Ontology-based approach for unsupervised and adaptive focused crawling

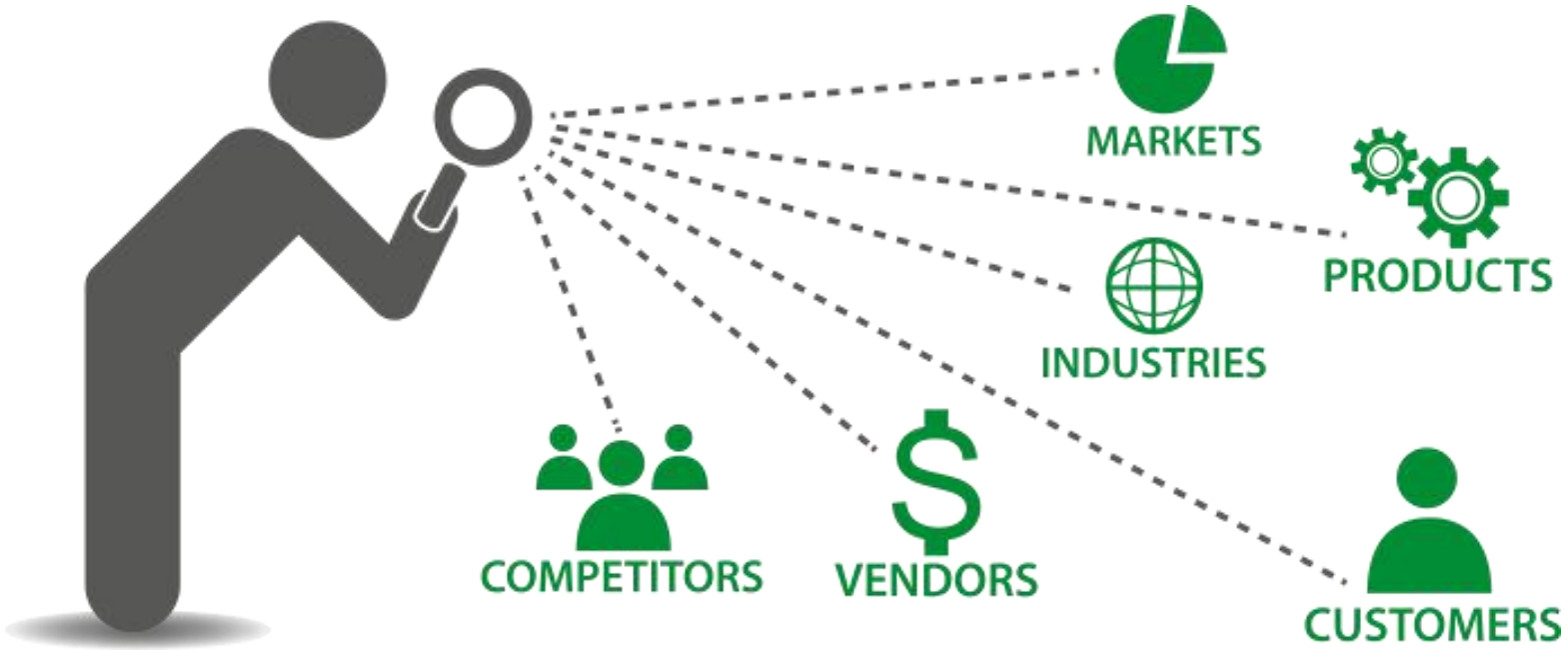
Thomas HASSAN, Christophe CRUZ, Aurélie Bertaux
thomas.hassan@u-bourgogne.fr

Le2i FRE2005, CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté
Dijon, France

