Using DL-Reasoner for Hierarchical Multilabel Classification applied to Economical e-News

David WERNER
Universit de Bourgogne
LE2I, CNRS
France
david.werner@u-bourgogne.fr

Nuno SILVA
GECAD and School of
Engineering Polytechnic of Porto
Portugal
nps@isep.ipp.pt

Christophe CRUZ
and Aurelie BERTAUX
Universit de Bourgogne
LE2I, CNRS
France
{christophe.cruz, aurelie.bertaux}@u-bourgogne.fr

Abstract—**This work is part of a global project to develop a recommender system of economic news articles. Its objectives are threefold: (i) automatically multi-classify the economic new articles, (ii) recommend the articles by comparing the profiles of the users and the multi-classification of the articles, and (iii) managing the vocabulary of the economic news domain to improve the system based on the seamlessly intervention of the documentalists. In this paper we focus on the automatic multi-classification of the articles and the respective description and justification to the documentalists. While several multi-classification solutions exist they are not automatically adaptable to the problem in hands as their description of the resulting multi-classification lacks substantial correlation with the documentalists perspective. In fact, we need to consider not only the automatic classification but also the supervision of the classification and its evolution based on the documentalists supervision of the automatic classification. Accordingly, it is necessary to provide a mechanism that bridges the gap between the automatic classification mechanisms and the documentalists thesaurus, in order to support their seamless supervision of classification and of thesaurus management. Ontologies are central to our proposal, as they are used to represent and manage the thesaurus, to describe the content of the articles, and finally to automatically multi-classify them via inference process. Also, we adopt a machine learning approach for generating a prediction model for supporting the automatic classification. This paper presents a proposal for enriching the documentalist-oriented ontology with the model prediction rules, which provides the necessary capabilities to the DL reasoner for automatic multi-classification.**

Keywords—multi-classify, recommender system, ontology economical e-news, machine learningmulti-classify, recommender system, ontology economical e-news, machine learning.

I. Introduction

The decision-making process in the economic field requires the centralization and intake of a large amount of information. The aim is to keep abreast with current market trends. Thus, contractors, businessmen and salespersons need to continuously be aware of the market conditions. This means to be up-to-date regarding ongoing information and projects undergoing development. With the help of economic monitoring, prospects can be easily identified, so as to establish new contracts. Our tool is specialized in the production and distribution of press reviews about French regional economic actors. The overload of news information is a particular case of information overload, which is a well-known problem, studied by Information Retrieval and Recommender Systems research fields. News recommender systems already exist [1], Athena [2], GroupLens [3] or News Dude [4]. Some of these systems use domain knowledge to improve the recommendation task [1] [2]. To achieve this goal, a content-based recommender system is being developed. A recommender system is necessary for the item ranking and a content-based approach is required to analyze the content of each article to structure and preserve information content. The results of the analysis enable linking the domain knowledge to the articles to improve the recommendation task [1] [2]. Content based recommender systems typically follow a two-step process:

(i) *the indexing of articles (also known as* classification*) and users (also known as* profiling*)*

(ii) *the comparison process which consists to compare the classification of the articles and the profiles of the users. The latter computes the article relevance with regards to the user profile.*

In order to capture this economical context, we are moving towards a customized review for each user and towards an opinion survey on magazine readers that cover a broad array of subjects, including news services. Criteria for a relevant customization of the review were identified as a result of this survey as well as expert domain knowledge. These criteria are economic themes (i.e. main economic events), economic sectors, temporal and localization information. Our domain knowledge is based on expert-defined thesaurus and an ontology allowing managing the main concepts and relations.
As consequence of this effort the complete production process of the review was redesigned to produce and to automatically distribute a customized review for each user. So, the aim of the overall system is to manage all news articles produced, and provide the most relevant article for each customer. Yet, while the comparison process is automated, the article classification process remains manual. While this is a time consuming process, it ensures the quality of the recommendation as long as the documentalists know the economical context and perform the classification. The aim of this paper twofold:

(i) *proposing an automatic classification processes that mimics the documentalists process*

(ii) *provide to the documentalists the description and justifications of the automatic classification to support their supervision and feedback in a seamless way.*

Consequently, time is saved and the economical context evolution is kept in the business process.
The paper is organized as follows. First, we present the background research work and related work. Then, we present the overall system proposal, by describing the core definitions, principles and decisions. The fourth section presents the ontology enrichment from the prediction model, as well as the ontology population and automatic classification processes. Then we present experimentations

and we summarize the contributions and point out future research directions.

