Processus de classification multi-label hiérarchique sémantique pour le Big Data

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

PLAN

1. Contexte, Problématique
2. Prototype
   - Contraintes, objectifs
   - Architecture
3. Processus
4. Travaux futurs
   - Evaluations
   - Améliorations
5. Conclusion

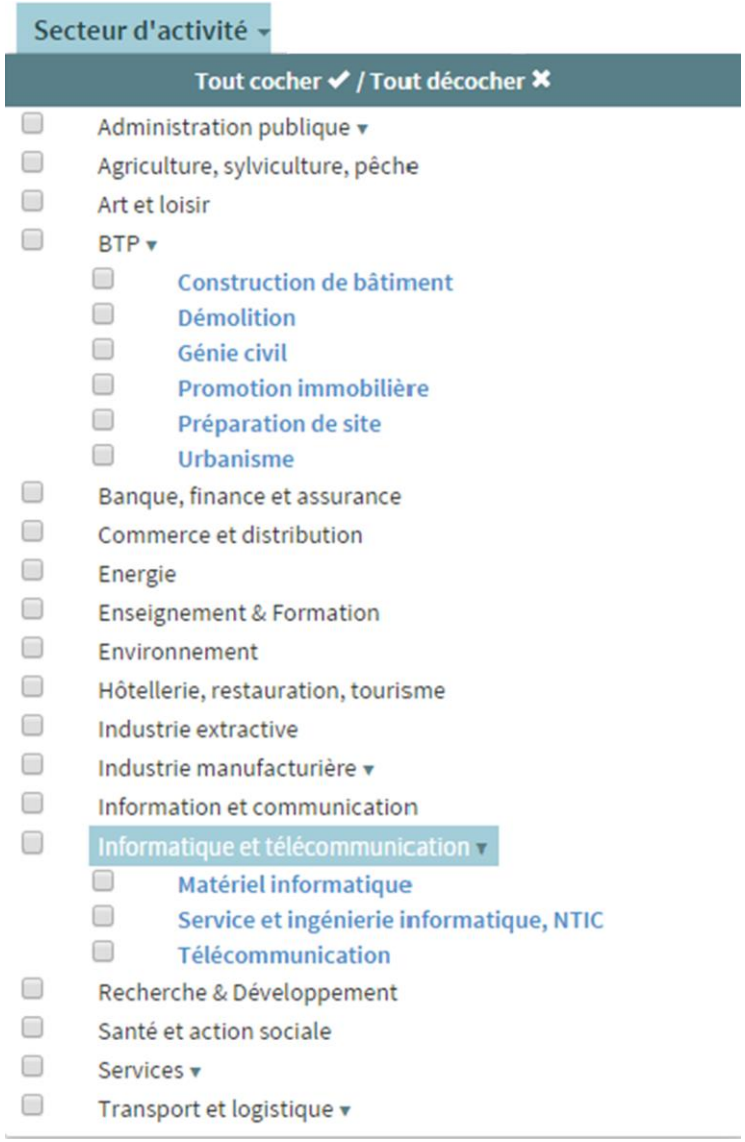
Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

CONTEXTE

FIRSTECOPRO'FIL
BOOSTEZ VOTRE BUSINESS AVEC LA VEILLE PERSONNALISÉE

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

CONTEXTE



Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

CONTEXTE

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

CONTEXTE

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

PROBLEMATIQUE

Problème :

Réduire la quantité d'informations à traiter, tout en gardant une représentation proche des connaissances des experts.

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

PROBLEMATIQUE

Littérature :

- **Systèmes existants restreints à un cadre précis**
- **Pas de système générique**
- **Les systèmes n'évoluent pas en fonction des données**

Verrou scientifique :

Pas de processus d'analyse de données générique, à la fois adapté aux grands volumes de données et à des connaissances métier.

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Contraintes

- **Traitement générique**

  ➔Adaptation à différents cas d'utilisation

- **Evolution du processus en fonction des données**

  ➔ Apprentissage continu
  ➔ Montée en charge

- **Recherche d'informations de Valeur pour les entreprises**

  ➔ Utilisation de la connaissance métier
  ➔ Découverte d'informations dirigée par les données