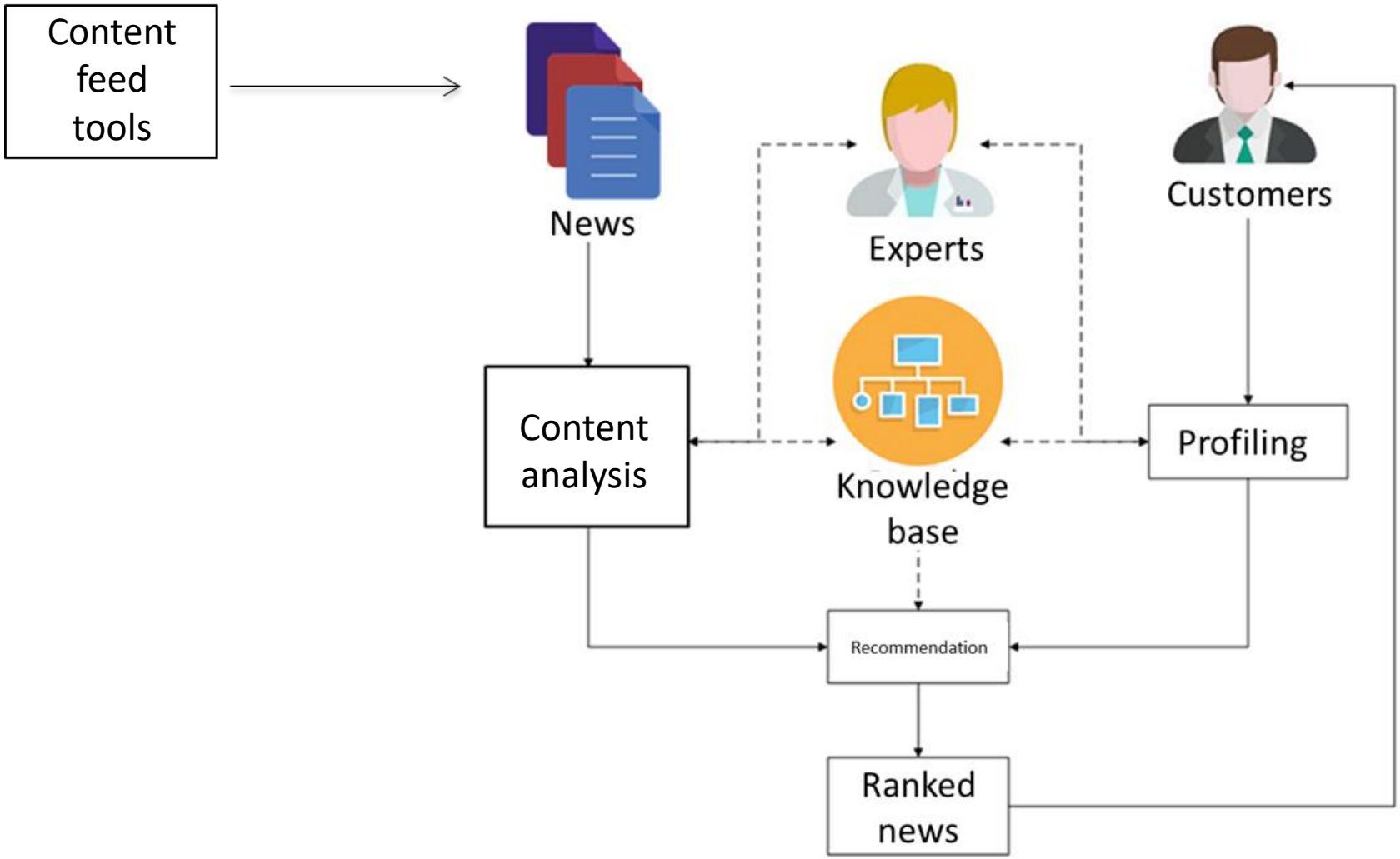
Outline

- **Context**
  - Industrial context
  - Problem statement
- Proposed solution
  - Background
  - Architecture
- Evaluation
  - Scaling
  - Performance
- Conclusion and future work

Competitive intelligence

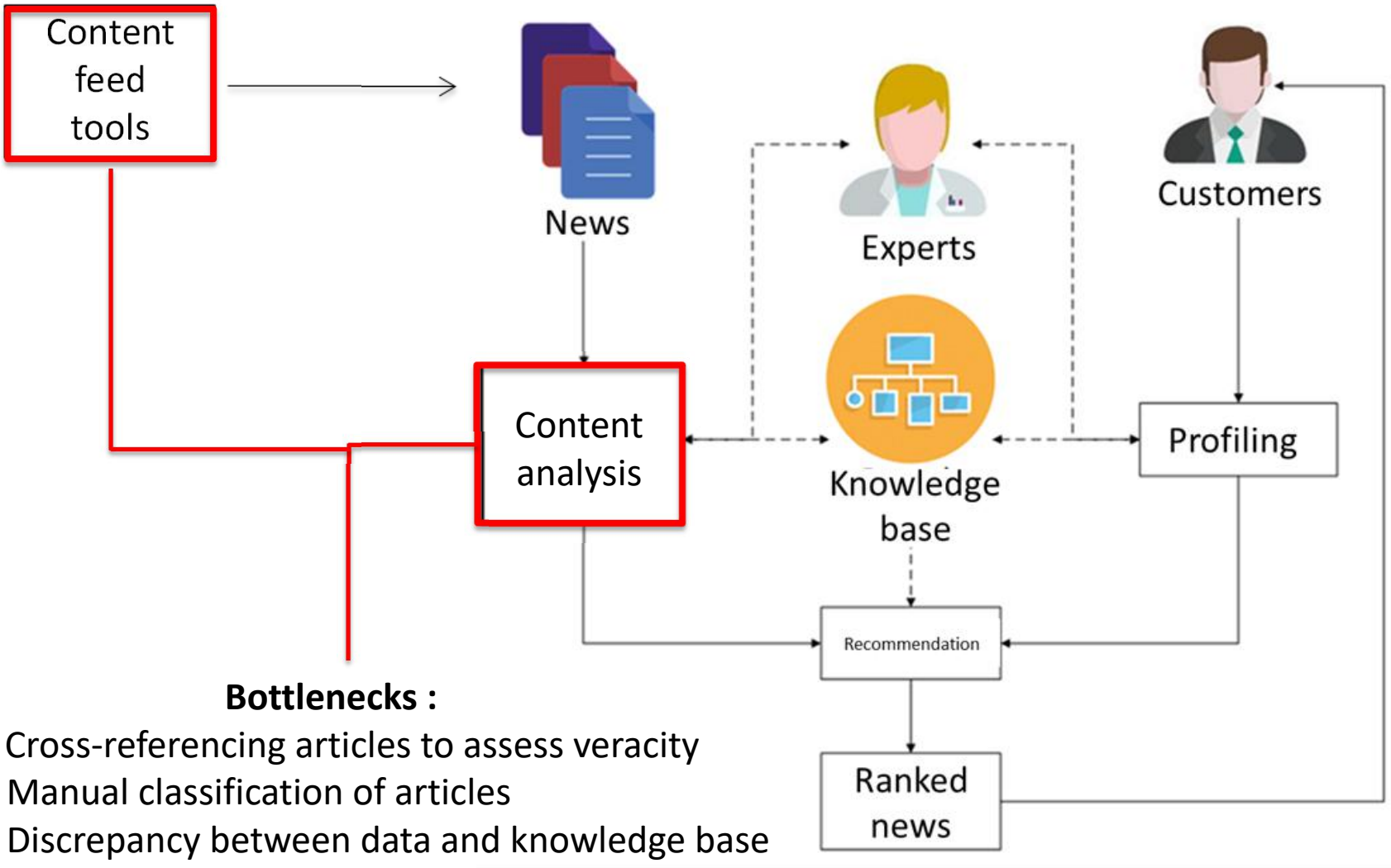


Industrial context



FIRSTECO
VOS DÉCISIONS PRENNENT UN TEMPS D'AVANCE

Problem statement



Bottlenecks :