II. State of the art and our hierarchical multi-label classification

This section presents the background research context and from there describes the related work. The large amount of information on the web, company information systems, Digital Libraries, Selling websites and so on, is a well-known fact. The recommender systems aims at providing for each user the better items according to his/her needs. Items can be websites, news articles, books, video, music, washing machine, etc. In the recommender system literature two paradigms are distinguished. First, content-based recommender systems try to recommend items similar to those a given user has liked in the past. Second, Collaborative filtering recommender systems identify users whose preferences are similar to those of the given user and recommend items they have liked [5]. Some subtasks should be performed and the first is named the indexing task. It is possible to distinguish two cases, the indexing of items, by content analysis, and the indexing of the profiles, via profiles learning (which generally includes a study of the behavior for implicit feedbacks or proposed to the user to giving explicit feedbacks as I like button). Both can be seen as a multi-classification task. The second task, comparison, consists in filtering each item relative to a given profile [6]. Usually, both types of recommender systems are combined to define a so-called Hybrid recommender systems and to overcome drawbacks of each case. The survey of K. Nagewara Rao [7] proposed a general comparison of the main advantages and drawbacks of each kind of Recommender System (e.g. content based or collaborative filtering). In this paper we focus on content-based recommender system, and on the indexing subtask of textual items, based on a controlled vocabulary.
The indexing (or multi-classification) task consists in selecting a set of keyword from the controlled vocabulary and associating it with an item (content analysis) or a user profile (profiling).
As it is presented in the survey about controlled vocabularies from [8], more and more companies use controlled vocabularies in their information system. Different kinds of structure are used to manage vocabularies, i.e. from the lowest to the richest semantic definition: glossary, taxonomy, thesaurus, ontology. A lot of companies plan to use ontologies in their applications [8]. While acquiring, managing and maintaining controlled vocabularies are important yet relatively easy tasks for the documentalist, the ontology approach to model the domain knowledge is hardly accessible to the documentalist due to the complexity of the logical structure. Actually, this model differs from his/her habits, which is much more related to a taxonomy (subsumption relations between entities) or a thesaurus (different relations between entities, notably words). Through the remaining of the text we will use the terms documentalist and expert interchangeably.
An hybrid domain knowledge representation has been adopted, in which the skeleton of the domain knowledge is delivered by an ontology that captures the process and application needs, complemented by a set of thesaurus that capture the domain knowledge of the experts. The thesaurus has been captured in a very light, expert-oriented fashion, with minimal formal semantics and consistency obligations. For that, the experts have been provided with only some types of relations as defined in SKOS [9]. On the contrary, for the authoring of this ontology (Fig. 1) the strict semantics of OWL DL has been followed.

Each of the resulting thesaurus were appended to the Theme and Sector classes, by interpreting (i) each label as a class, and (ii) the skos:broader relationships as owl:subClassOf relationships, thus forming a taxonomy (hierarchy of subsumption relationships) of labels (i.e. a taxonomic thesaurus). Here is a partial definition in DL syntax:

- $EconomicEntity \sqsubseteq Thing$
- $Location \sqsubseteq Thing$
- $Theme \sqsubseteq EconomicEntity$
- $Organisation \sqsubseteq EconomicEntity$
- $Theme \sqsubseteq EconomicEntity$
- $Sector \sqsubseteq EconomicEntity$
- $City \sqsubseteq Location$
- $FirstOD \sqsubseteq Location$***, i.e. it is the First Order Division, in France.***
- ***...***
- $Profile \sqsubseteq Thing \sqcap \exists isInteresedIn.Location \sqcap \exists isInterestedIn.EconomicEntity$
- $Article \sqsubseteq Thing \sqcap \exists isAbout.Location \sqcap \exists isAbout.EconomicEntity$
- $Offshoring \sqsubseteq International \sqsubseteq Expanding \sqsubseteq Theme$
- $Pneumatic \sqsubseteq Automobile \sqsubseteq Transport \sqsubseteq Sector$
- ***...***

The other relations used in the thesaurus were copied unchanged to the ontology to be used solely in the recommendation process as input for the evaluation of the semantic distance between articles and profiles. Furthermore, despite these thesaurus relations have no ontology-based semantic and inference usefulness, they are very important for the expertss tasks as allow him/her to better and faster understand the context of the word/label to use in an article or profile classification. As consequence, there is no semantic constraints associated with any of the terms of the thesaurus that promotes rich ontology-based inference. The comparison of the item classification with the user profile is performed using the classification of both according to the thesaurus (Fig. 2). More details about the comparison process can be found in the [10].