- **Mise en corrélation des informations**

  ➔ Rapprochement des données similaires

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Contraintes

Passage à l'échelle

Stocker et analyser de grands volumes de données

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Contraintes

Hétérogénéité (Variété) des données

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Contraintes

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <rss version="2.0">
  - <channel>
      <title>Hobbyman.se</title>
      <link>http://www.hobbyman.se/</link>
      <description>Tips och Trix för hemmet</description>
      <language>en-us</language>
      <generator>Nucleus CMS v3.23</generator>
      <copyright>©</copyright>
      <category>Weblog</category>
      <docs>http://backend.userland.com/rss</docs>
    - <image>
        <url>http://www.hobbyman.se//nucleus/nucleus2.gif</url>
        <title>Hobbyman.se</title>
        <link>http://www.hobbyman.se/</link>
      </image>
    - <item>
        <title>Så här gör du ditt eget t-shirt tryck med strykjärn...</tit
        <link>http://www.hobbyman.se/index.php?itemid=10</link>
      - <description>
        - <![CDATA[
            <table border="0"><tr><td valign="0">Fyller din kom
            inte vad du ska ge i present? Ge en personlig prese
            T-shirt (vit eller färgad)<br />
```

Items

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Contraintes

Intégration des connaissances métier

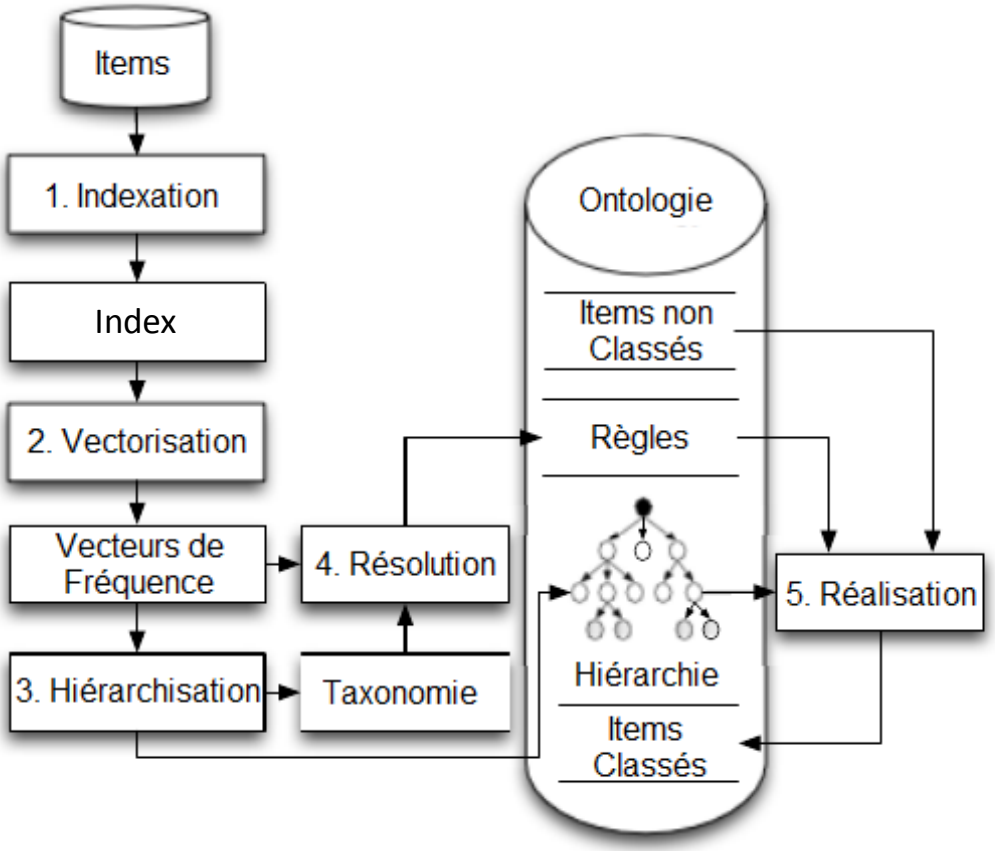
Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Objectifs

- **Traitement des données brutes, réduction de la dimension**
- **Recherche d'informations pertinentes**
- **Création d'un modèle de classification évolutif (ontologie)**
- **Classification des nouveaux items**

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

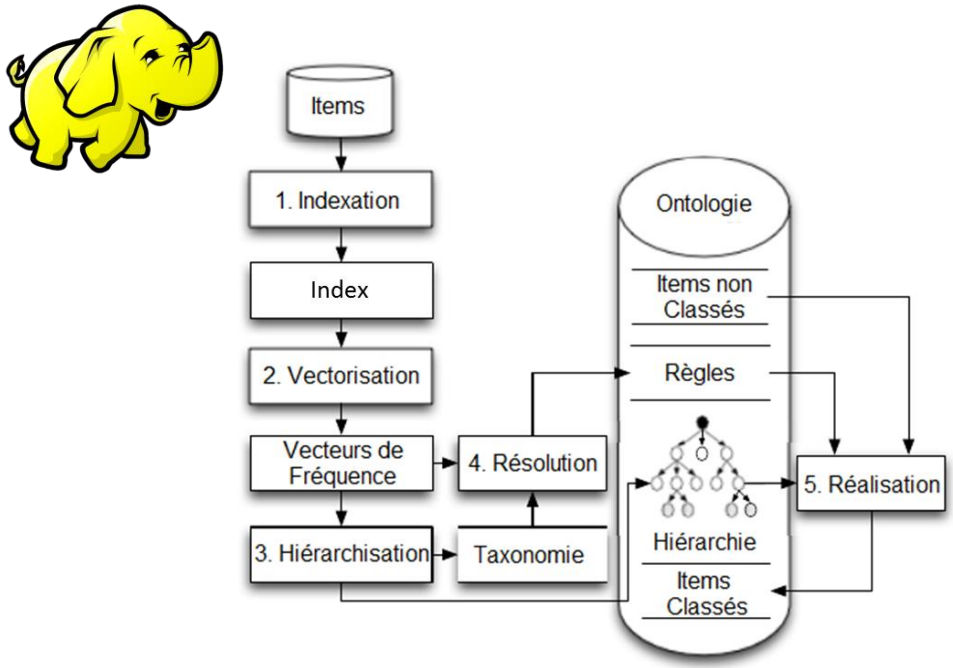
Architecture : processus à 5 phases distinctes



Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Montée en charge :

Utilisation du framework Hadoop et du modèle MapReduce



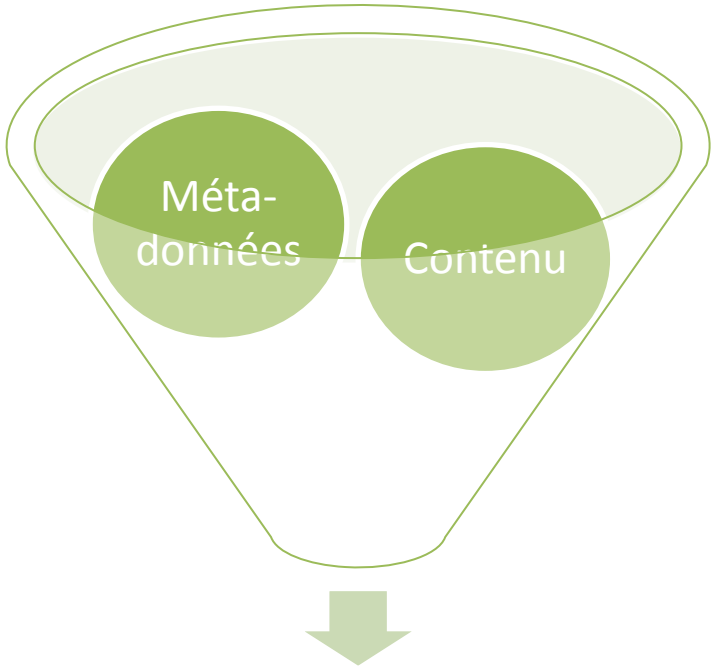
➔ **Comment distribuer chacune des phases du processus ?**

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Indexation

Génération d'un index des termes



Parsing, Traitement du Langage Naturel*

