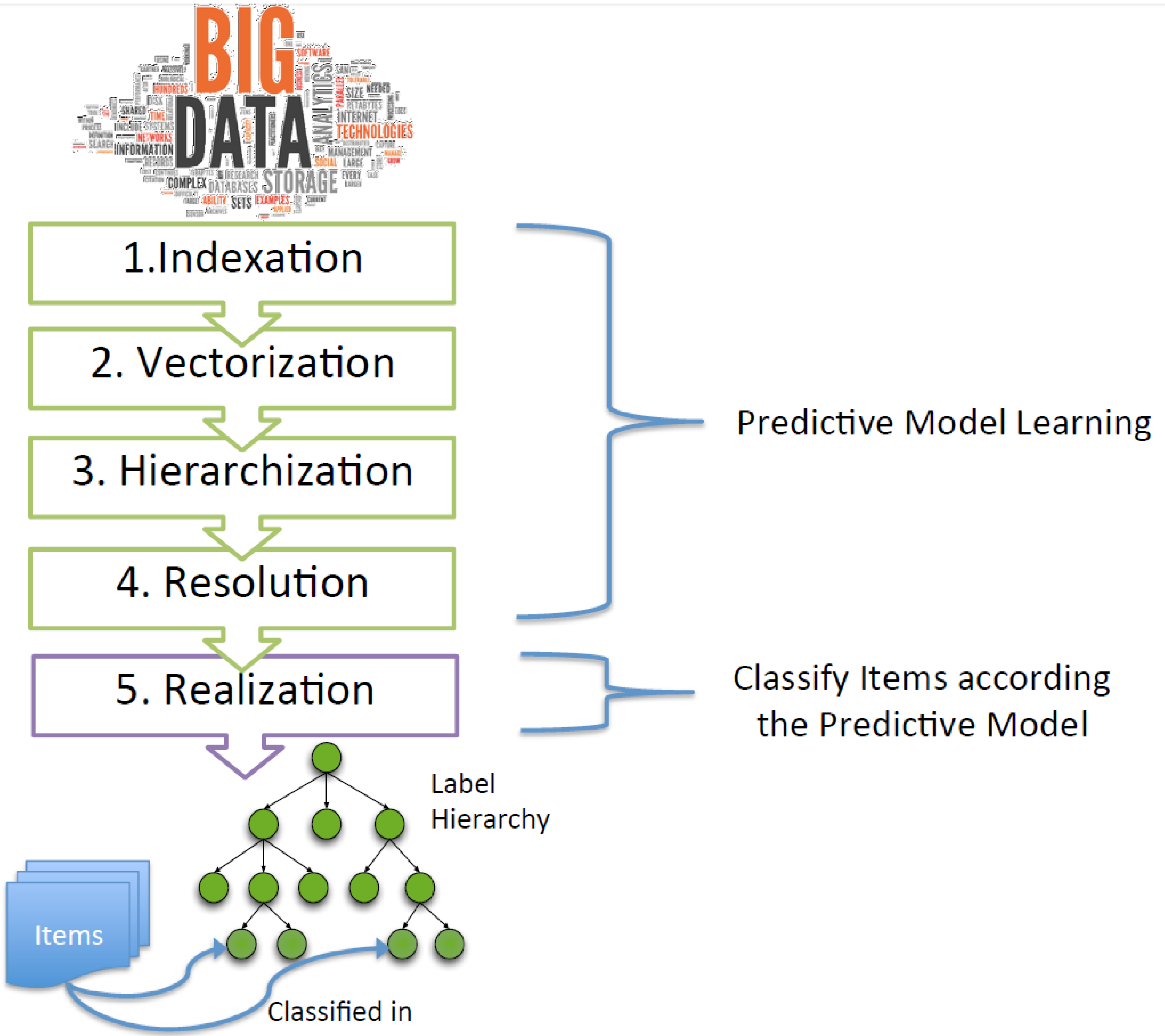


Global Problem

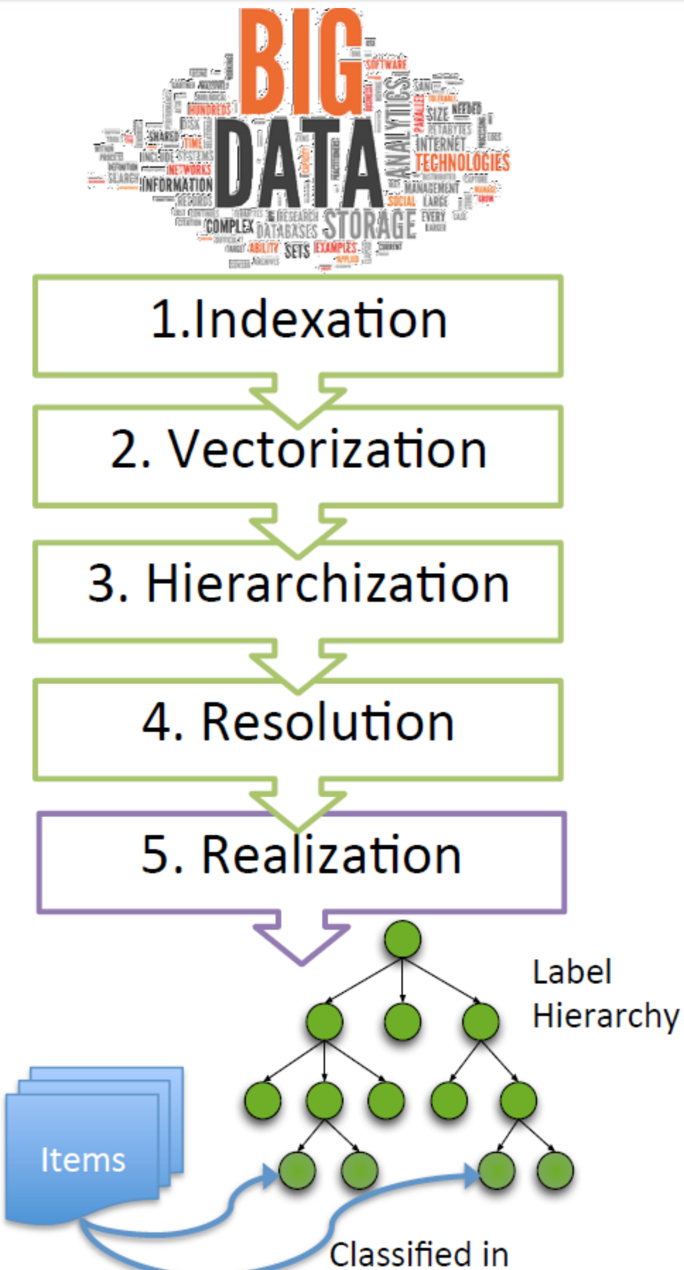
- Why ontologies
  - Ontologies are the most accepted way to represent semantics in the Semantic Web and a good solution for intelligent computer systems that operate close to the human concept level, bridging the gap between human conceptions and computational requirements.
- Semantic HMC
  - Ontology-described predictive model
  - Learned using Big Data technologies
  - Rule-based Web Reasoning to perform classification