In order to make the classification process automatic, the system has to associate a set of labels from the taxonomic thesaurus to each article. This cannot be done without the two following process: (i) a text analysis process to extract keyword and other features from texts and (ii) a machine learning process to learn the classification process from examples. Feature extraction processes ranges from simple term extraction process like tf-idf [11] to text-based semantic-aware processes, e.g. term extraction from text based on (i) information retrieval methods [12], or (ii) based on NLP works [13] [14]. It is possible to use different degree of text processing tools (and preprocessing), to extract noun phrase (i.e. tokenizer, part of speech tagger, handmade patterns and even parsers). This process allows us to interlink a set of terms (named features in section 3) from the article term extraction process (content analysis) and the set of taxonomic keywords (named label in section 3) for the articles indexing task.
Machine learning process has two primary goals: prediction and description. Prediction is concerned with using features of previously classified examples (e.g. documents or any other resources that can be analyzed and classified) to predict the unknown classification (i.e. labels). Description on the other hand focuses on finding human-interpretable patterns that describe the performed classification.
Two main categories of label-classification prediction can be enumerated: the single-label and the multi-label classification. Single-label classification aims to learn a prediction model from a set of examples that are related with a single label from a set of disjoint labels. In multi-label classification instead, the examples are associated with a set of labels [15]. Multi-label classification faced increased attention in the last decade, overcoming the single-label

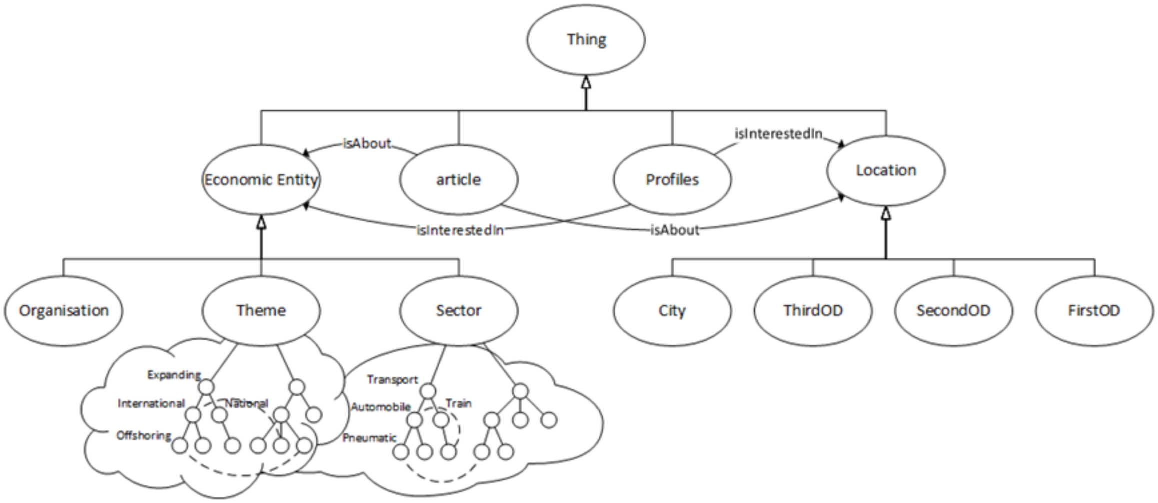


Fig. 1: Domain ontology with the taxonomic thesaurus.