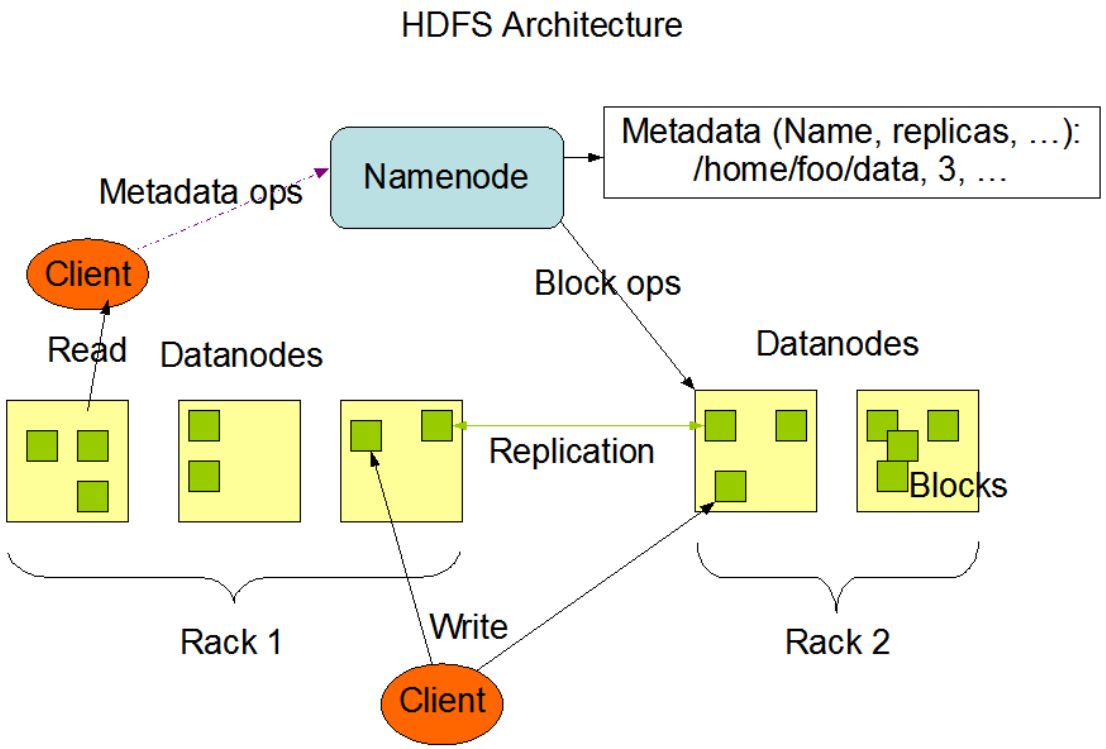
- Réduction de la dimension
- Suppression des données inutiles

* Tokenisation, Lemmatisation, Suppression des mots vides

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Indexation

Distribution de l'index sur un système de fichiers HDFS



Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Vectorisation

Détection des termes pertinents basée sur l'index.

Mesures statistiques telles que TF-IDF :

- Uni-grams (termes)
- N-grams (collocations)

➔ Définition des concepts basée sur les termes les plus fréquents

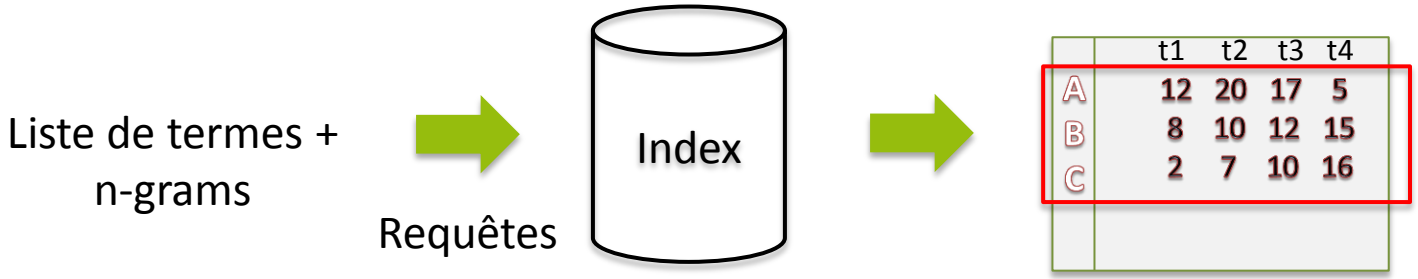
	t1	t2	t3	t4
A	12	20	17	5
B	8	10	12	15
C	2	7	10	16

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Vectorisation

Utilisation d'implémentations existantes de l'algorithme de collocations (MapReduce)

Requêtes sur l'index pour générer les vecteurs de fréquence (1 vecteur par concept) :

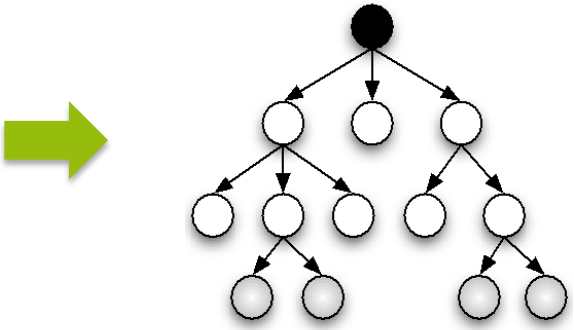


Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Hiérarchisation

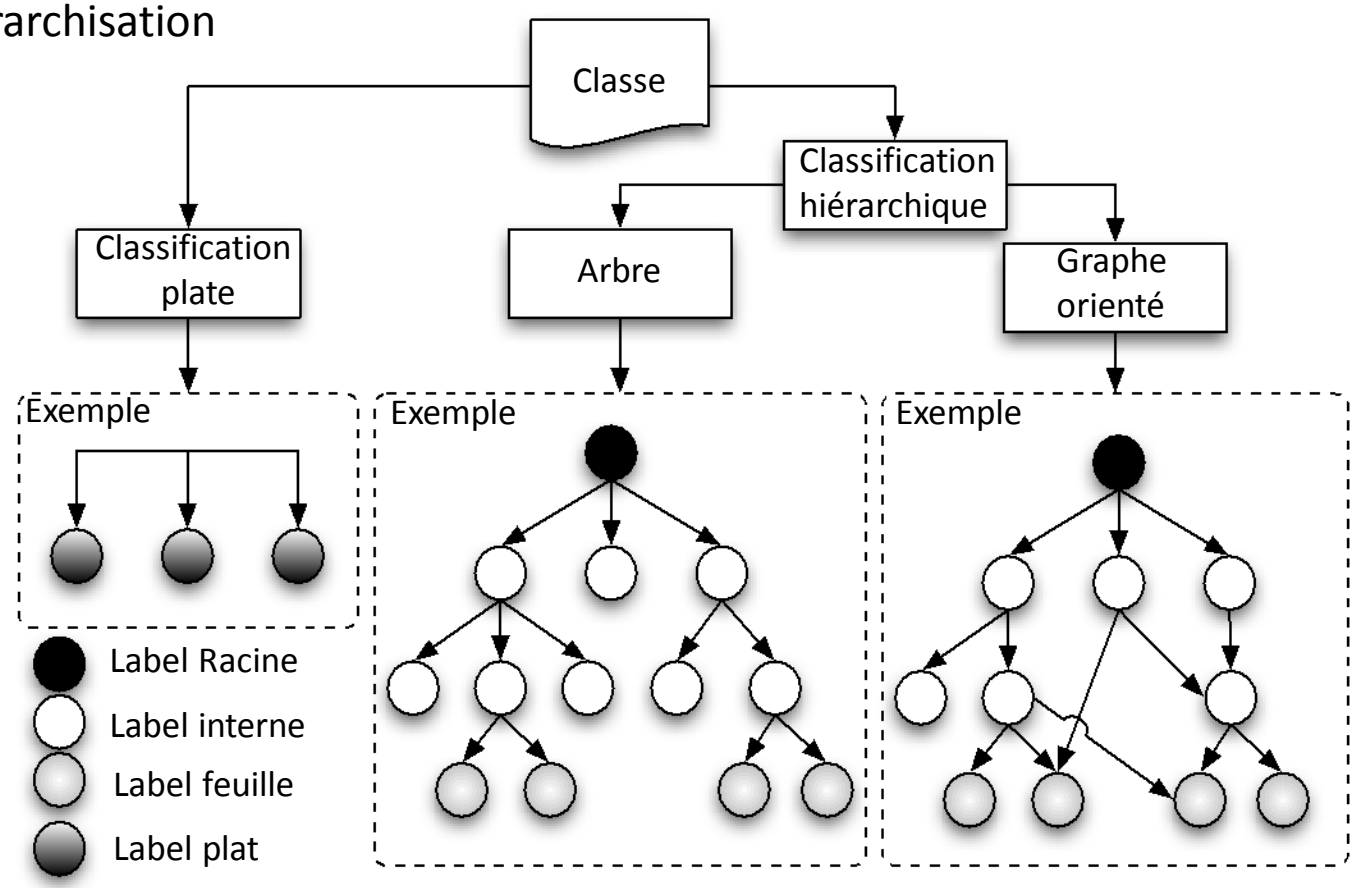
Utilisation de la matrice de fréquence pour générer une hiérarchie de concepts

	t1	t2	t3	t4
A	12	20	17	5
B	8	10	12	15
