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**
Industrial context

Competitive intelligence
Industrial context

Content feed tools

Content analysis

News

Experts

Knowledge base

Profiling

Recommendation

Ranked news

Customers
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?
• 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
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 used 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 useful modules

Diagram from: https://nutch.wordpress.com/
## Architecture: classification module

Classification model construction based on probability distribution of features

| $P_c(l|l)$ | term₁ | term₂ | term₃ | term₄ | term₅ | term₆ | term₇ |
|------------|-------|-------|-------|-------|-------|-------|-------|
| label₁     | 0     | 0     | 5     | 0     | 5     | 25    | 25    |
| label₂     | 0     | 75    | 0     | 0     | 0     | 75    | 5     |
| label₃     | 0     | 0     | 75    | 0     | 25    | 0     | 0     |
| label₄     | 5     | 25    | 25    | 0     | 5     | 93    | 25    |
| label₅     | 95    | 0     | 0     | 0     | 60    | 0     | 5     |
| label₆     | 0     | 60    | 0     | 95    | 0     | 0     | 90    |
| label₇     | 5     | 98    | 5     | 60    | 25    | 0     | 79    |
Multi-label Hierarchical Classification

Architecture: classification module

Objective: content-based classification of items

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

Outputs a vector of labels (relevant concepts of the ontology) for each item
Use the context-graph approach to estimate relevance of unseen links. Computes similarity with fetched items based on the distance to relevant items.

Integration with the crawler

Architecture: classification module
Architecture: maintenance module

Objective: maintain a cooccurrence matrix of features

| \( P_C(i|j) \) | term₁ | term₂ | term₃ | term₄ | term₅ | term₆ | term₇ |
|--------------|-------|-------|-------|-------|-------|-------|-------|
| label₁       | 0     | 0     | 5     | 0     | 5     | 25    | 25    |
| label₂       | 0     | 75    | 0     | 0     | 0     | 75    | 5     |
| label₃       | 0     | 0     | 75    | 0     | 25    | 0     | 0     |
| label₄       | 5     | 25    | 25    | 0     | 5     | 93    | 25    |
| label₅       | 95    | 0     | 0     | 0     | 60    | 0     | 5     |
| label₆       | 0     | 60    | 0     | 95    | 0     | 0     | 90    |
| label₇       | 5     | 98    | 5     | 60    | 25    | 0     | 79    |
Outline

• Context
  ▪ Industrial context
  ▪ Problem statement

• Proposed solution
  ▪ Background
  ▪ Architecture

• Evaluation
  ▪ Scaling
  ▪ Performance

• Conclusion and future work
Distributed architecture to deal with scaling
Distributed architecture to deal with scaling
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