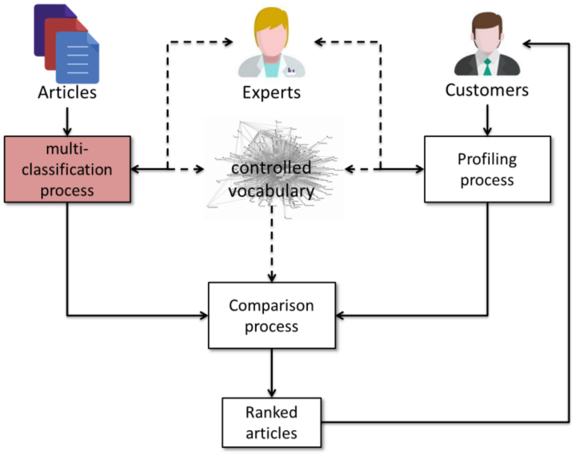


Fig. 2: Our ontology-based recommender system.

classification previous dominance, but it was only much recently that hierarchical multi-label classification (HMC) approaches received the desired attention. Even so, some of the so-called HMC approaches do not follow a strict hierarchical semantics (in the sense of subsumption), but a clustering approach. This is the case of the state of the art "hierarchical" multi-label approach HOMER [16] and that of [17] that uses Predictive Clustering Tree (PCT) framework. However, unlike HOMER, the approach described in [17] is constrained by the taxonomy or DAG underlying the training and testing datasets. This is also the case of other works, notably in the area of bioinformatics [18] [19].

Both [17] and [18] are very interested in the description of the predictions to the user. In [20] the authors propose an iterative and interactive (between AI methods and domain experts) approach to achieve prediction and description (which are usually hard to fulfill), considering domain expert knowledge and feedback. Unlike us however, in [20] the authors do not aim to automatically multi-classify the items but only to improve the ontology, which means that the resulting ontology is not used for automatic classification of items.

In [23] the authors propose an approach to build ontologies using data mining results upon databases. The result is the enrichment of the ontology with new concepts and datatype properties, which is far from the required specification of classes. Our goal instead is to enrich the already existent taxonomic thesaurus with constraints capturing the prediction model allowing the user to perceive the taxonomic thesaurus, rules and the adopted features.

In [21] the authors are concerned with automatically create an ontology from the text documents without any prior knowledge about their content. For that they use an iterative and interactive 4-phases process. Unlike [22], [23] that construct thesaurus from the learning examples, in this project/paper the thesaurus-based taxonomy already exists and should be applied both in the automatic classification and description.

This paper does not aim at improving the state of the art in multi-classification, nor in ontology learning from text, but instead to propose a method to semantically enrich the ontology by adopting machine learning processes in order to both classify and describe classification, so the gap between the experts perspective and the classification rules representation is reduced.

Our method is based on four following steps.

1) *The* **vectorization** *phase allows generating the matrix of term frequencies from a learning set.*
2) *The* **resolution** *allows creating logical constraints (rules) associated to the keywords of the taxonomy (controlled vocabulary) using named entities extracted in the vectorization phase. This phase generates a flat ontology.*
3) *The* **hierarchization** *allows to generate a class hierarchy of subsumption of the ontology used to label documents.*
4) *The realization allows searching and deducing the most specific classes of documents to be classified which consists in generating the multi-classification.*

Phases 3 and 4 are done using standard reasoner such as FaCT++, HermiT, Pellet.

III. Definitions

In this section we give some definitions necessary to understand notions used by our system. First we present fundamental definitions and then we define the four categories of documents.

A. Fundamental definitions

Definition 1 : Let W_j be a named entity belonging to the set of relevant words extracted from learning thesaurus.

Definition 2 : ***Let Tax_i be the name of a classe of the taxonomy.***

Definition 3 : ***The frequency of occurrence TF_{ij} of a term is the percentage of occurrence of W_j for documents labelled with Tax_i label.***

Definition 4 : ***The threshold α is such $\alpha < TF_{ij}$***

Definition 5 : ***Let* alpha terms set *containing $\omega_\alpha^{Tax_i}$ terms, the set of W_j terms having TF_{ij} greater than α threshold for the term Tax_i of the taxonomy.***

$$\omega_\alpha^{Tax_i} = \left\{ \bigcup_j \{W_j\} \mid TF_{ij} > \alpha \right\}$$