C	2	7	10	16



Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Hiérarchisation



Tsoumakas, G., Katakis, I. & Vlahavas, I., 2010. Mining multi-label data. In *Data mining and knowledge discovery handbook*. Springer, pp. 667–685.

Cerri, R., Barros, R.C. & de Carvalho, A.C.P.L.F., 2014. Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences*, 80(1), pp.39–56. Available at: http://www.sciencedirect.com/science/article/pii/S0022000013000718.

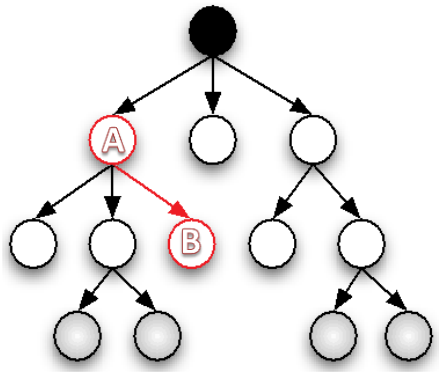
Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Hiérarchisation

Algorithme de subsomption

	t1	t2	t3	t4
A	12	20	17	5
B	8	10	12	15
C	2	7	10	16

Liens de parenté :



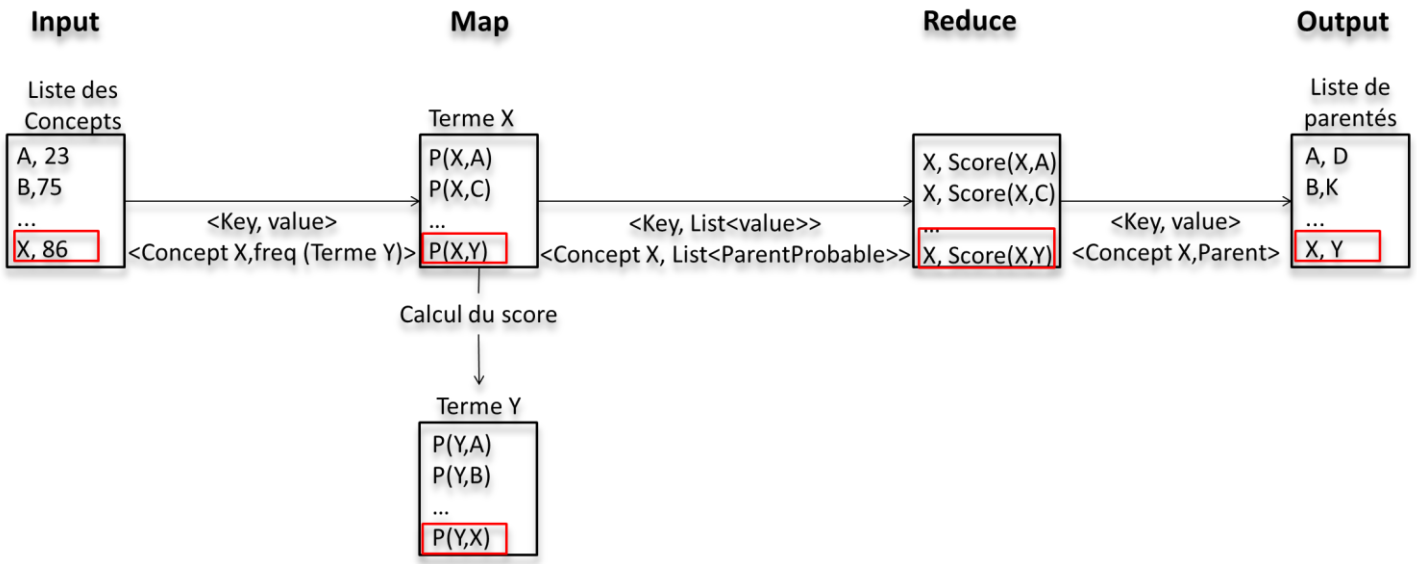
$$P(p|x) \geq t, P(x|p) < t$$

De Knijff, Jeroen and Frasincar, Flavius and Hogenboom, Frederik, 2013. Domain taxonomy learning from text: The subsumption method versus hierarchical clustering, Data & Knowledge Engineering, pp.54-69

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Hiérarchisation

Algorithme de subsomption MapReduce



De Knijff, Jeroen and Frasincar, Flavius and Hogenboom, Frederik, 2013. Domain taxonomy learning from text: The subsumption method versus hierarchical clustering, Data & Knowledge Engineering, pp.54-69