Proposition: « Semantic HMC »

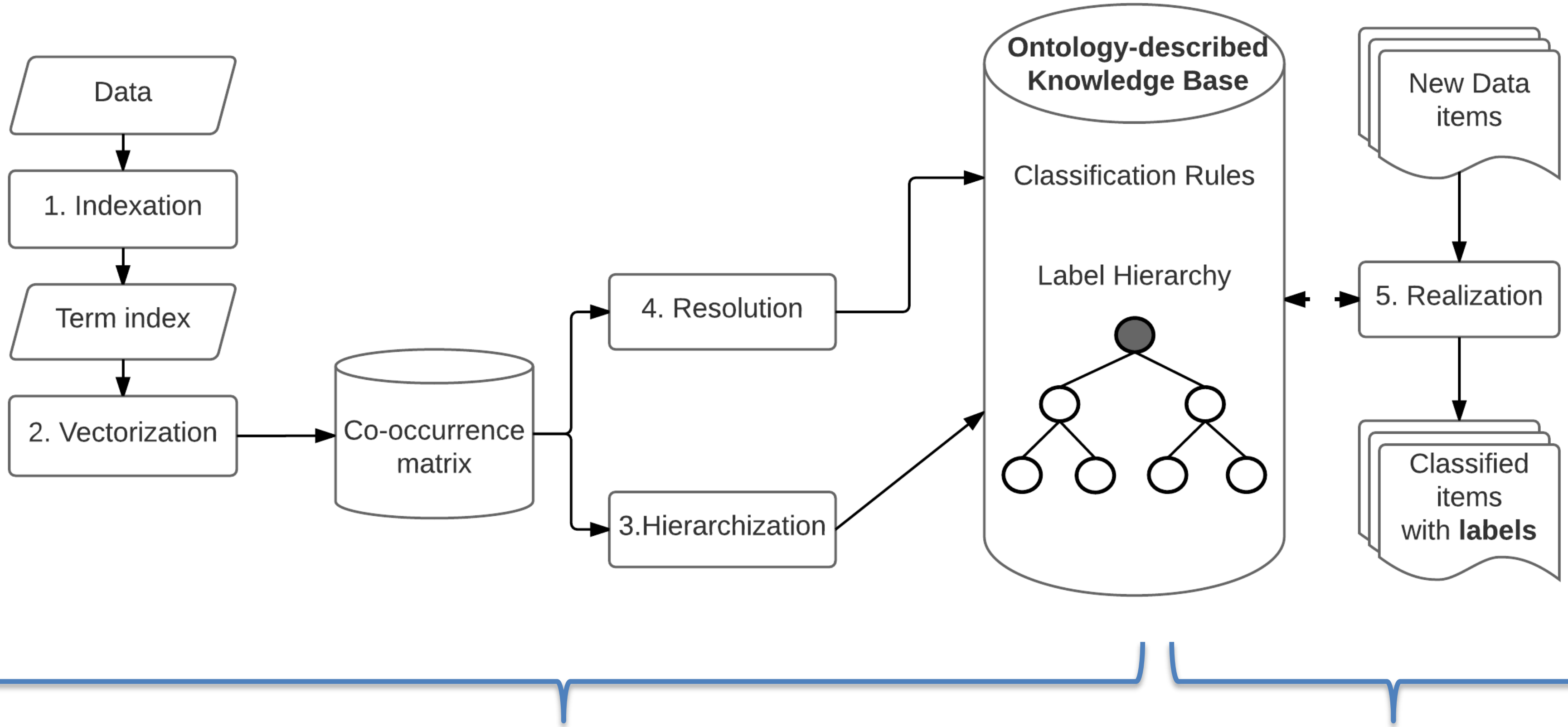


Proposition: « Semantic HMC »



- Indexation
  - Extract terms
  - Index the items
- Vectorization
  - Calculate term frequency vectors
  - Co-occurrence matrix
- Hierarchization
  - Label selection
  - Hierarchical relations
- Resolution
  - Classification rules creation
- Realization
  - Ontology population
  - Rule-based Web Reasoning to classify items

Proposition : « Semantic HMC »