- Cross-referencing articles to assess veracity
- Manual classification of articles
- Discrepancy between data and knowledge base

High time cost for experts, possible loss of information

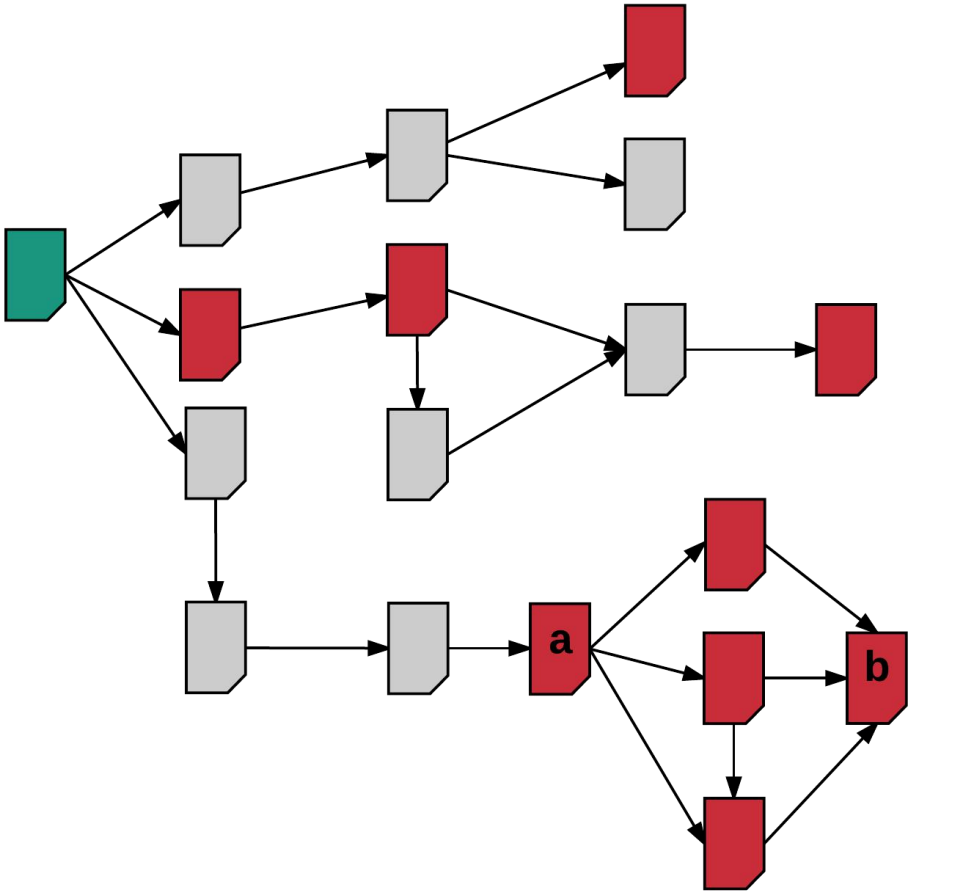
Problem statement

- How to specialize feed tools with domain-specific knowledge ?
- How to optimize content gathering to find most relevant items fast ?
- How to expand information sources horizon ?

Outline

- Context
  - Industrial context
  - Problem statement
- **Proposed solution**
  - Background
  - Architecture
- Evaluation
  - Scaling
  - Performance
- Conclusion and future work

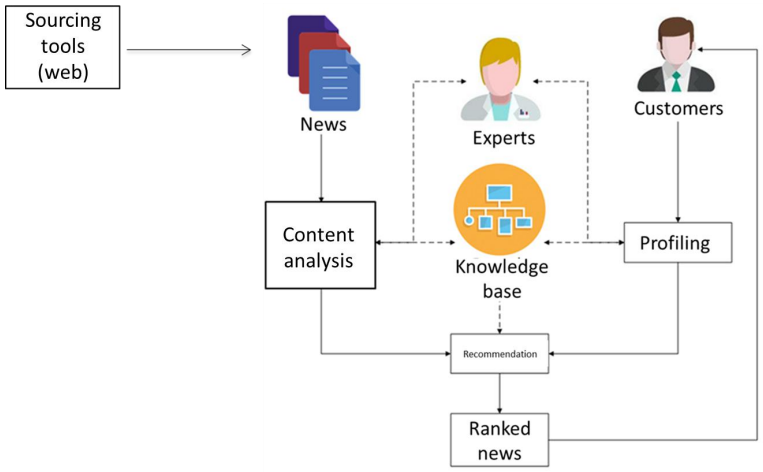
Background : focused crawler

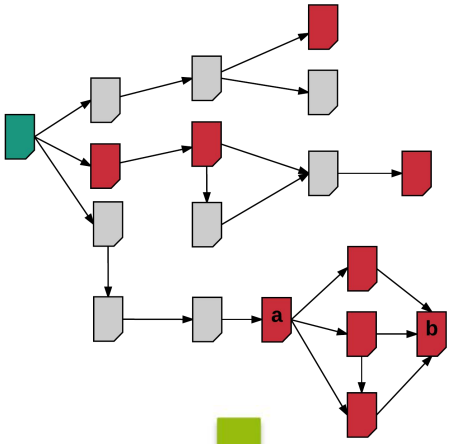


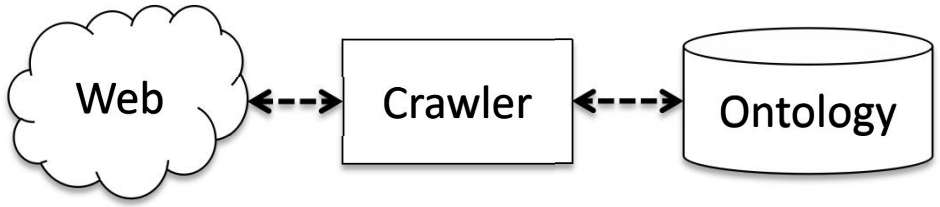
Relevant Irrelevant Seed item Inlink

a : "Hub" **b :** "Authority"