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Résolution

Génération de règles SWRL simples, basées sur la matrice de fréquence :

	t1	t2	t3	t4
A	12	20	17	5
B	8	10	12	15
C	2	7	10	16

Item(?d)^Word(?w1)^hasWord(?d,?w1) ➔ Concept(?d1)

Enrichissement de l'ontologie

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

Résolution

Génération des règles pour un concept est indépendante

➔ Approche « diviser pour régner »

	t1	t2	t3	t4
A	12	20	17	5
B	8	10	12	15
C	2	7	10	16

Distribution des opérations par vecteur

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

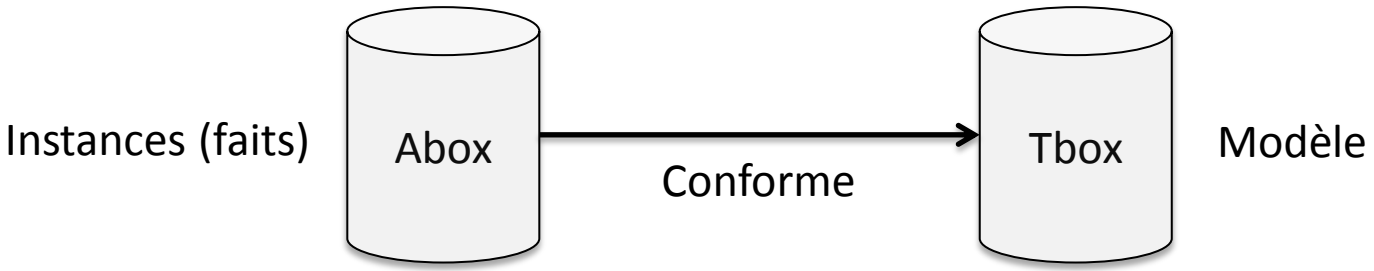
Réalisation

Utilisation des règles SWRL pour classer les nouveaux items

Item(?d)^Word(?w1)^hasWord(?d,?w1) ➔ *Concept(?d1)*

➔ Moteur d'inférence déduit les concepts les plus appropriés pour chaque item.

Assimilation des items en tant qu'individus de l'ontologie (Abox) :



Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

PROCESSUS

Réalisation

Coût calculatoire important : inférence

➔ Utilisation d'un moteur d'inférence basé sur les règles