Unsupervised **ontology learning**

Rule-based Classification (Web Reasoner)

Outline

Context

- Global problem
- The Semantic HMC

Specific Problem

- Proposed Solution

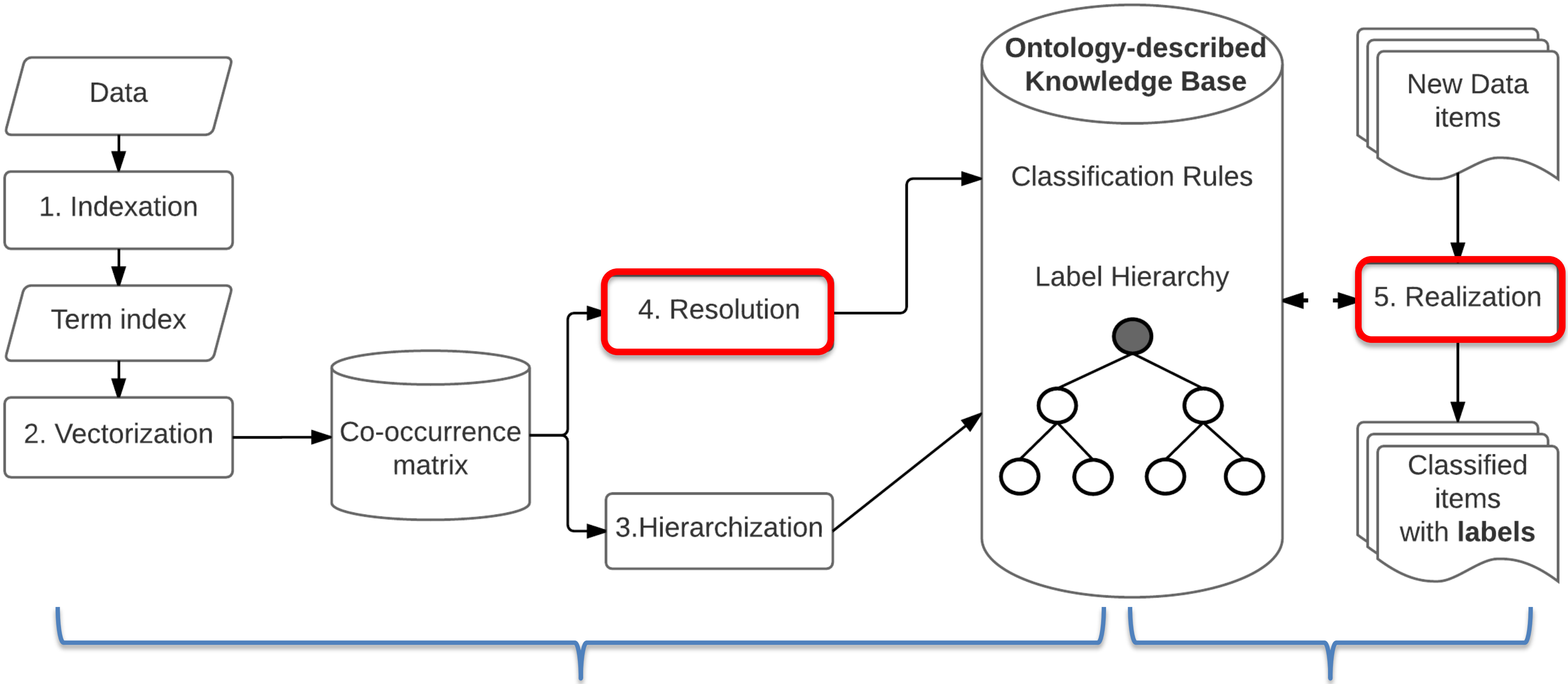
Implementation

- Setup
- Results

Conclusion and future work

Specific Problem

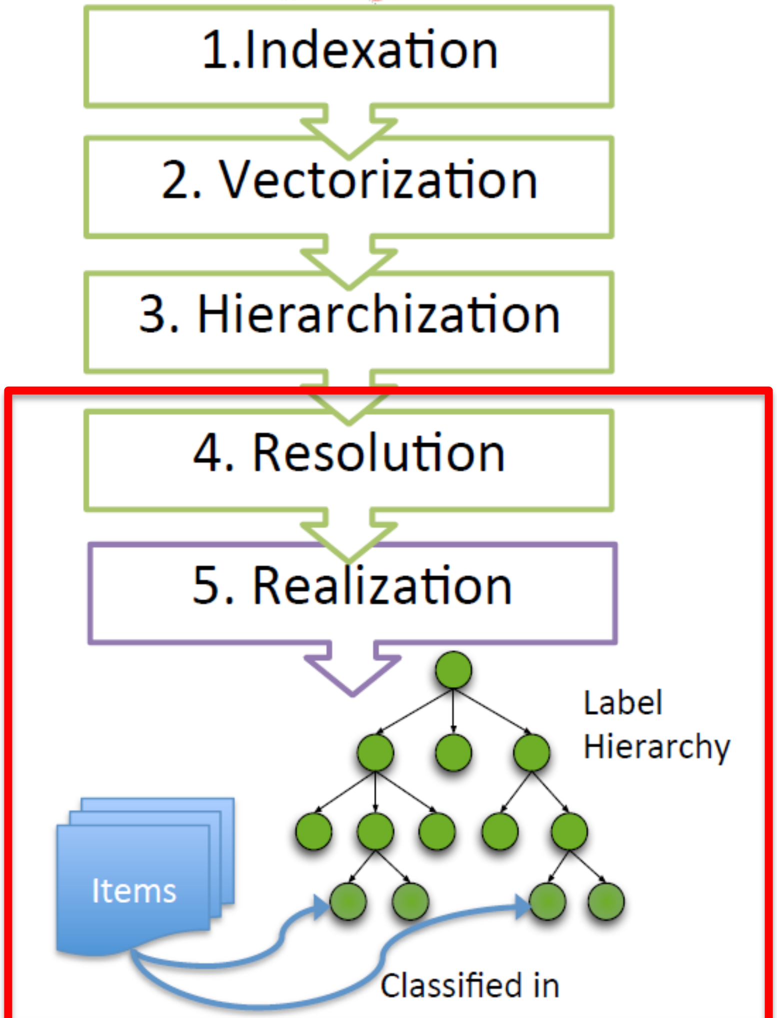
Rule-based reasonning to perform **Classification**