Background : focused crawler + semantics







Efficient content gathering

Relevant content analysis

Limitations

1) Dynamic data VS **static** ontology :

Discrepancy between ontology-based classifier and actual web data

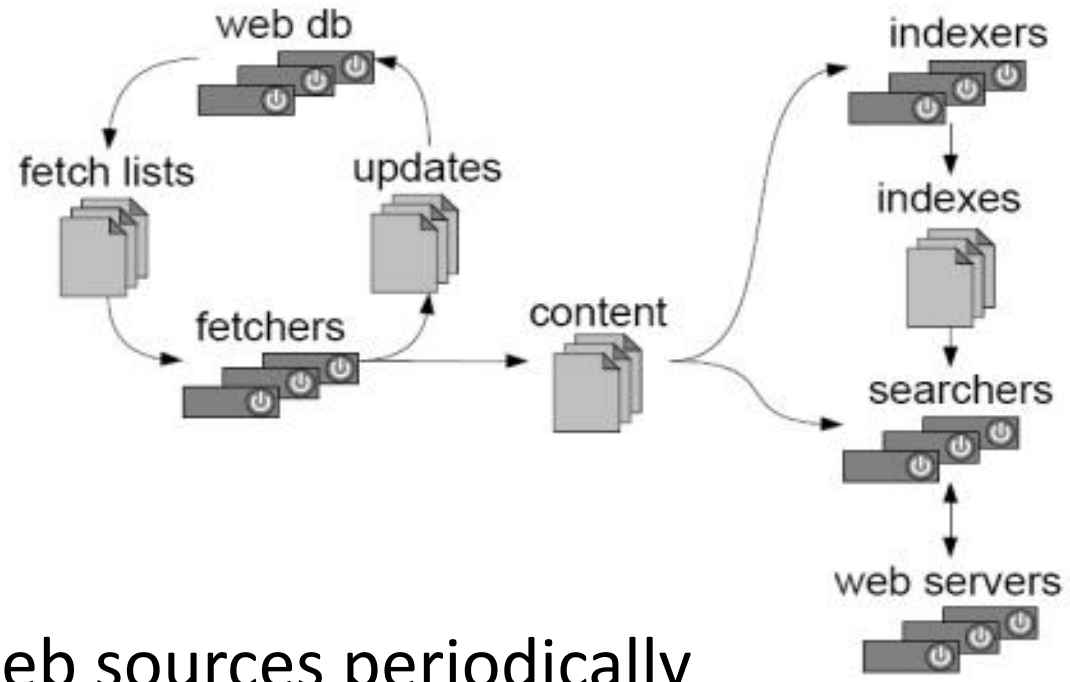
2) Crawler should improve from **experience :**

Both **content** and **graph** mining should be useed to enhance crawling performance

Objectives : **adapt** both crawling experience and content analysis **over time** to **accelerate crawling** and improve **relevance**

Architecture : baseline implementation

Based on Nutch, hadoop-based distributed crawler



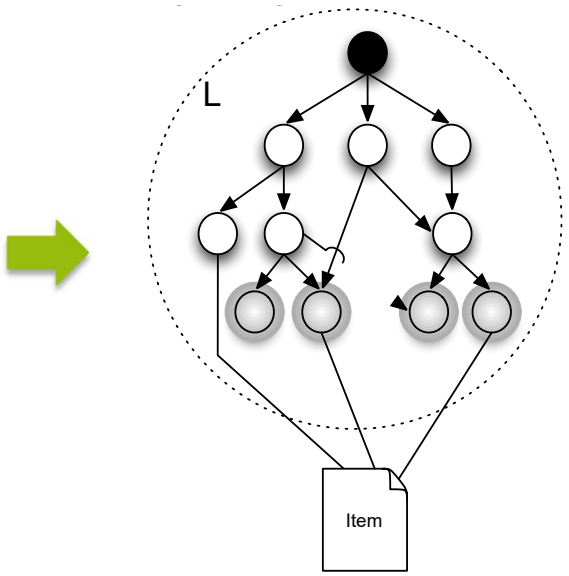
- Crawl web sources periodically
- High throughput, fault tolerance
- Integrate usefull modules

Diagram from : https://nutch.wordpress.com/

Architecture : classification module

Classification model construction based on probability distribution of features

$P_C(i\|j)$	$term_1$	$term_2$	$term_3$	$term_4$	$term_5$	$term_6$	$term_7$
$label_1$	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



Multi-label Hierarchical Classification

Architecture : classification module

Objective : content-based **classification** of items

MENU **Le Point**

Groupe PSA lève 500 M EUR dans sa première émission obligataire depuis 2013

AFP

Publié le 11/04/2016 à 09:48 | AFP

ABONNEZ-VOUS À PARTIR DE 1€

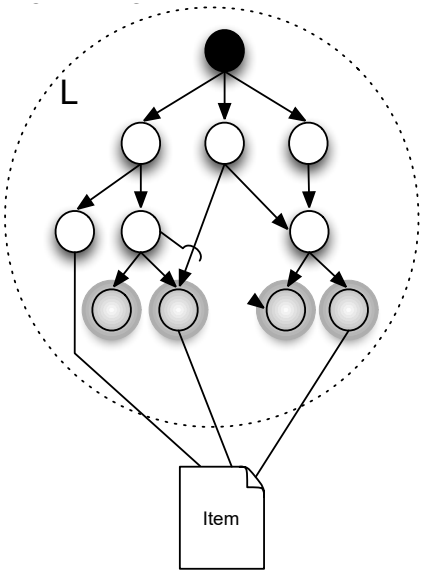
Le constructeur automobile PSA a annoncé lundi avoir levé 500 millions d'euros à l'occasion de sa première émission obligataire depuis 2013, une opération destinée à refinancer sa dette.