Definition 6 : ***The threshold β is such $\beta \leq TF_{ij} \leq \alpha$***

Definition 7 : ***Let* beta terms set *containing $\omega_\beta^{Tax_i}$ terms, the set of W_j terms having TF_{ij} greater or equal to β threshold and lower or equal to α threshold for the term Tax_i of the taxonomy.***

$$\omega_\beta^{Tax_i} = \left\{ \bigcup_j \{W_j\} \mid \beta \leq TF_{ij} \leq \alpha \right\}$$

Definition 8 : ***Let $\overrightarrow{d}$ be a boolean vector such each component of the vector corresponds to the occurring or not of term W_j into the document.***

$$\overrightarrow{d} = (a_1, ..., a_m) \mid m\, is\, the\, number\, of\, terms\, W_j\, and\, a_i \in \{0, 1\}$$

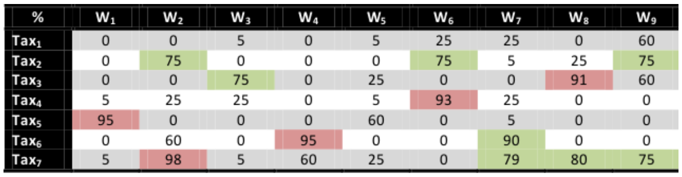
Definition 9 : ***Let* alpha documents set *$D_\alpha^{Tax_i}$ of a document $\overrightarrow{d}$ be the set of W_j terms having TF_{ij} greater than α threshold for the term Tax_i of the taxonomy such as a_i component of $\overrightarrow{d}$ is not null.***

$$D_\alpha^{Tax_i} = \left\{ \bigcup_j \{W_j\} \mid TF_{ij} > \alpha \quad and\; a_i \neq 0 \right\}$$

Definition 10 : ***Let* beta documents set *$D_\beta^{Tax_i}$ of a document $\overrightarrow{d}$ be the set of W_j terms having TF_{ij} greater than β threshold and lower or equal to α threshold for the term Tax_i of the taxonomy such as a_i component of $\overrightarrow{d}$ is not null.***

$$D_\beta^{Tax_i} = \left\{ \bigcup_j \{W_j\} \mid \beta \leq TF_{ij} \leq \alpha \quad and\; a_i \neq 0 \right\}$$

Table 3 presents examples of term frequencies into documents and their belonging to alpha and beta terms sets of the taxonomy. Red cells represent alpha terms set and green cells represent beta terms set.



%	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9
Tax_1	0	0	5	0	5	25	25	0	60
Tax_2	0	75	0	0	0	75	5	25	75
Tax_3	0	0	75	0	25	0	0	91	60
Tax_4	5	25	25	0	5	93	25	0	0
Tax_5	95	0	0	0	60	0	5	0	0
Tax_6	0	60	0	95	0	0	90	0	0
Tax_7	5	98	5	60	25	0	79	80	75

Fig. 3: W_j belonging to the set $\alpha = 90$ are in red, and those belonging to the set $\beta = 75$ are in green.

B. Categories of documents

Taking into account the two α and β thresholds we are considering 4 categories of documents :

1) ***alpha set is not empty and beta set is empty such as :***

$$|\omega_\alpha^{Tax_i}| > 0 \wedge |\omega_\beta^{Tax_i}| = 0$$

2) ***alpha set is empty and beta set is not empty such as :***

$$|\omega_\alpha^{Tax_i}| = 0 \wedge |\omega_\beta^{Tax_i}| > 0$$

3) ***alpha and beta sets are not empty such as :***

$$|\omega_\alpha^{Tax_i}| > 0 \wedge |\omega_\beta^{Tax_i}| > 0$$

4) ***alpha and beta sets are empty such as :***

$$|\omega_\alpha^{Tax_i}| = 0 \wedge |\omega_\beta^{Tax_i}| = 0$$

IV. TRANSLATION RULES FOR ONTOLOGIES

This section describes how the alpha and beta sets of each term of the taxonomy are translated into logical constraints in an ontology classification like Description Logic. First, it is necessary to define the core of the ontology by using primitive concepts and defined concepts as explained in Section IV-A. This starting ontology will be enriched with the rules as defined in Section IV-B. With this enhancement, the inference engine is capable to deduce from this TBox the classification of the most specific subsumers for concepts defined (Section IV-B1). On an other hand, when the ontology is populated with the new documents to be labeled by adding them to assertional level or ABox, the inference engine will be able to find the most specific classes of documents. This phase is described in Section IV-B2.