Unsupervised **ontology learning**

Rule-based Classification

Specific Problem



- *Resolution:* Learn **classifications rules** from **large volumes** of unstructured text

  Distributed method that exploits the coocurrence matrix

- *Realization:* classify **large volumes** of new **data items**

  Classification using a Web Reasonner

Proposed solution: rule learning (Resolution)

Learning **Alpha** and **Beta** sets

P_C(i|j)	$term_1$	$term_2$	$term_3$	$term_4$	$term_5$	$term_6$	$term_7$
$label_1$	0	0	5	0	5	25	25
$label_2$	0	75	0	0	0	75	5
$label_3$	0	0	75	0	25	0	0
$label_4$	5	25	25	0	5	93	25
$label_5$	95	0	0	0	60	0	5
$label_6$	0	60	0	95	0	0	90
$label_7$	5	98	5	60	25	0	79

Coocurrence:

$$P_C(term_i|term_j) = \frac{cfm\ (term_i, term_j)}{cfm(term_j, term_j)}$$

Alpha set:

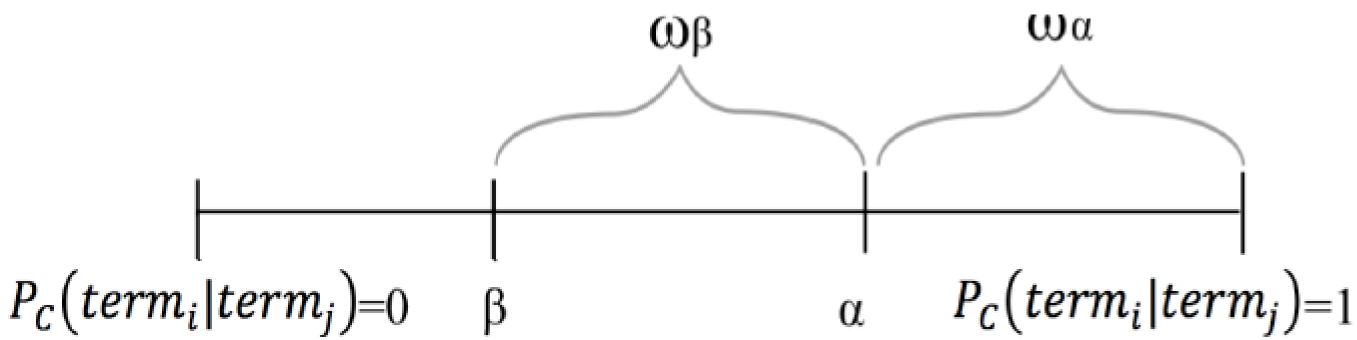
$$\omega_\alpha^{t_i} = \{t_j | \forall t_j \in Term{:}\ P_C(t_i|t_j) > \alpha\}$$

Beta set:

$$\omega_\beta^{t_i} = \{t_j | \forall t_j \in Term{:}\ \beta \leq P_C(\ t_i|t_j) \leq \alpha\}$$

Proposed solution: rule learning (Resolution)

Learning **Alpha** and **Beta** sets



Alpha set: $\omega_{\alpha}^{t_i} = \{t_j | \forall t_j \in Term\colon\ P_C(t_i|t_j) > \alpha\}$

Beta set: $\omega_{\beta}^{t_i} = \{t_j | \forall t_j \in Term\colon\ \beta \leq P_C(\ t_i|t_j) \leq \alpha\}$

Proposed solution: rule learning (Resolution)

Example:

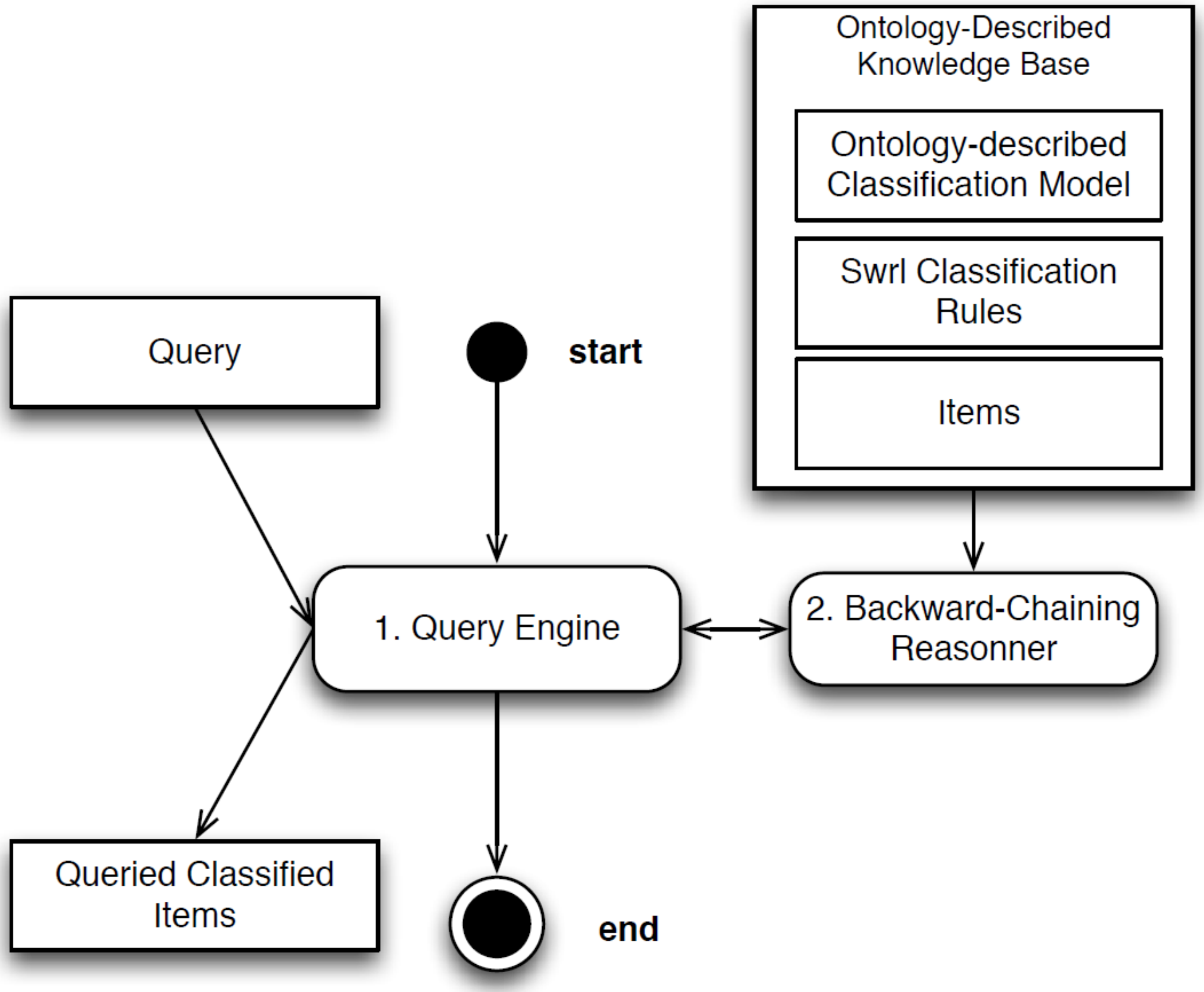
%	$term_1$	$term_2$	$term_3$	$term_4$	$term_5$	$term_6$	$term_7$
$label_1$	0	0	5	0	5	25	25
$label_2$	0	75	0	0	0	75	5
$label_3$	0	0	75	0	25	0	0
$label_4$	5	25	25	0	5	93	25
$label_5$	95	0	0	0	60	0	5
$label_6$	0	60	0	95	0	0	90
$label_7$	5	98	5	60	25	0	79

$$\omega_{\alpha}^{t_i} = \{t_j | \forall t_j \in Term\colon\ P_C(t_i|t_j) > \alpha\},\ \alpha = 91$$

$$\omega_{\beta}^{t_i} = \{t_j | \forall t_j \in Term\colon\ \beta \leq P_C(t_i|t_j) \leq \alpha\},\ \beta = 70$$