Lancée vendredi, l'opération "a été sur-souscrite 7,6 fois", a souligné le groupe dans un communiqué.

"Profitant de conditions de marché favorables, cette opération d'une maturité de 7 ans (avril 2023) permet d'allonger la maturité de la dette du groupe à des coûts historiquement bas (coupon annuel de 2,375%)", a-t-il détaillé.

Cette levée de fonds "illustre la réussite de notre plan de reconstruction économique +Back in the Race+, et la confiance des investisseurs dans notre nouveau plan de croissance rentable +Push to Pass+ présenté le 5 avril dernier", a commenté le directeur financier de l'entreprise, Jean-Baptiste de Chatillon, cité dans le communiqué.

Après avoir frôlé le dépôt de bilan en 2013-2014, PSA a lancé le plan de restructuration drastique "Back in the race" pour moderniser ses usines, réduire ses coûts et ses stocks.



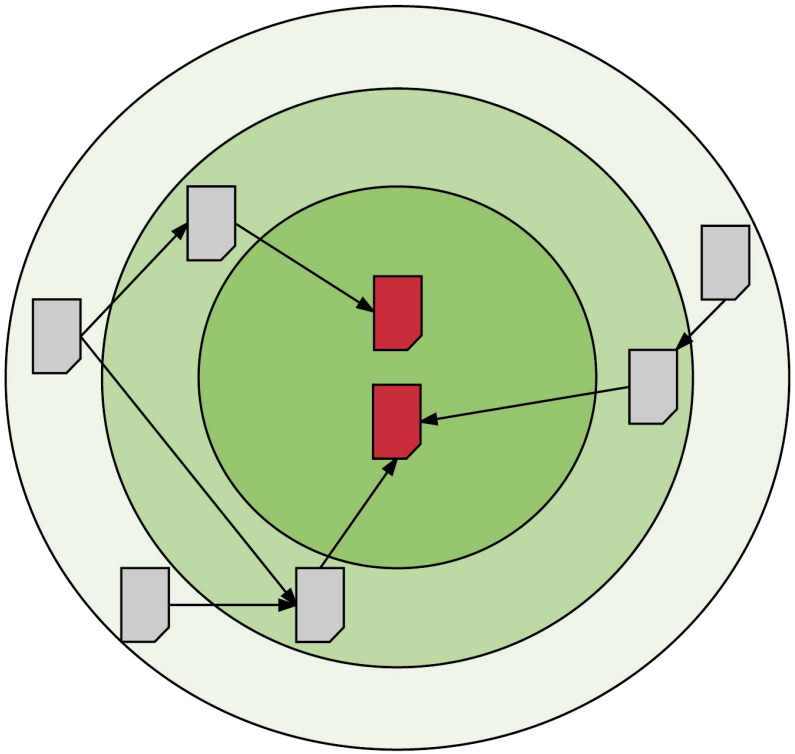
Multi-label Hierarchical Classification

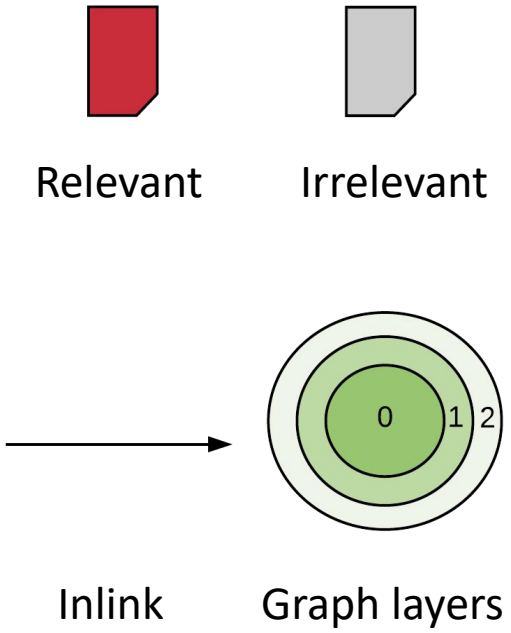
Each document represented as a **vector of terms** it contains (Lucene)

Outputs a **vector of labels** (relevant concepts of the ontology) for each item

Architecture : priority module

Use the context-graph approach to estimate relevance of unseen links. Computes similarity with fetched items based on the distance to relevant items