Approches pour distribuer le raisonnement à l'échelle du Big Data se développent

Chevalier, J. (2013). A Linked Data reasoner in the Cloud. In *The Semantic Web: Semantics and Big Data* (pp. 722-726). Springer Berlin Heidelberg.

Urbani, J., Kotoulas, S., Maassen, J., Van Harmelen, F., & Bal, H. (2012). WebPIE: A Web-scale parallel inference engine using MapReduce. *Web Semantics: Science, Services and Agents on the World Wide Web*, *10*, 59-75.

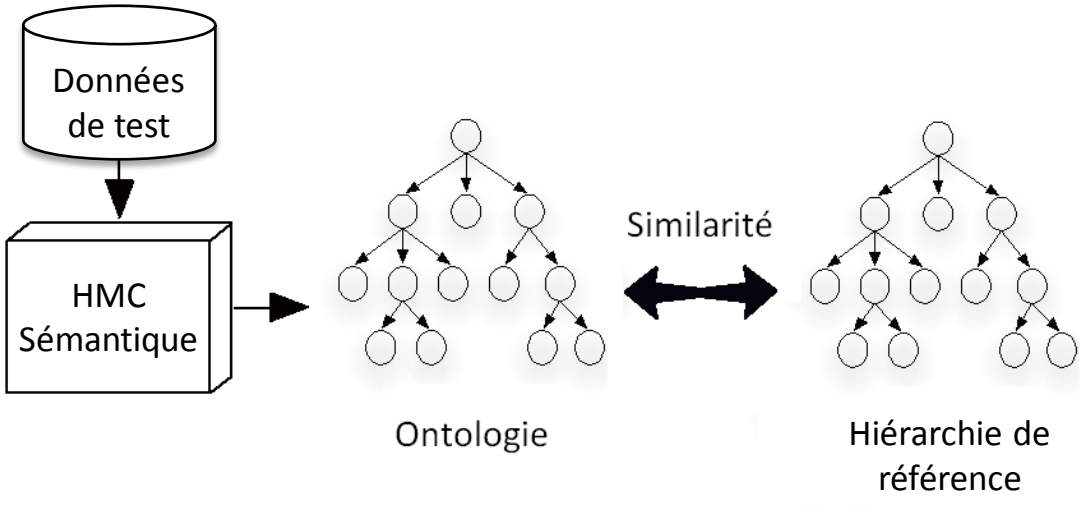
Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

TRAVAIL FUTUR

Evaluations

Qualité de la hiérarchie : utilisation de jeux de données standards

➔ Comparaison entre la hiérarchie générée et la hiérarchie de référence (étalon)



Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

TRAVAIL FUTUR

- Evaluations

Performance du processus

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

TRAVAIL FUTUR

Améliorations

Extraction des concepts :

- Lien avec des bases de connaissance lexicales (type Wordnet)
- Lien avec des bases de connaissances générales
- Lien avec la taxonomie d'entreprise.
- Extraction d'événements complexes basés sur les bases de connaissances.