Proposed solution: classification with web reasoner

Classification at **query-time** using **backward-chaining**



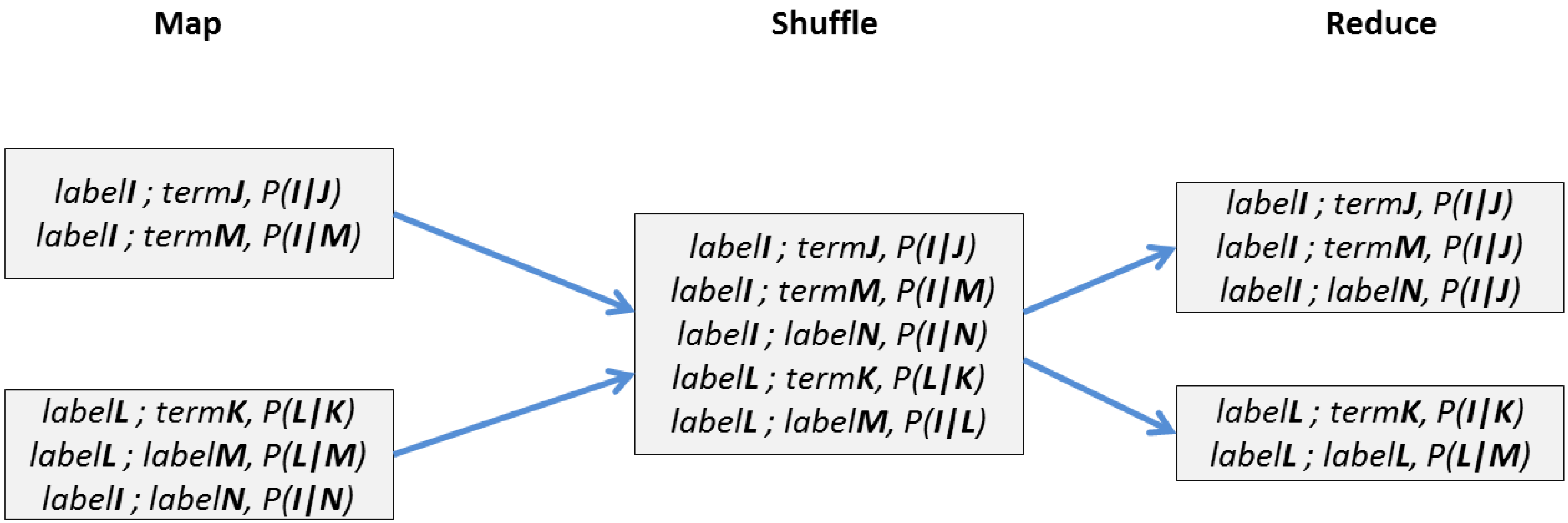
Core Ontology

DL concepts	**Description**
$Item \sqsubseteq \exists hasTerm.Term$	Items to classify (e.g. doc) has terms
$Term \sqsubseteq \top$	Terms (e.g. word) extracted from items
$Label \sqsubseteq Term$	Labels are terms used to classify items
$Label \sqsubseteq \forall broader.Label$	Broader relation between labels
$Label \sqsubseteq \forall narrower.Label$	Narrower relation between labels
$broader \equiv narrower^{-}$	Broader and narrower are inverse
$Item \sqcap Term = \emptyset$	Items and Terms are disjoint
$Item \sqsubseteq \forall isClassified.Label$	Relation that links items with labels

Outline

- Context
  - Global problem
  - The Semantic HMC
- Specific Problem
  - Proposed Solution
- **Implementation**
  - Setup
  - Results
- Conclusion and future work

Distributed process using mapreduce:



OWL API used to generate SWRL rules from the output

$$Item(?it), Term(term_i), Label(term_i), hasTerm\left(?it, term_j\right) \rightarrow isClassified(?it, term_i)$$

Implementation: rule creation

Generated rules Exemple

Alpha rules
$Item(?\,it), Term(\mathrm{t}_1), Label(term_i), hasTerm(?\,it, t_1) \rightarrow isClassified(?\,it, term_i)$
$Item(?\,it), Term(\mathrm{t}_2), Label(term_i), hasTerm(?\,it, t_2) \rightarrow isClassified(?\,it, term_i)$

Beta rules
$Item(?\,it), Term(\mathrm{t}_1), \mathrm{Term}(t_2), Label(term_i), \mathrm{hasTerm}(?\,it, t_1), hasTerm(?\,it, t_2) \rightarrow isClassified(?\,it, term_i)$
$Item(?\,it), Term(\mathrm{t}_1), \mathrm{Term}(t_3), Label(term_i), hasTerm(?\,it, t_1), hasTerm(?\,it, t_3) \rightarrow isClassified(?\,it, term_i)$
$Item(?\,it), Term(\mathrm{t}_2), \mathrm{Term}(t_3), Label(term_i), hasTerm(?\,it, t_2), hasTerm(?\,it, t_3) \rightarrow isClassified(?\,it, term_i)$

Implementation: Classification at query-time

Stardog used as a scalable triple-store (compatible with **backward-chaining** inference as well as **SWRL** rules inference)

Rule selection process developped in Java interacting with Stardog to optimize query performance

hadoop

Stardog

Implementation: test environment

Dataset

WIKIPEDIA
The Free Encyclopedia

Sub-Dataset	Number of articles
Wikipedia 1	174900
Wikipedia 2	407000
Wikipedia 3	994000

Cluster

Google Cloud Platform

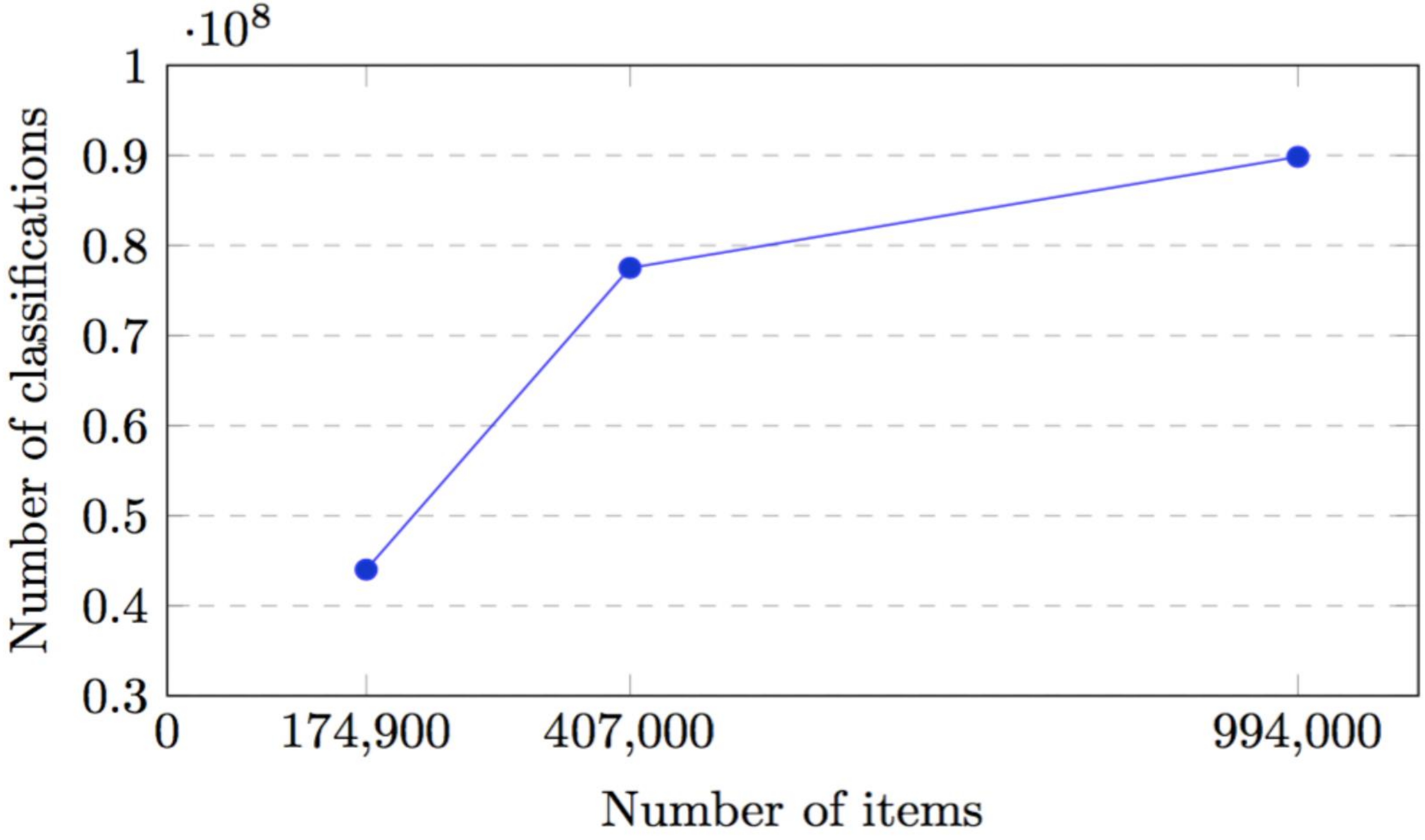
Resource type	Description
Number of nodes	4
CPU (per node)	Intel Xeon E5 v2
RAM (per node)	7.5GB
Disk (per node)	500GB