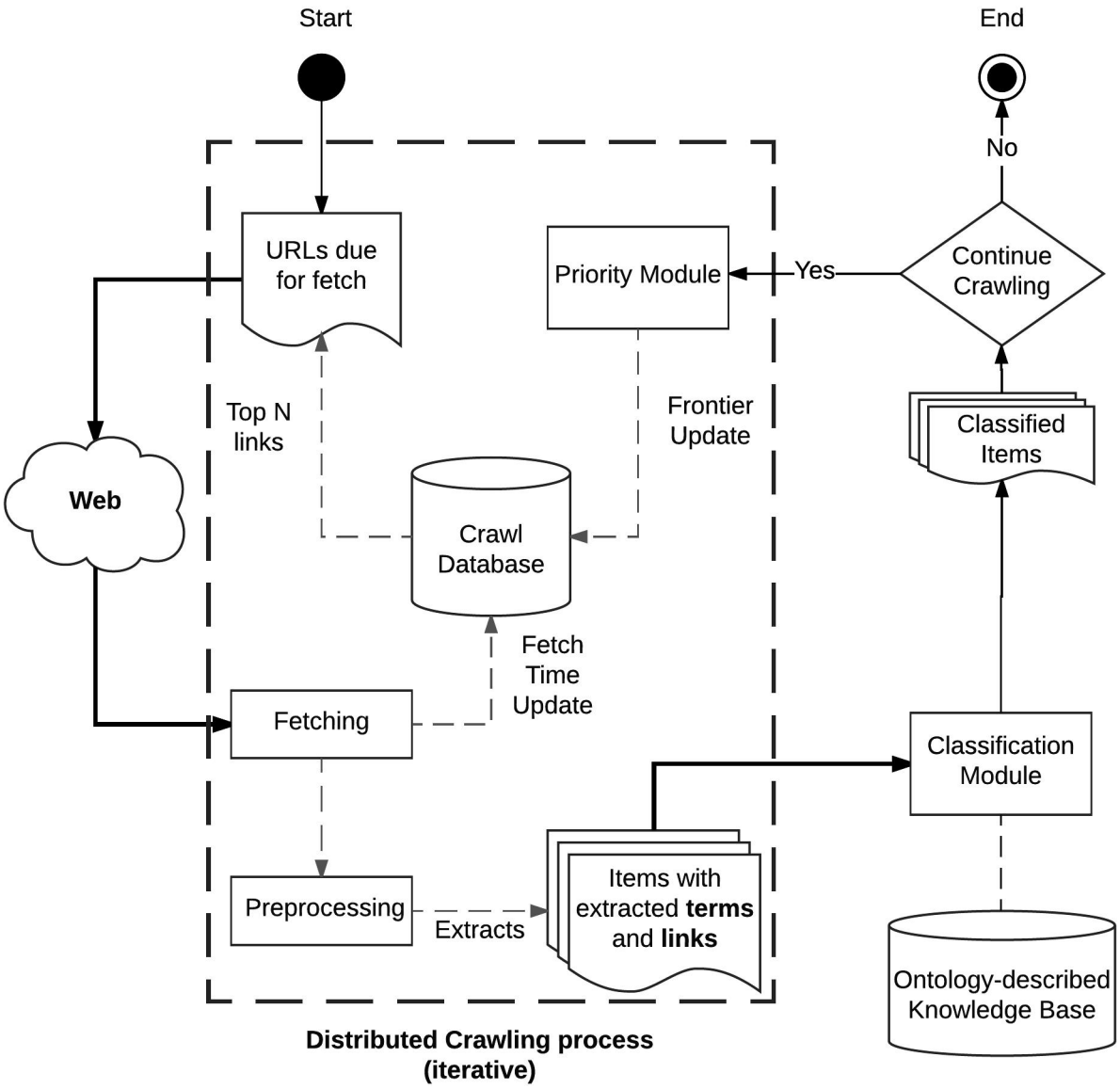




Diligenti, et al., 2000. Focused Crawling Using Context Graphs. In *VLDB* (pp. 527-534).