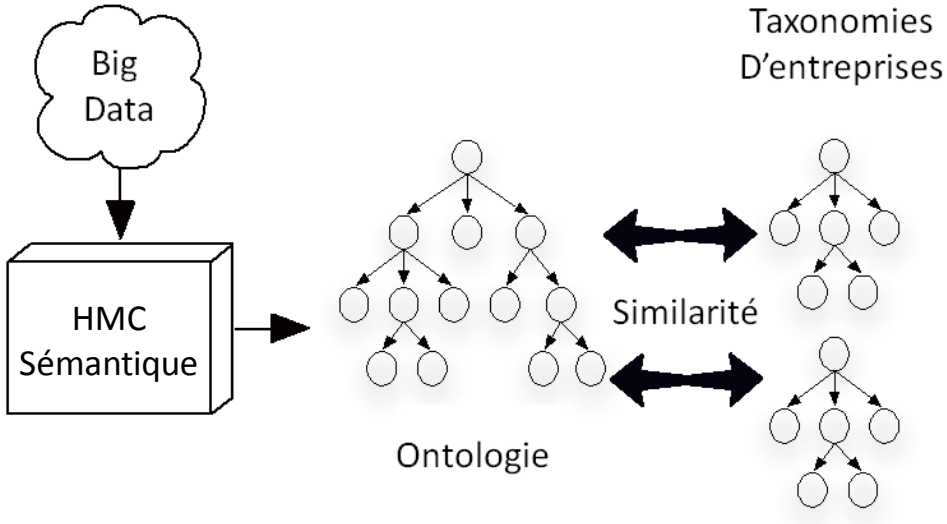
Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

TRAVAIL FUTUR

Améliorations

Rapprochement avec la connaissance métier :

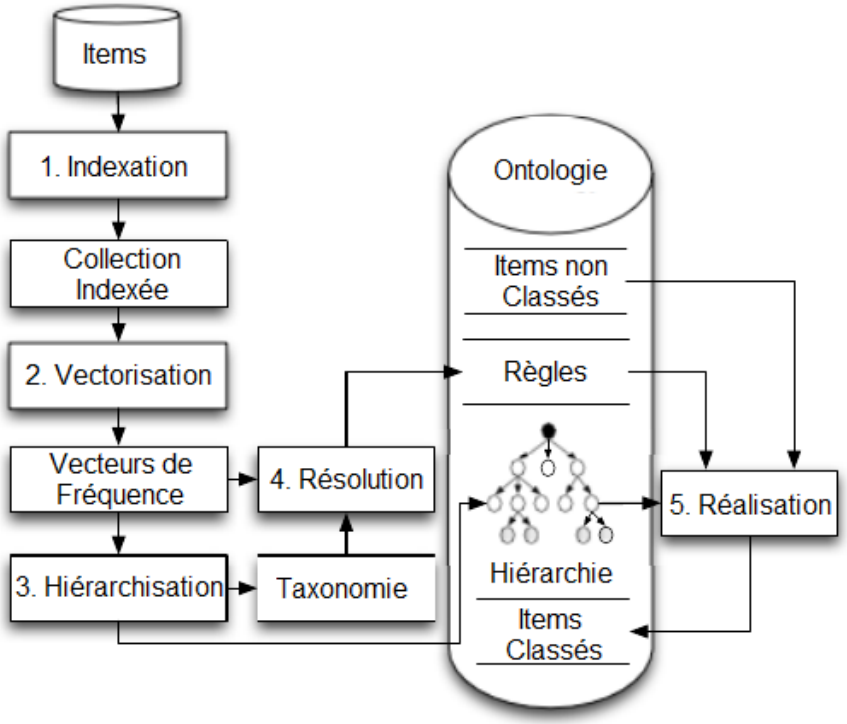
Comparaison entre les concepts de l'ontologie générée et la taxonomie d'entreprise.



Validation des concepts générés par un expert du domaine.

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

CONCLUSION



Processus d'analyse des données générique, adapté à des données et des connaissances variées

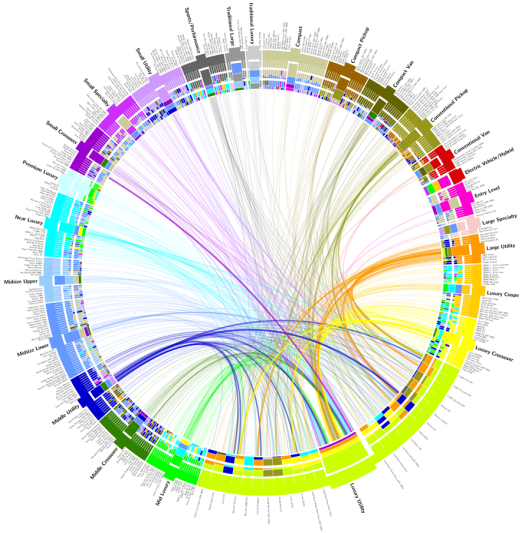
Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

CONCLUSION

A terme :

Création d'outils d'analyse et d'aide à la décision à destination des documentalistes

Recoupement d'informations



Informations déjà parues

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France

CHECKSEM

SEMANTICINTELLIGENCERESEARCH

RESEARCH & BUSINESS APPROACH

Thomas Hassan – thomas.hassan@gmail.com - Equipe de projet Checksem – Laboratoire Electronique Informatique et Image (LE2I – UMR CNRS 6306) IUT Dijon-Auxerre – Université de Bourgogne, BP 47870, 21078 Dijon Cedex, France