Implementation: parameter setup

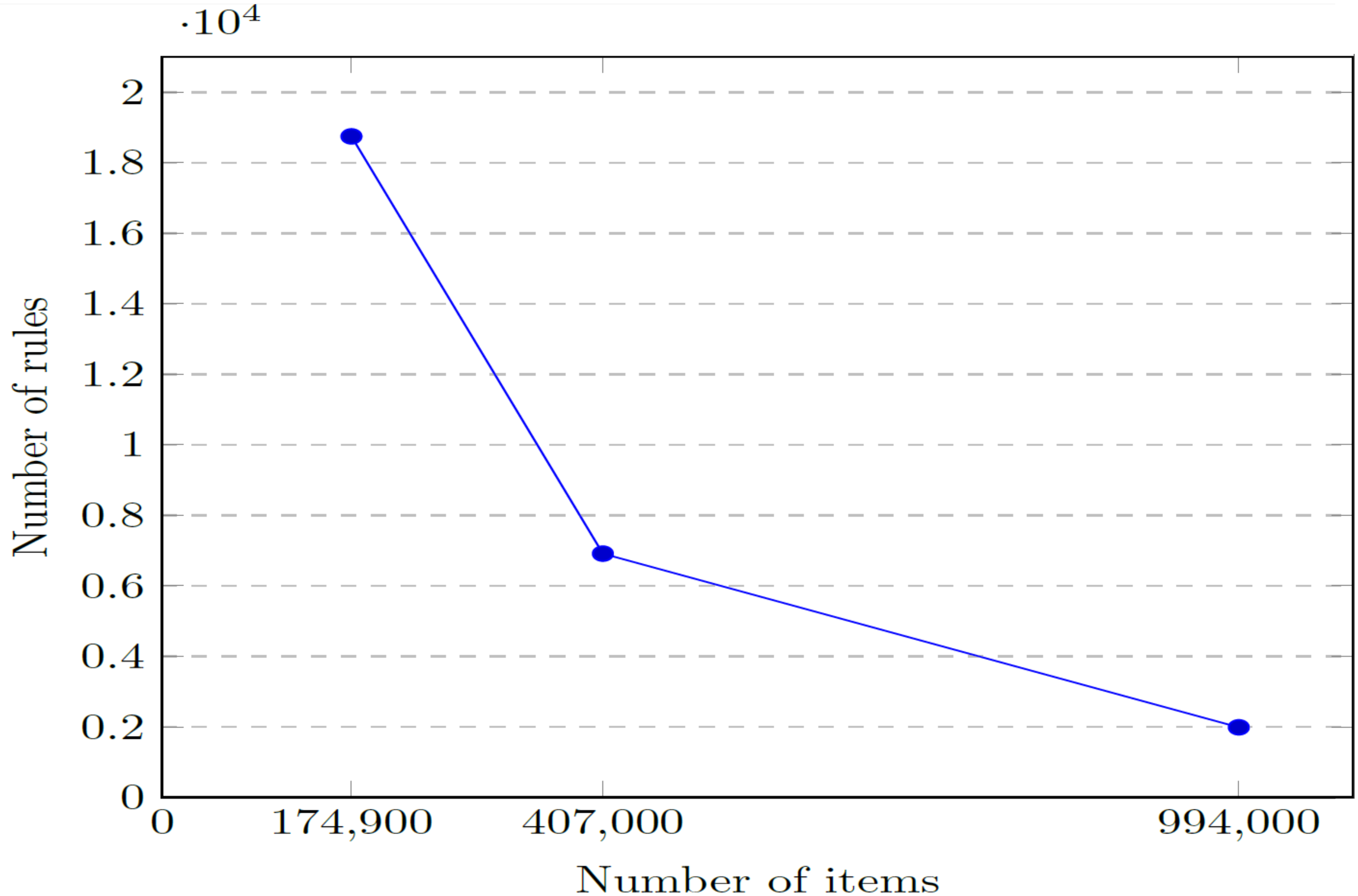
Parameter	Step	Value
Alpha Threshold	Resolution	90
Beta Threshold		80
Term ranking (n)		5
p		0.25
Term Threshold (γ)	Realization	2