Architecture : classification module

Integration with the crawler

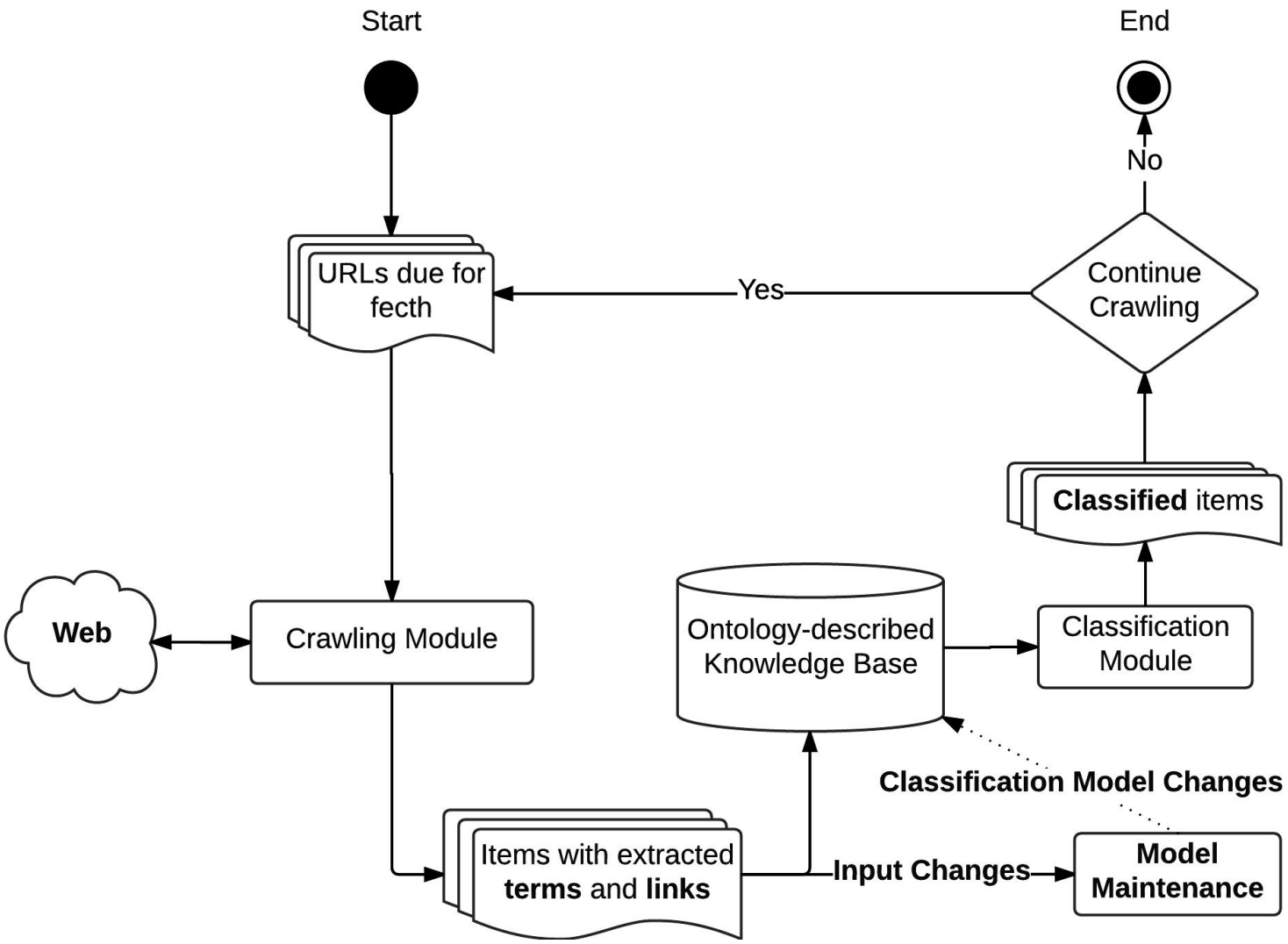


Architecture : maintenance module

Objective : maintain a cooccurrence matrix of features

$P_C(i \mid j)$	$term_1$	$term_2$	$term_3$	$term_4$	$term_5$	$term_6$	$term_7$
$label_1$	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

Architecture : maintenance module

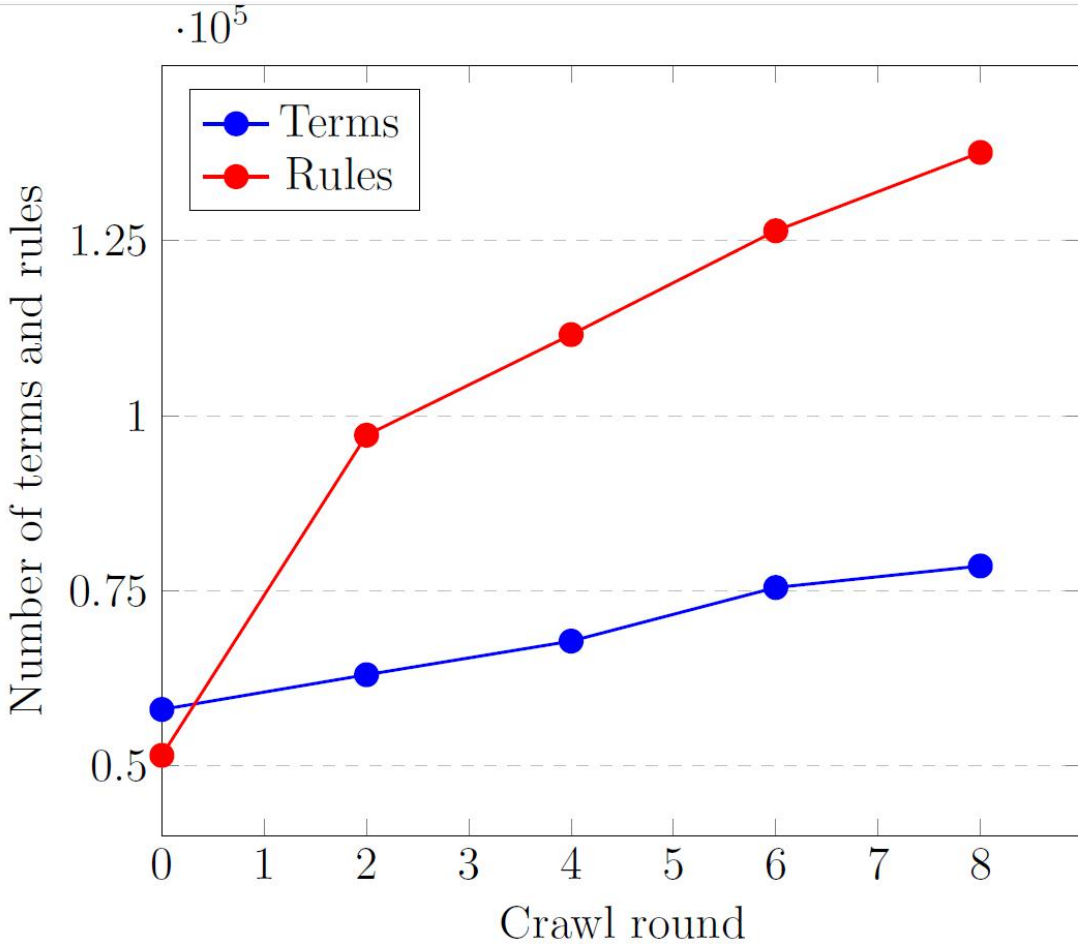


Outline

- Context
  - Industrial context
  - Problem statement
- Proposed solution
  - Background
  - Architecture
- **Evaluation**
  - Scaling
  - Performance
- Conclusion and future work