A. Defining the core of the ontology

We need to define three primitive concepts and a relationship. These concepts are the* Word *concept to define a term appearing in the documents(W_j), the* Doc *concept to define a document to be labeled ($\overrightarrow{d}$ vector)and the* Tax *concept to define the terms of the taxonomy used to label documents (Tax_i). Only one role is necessary in our core ontology: the role* hasWord *which can link a document to all of its terms.

- $Word \sqsubseteq \top$
- $Doc \sqsubseteq \top$
- $Tax \sqsubseteq \top$
- $Doc \equiv \exists hasWord.Word$

- $Tax_1 \sqsubseteq Tax$
- $Tax_2 \sqsubseteq Tax$
- ...
- $Tax_n \sqsubseteq Tax$

B. Defining enrichment rules

These rules allow the enrichment of the ontology from the matrix of term frequencies.
For this, we consider only three categories to simplify the problem and reduce the size of logical constraints. These cases are: **Beta is empty, Alpha is empty** ***and*** **both Alpha and Beta are non-empty.**

1) Empty beta category: ***In this case, only alpha set is considered:***

$$Tax_i \equiv \bigcap_{j=1}^{n} hasWord : W_j,\;\; W_j \in \omega_\alpha^{Tax_i}$$

In order to label a document with Tax_i ***term of the taxonomy, the document has to possess at least one of the terms of the alpha set*** $\omega_\alpha^{Tax_i}$.

2) Empty alpha category: ***In this case, only beta set is considered:***

$$Tax_i \equiv \left(\bigcup_{j=1}^{n} hasWord : W_j\right) \sqcap \geq \delta hasWord.Word$$

with $W_j \in \omega_\beta^{Tax_i}$, $\delta = \lceil |\omega_\beta^{Tax_i}| * p \rceil$, ***and*** $0 \leq p \leq 0.5$. ***In order to label a document with*** Tax_i ***term of the taxonomy, the document has to possess a number of the terms of the beta set*** $\omega_\beta^{Tax_i}$ ***greater than*** δ. ***This value is calculated based on a percentage based on p value. For example, if we want a set of terms at least equal to 30% of the terms*** $\omega_\beta^{Tax_i}$ ***and that the cardinality of*** $\omega_\beta^{Tax_i}$ ***equals 9, then*** $\delta = 3$.

3) Non-empty alpha and beta category: ***On one hand, we consider alpha set as defined in Section IV-B1, and on the other hand we consider the beta set as defined in Section IV-B2, but with a value q = p*2. It corresponds to*** $\delta = \lceil |\omega_\beta^{Tax_i}| * q \rceil$, ***with*** $0 \leq q \leq 1$ ***and*** $q = p * 2$.

In order to label a document with the term Tax_i ***of the taxonomy, the document must have a number of terms in the beta set*** $\omega_\beta^{Tax_i}$ ***greater than*** δ. ***This value is calculated based on a percentage defined by the p value. For example, if we want a set of terms at least equal to 60% of the terms of*** $\omega_\beta^{Tax_i}$ ***and that its cardinality equals 7 then*** $\delta = 4$.

C. Ontology populating rules definition

As written in definition 8, a document corresponds to a vector of terms such as:

$$\overrightarrow{d} = (a_1, ..., a_m) \mid m\, is\, the\, number\, of\, terms\, W_j\, and\, a_i \in \{0,1\}$$

For every vector, we define a set of assertions, a single concept assertion and set of role assertions for which the component of the vector is not null: $hasWord(d, W_j) | a_i \neq 0$.

D. Inference process

The inference engine processes can perform the following two phases of our method. It allows firstly to make the **classification** ***phase, and secondly to perform the*** **realization** ***phase.***

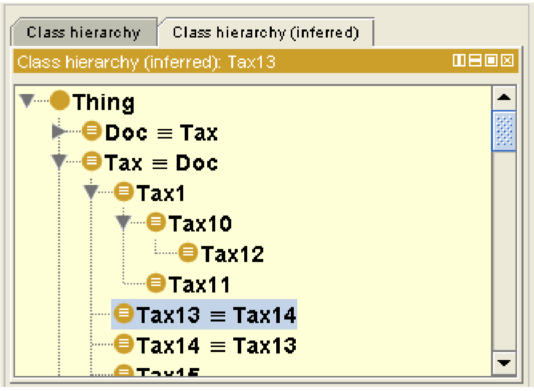


Fig. 4: Inferred class hierarchy showing inferred subsumptions and equivalences.