Results

Number of **classifications:** $Item \sqsubseteq \forall isClassified.Label$

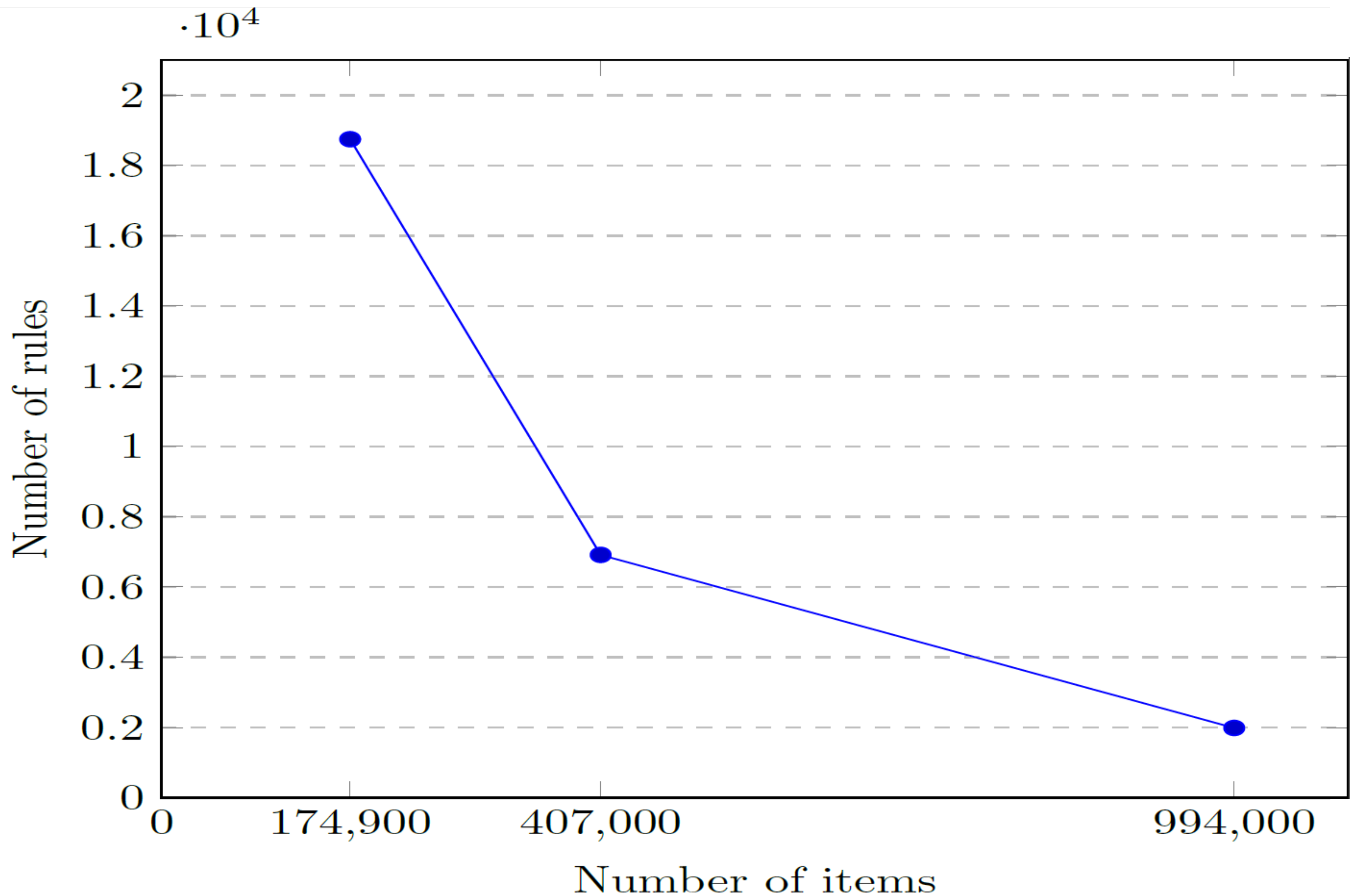


Number of **learned rules** (Alpha + Beta)



Results

Number of **learned rules** (Alpha + Beta) α = 90 β = 80



Outline

Context

- Global problem
- The Semantic HMC

Specific Problem

- Proposed Solution

Implementation

- Setup
- Results

Conclusion and future work

Conclusion

- A new unsupervised process to automatically classify items
  - A highly scalable rule learning method based on statistical and lexical approaches
  - A novel method to classify items using a web reasoner
- The process prototype was successfully implemented in a scalable and distributed platform to process Big Data
- Preliminary results show that the items are classified automatically by the reasonner

Ongoing and Future Work

- Quality Evaluation of the process: comparison with state-of-the art in classification
- Predictive performance evaluation based on cross-validation with large dataset
- Optimization of the process by exhaustive analysis of parameters' impact
- Application to classification of news articles on the web

CHECKSEM
SEMANTICINTELLIGENCERESEARCH

An unsupervised classification process for large datasets using web reasoning

Thank you !

Rafael PEIXOTO, Thomas HASSAN, Christophe CRUZ, Aurelie BERTAUX, Nuno SILVA
thomas.hassan@u-bourgogne.fr

Laboratoire LE2I – UMR CNRS 6306 – Université de Bourgogne

LE2I UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ cnrs dépasser les frontières Bourgogne Conseil régional CHECKSEM SEMANTICINTELLIGENCERESEARCH FIRSTECO VOS DÉCISIONS PRENNENT UN TEMPS D'AVANCE