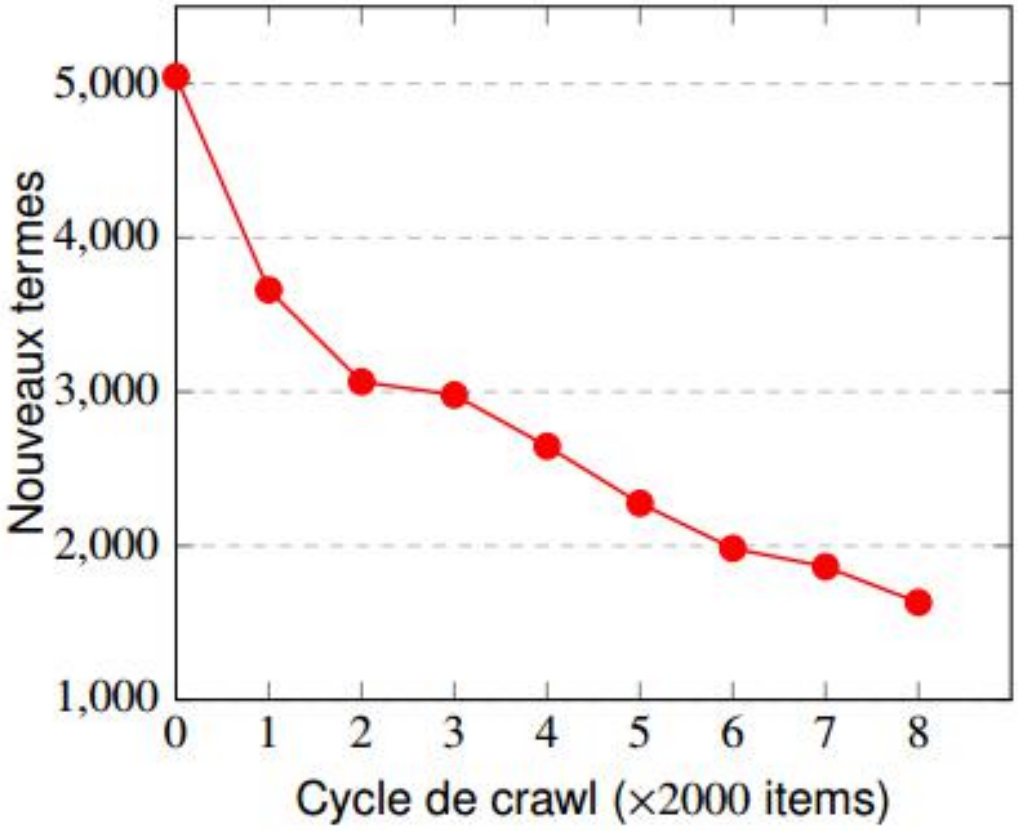
Scaling

Distributed architecture to deal with scaling



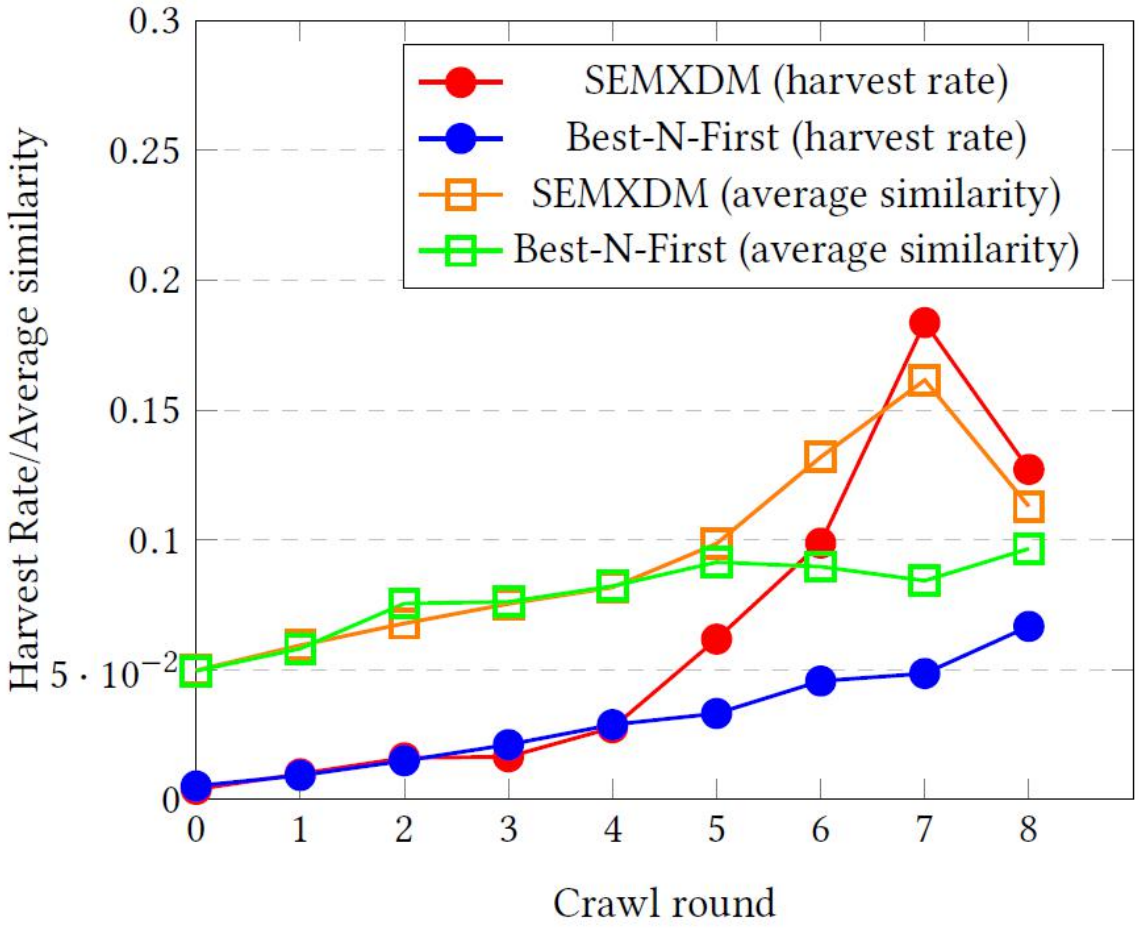
Scaling

Distributed architecture to deal with scaling



Quality Evaluation

Comparison with standard Best-N-First using only cosine similarity



Outline

- Context
  - Industrial context
  - Problem statement
- Proposed solution
  - Background
  - Architecture
- Evaluation
  - Scaling
  - Performance
- **Conclusion and future work**

Conclusion

- An approach for unsupervised ontology-based focused crawling
  - Performs cross-referencing of web items
  - Ontology-based classification model for accurate item classification
  - Adaptation and evolution of the model using web content and web graph mining
- Future work
  - Evaluation of the architecture in industrial context
  - Leverage scalability issues of the maintenance process.
  - Active learning integration in the maintenance process (expert feedback)

Ontology-based approach for unsupervised and adaptive focused crawling

Thank you !

Thomas HASSAN, Christophe CRUZ, Aurélie Bertaux
thomas.hassan@u-bourgogne.fr

Le2i FRE2005, CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté
Dijon, France