1) Classification phase: ***The classification provides two types of results as illustrated in Fig 4. The first is the discovery of the most specific subsuming class such as*** Tax_{10} ***subsumes*** Tax_{12} ***and*** Tax_1 ***subsumes*** Tax_{10}. ***The second allows to infer equivalence classes when the logical constraints are equivalent, for example:*** Tax_{13} ***is equivalent to*** Tax_{14}. ***On one hand, this means that when a document is labeled with a class that has subsumers, this document will also be labeled by subsumant classes. On the other hand, when a document is labeled with a class that has the equivalence classes then this document is also labeled with equivalent classes.***
These two elements can achieve a multi-labeling, knowing that the terms of the taxonomy are hierarchical. Accordingly, this is a hierarchical multi-label classification (HMC) process.

It should be noticed that for purposes of optimization, we define a document as an instance of a term of the taxonomy even if a conceptualization point a view is wrong. However in this case we do not focus on the definition of business knowledge, but on the definition of a system with logical constraints in order to obtain answers to questions. The formal settings of description logic ensures that the reasoning problem is decidable and calculable.

2) Realization phase: ***The realization phase consists in finding all the most specific classes of individuals.***
This phase is carried out by the inference engine which enables to deduce all the more specific classes. It also allows to manage multi-labeling while adding subsuming and equivalent classes. As consequence, a document is multi-labeled according to a hierarchy. Fig. 5 presents the results of the realization phase: the document doc1 belongs to the class Tax_{12} ***and may be labeled with terms*** Tax_{12}, Tax_{10} ***and*** Tax_1.

V. SETTINGS

This section describes how to automatically determine the values of alpha and beta thresholds according to the desired number of terms in the logical constraints.
These two values are important because the size of the rules depends directly on them. On the one hand, if the threshold values are too high, then the logic constraints are too large involving failure to realize the phases of the classification and realization due to the polynomial complexity. On the other hand, if the value is too small, so most of the time alpha and beta sets may be empty involving the incapacity of the system to achieve its objective.

Definition 11 : ***Consider the sum of the cardinalities of alpha***

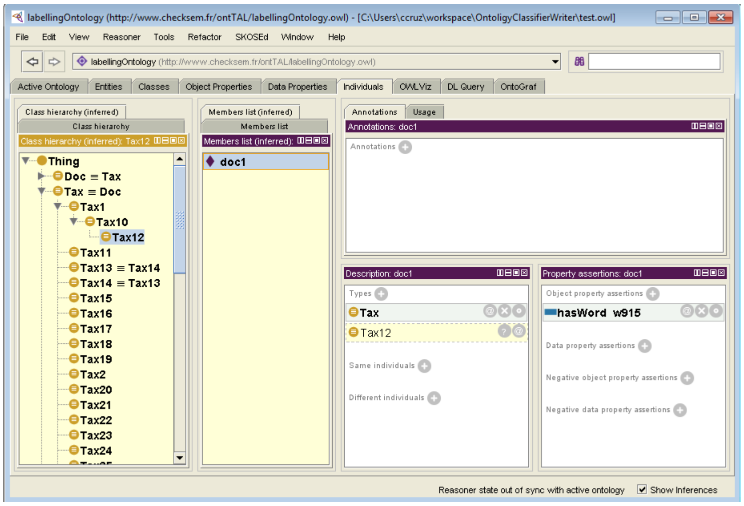


Fig. 5: Results of the realization phase.

Dataset	Exemples		Attributes		Label		
	Train	Test	Numeric	Nominal	Count	Cardinality	Density
delicious	12920	3185	0	500	983	19.020	0.019

TABLE I: The dataset in some numbers.

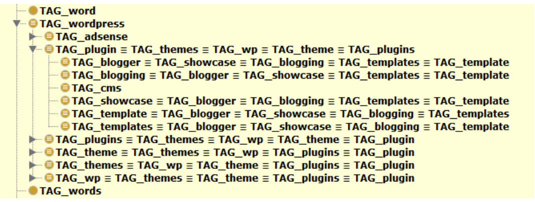

TAG_word
TAG_wordpress
TAG_adsense
TAG_plugin ≡ TAG_themes ≡ TAG_wp ≡ TAG_theme ≡ TAG_plugins
TAG_blogger ≡ TAG_showcase ≡ TAG_blogging ≡ TAG_templates ≡ TAG_template
TAG_blogging ≡ TAG_blogger ≡ TAG_showcase ≡ TAG_templates ≡ TAG_template
TAG_cms
TAG_showcase ≡ TAG_blogger ≡ TAG_blogging ≡ TAG_templates ≡ TAG_template
TAG_template ≡ TAG_blogger ≡ TAG_showcase ≡ TAG_blogging ≡ TAG_templates
TAG_templates ≡ TAG_blogger ≡ TAG_showcase ≡ TAG_blogging ≡ TAG_template
TAG_plugins ≡ TAG_themes ≡ TAG_wp ≡ TAG_theme ≡ TAG_plugin
TAG_theme ≡ TAG_themes ≡ TAG_wp ≡ TAG_plugins ≡ TAG_plugin
TAG_themes ≡ TAG_wp ≡ TAG_theme ≡ TAG_plugins ≡ TAG_plugin
TAG_wp ≡ TAG_themes ≡ TAG_theme ≡ TAG_plugins ≡ TAG_plugin
TAG_words

Fig. 6: Results of the realization phase, with the delicious dataset.

and beta sets defined as follow:

$$Sum(\alpha) = \sum_{i=1}^{n} |\omega_{\alpha}^{Tax_i}|$$

$$Sum(\beta) = \sum_{i=1}^{n} |\omega_{\beta}^{Tax_i}|$$

With n the number of terms in the taxonomy.

Definition 12 : Consider the average of the cardinalities of alpha and beta sets defined as follow:

$$Avg(\alpha) = \lceil \frac{Sum(\alpha)}{n} \rceil$$

$$Avg(\beta) = \lceil \frac{Sum(\beta)}{n} \rceil$$

With n the number of terms in the taxonomy.

Definition 13 : Consider the maximum value of the cardinalities of alpha and beta sets defined as follow:

$$Max(\alpha) = \arg_i max |\omega_{\alpha}^{Tax_i}|$$

$$Max(\beta) = \arg_i max |\omega_{\beta}^{Tax_i}|$$

With n the number of terms in the taxonomy.

It is possible to take into account the maximum number of terms in a logical rule, or its average. This by using functions $Max(\alpha)$ and $Max(\beta)$ functions and $Avg(\alpha)$ and $Avg(\beta)$ respectively. In the following algorithm, only the instance with the maximum value is shown. The algorithm for calculating the average value is to override the function by function Avg.
The following algorithm determines the values $\alpha$and β for a given Objective Value. This value is set at the beginning of the algorithm and corresponds to the maximum number of terms that logical rules must contain:

```
$\varepsilon = 0.25$
$\alpha = 0.5$
While ( Objective != $Max(\alpha)$)
    if $Objective > Max(\alpha)$ then $\alpha = \alpha - \varepsilon$
    if $Objective < Max(\alpha)$ then $\alpha = \alpha + \varepsilon$
    $\varepsilon = \varepsilon/2$
End While
$Result.Objective \rightarrow \alpha$

$\varepsilon = 0.25$
$\beta = 0.5$
While ( Objective != $Max(\beta)$)
    if $Objective > Max(\beta)$ then $\beta = \beta - \varepsilon$
    if $Objective < Max(\beta)$ then $\beta = \beta + \varepsilon$
    $\varepsilon = \varepsilon/2$
End While
$Result.Objective \rightarrow \beta$
```

$$Result.Objective : (\alpha, \beta)$$

This algorithm determines the optimum values for α and β for an objective value. For our prime evaluations exposed here, we used an objective average of 10 terms in all rules.

VI. First Evaluation of the Approach

In this section we present a preliminary evaluation of the approach. Due to the lack of real data for our platform, the first evaluation is based on the delicious dataset available on the Mulan project web site and already used in some multilabel-classification works [13]. It was extracted from the del.icio.us social bookmarking site on the 1st of April 2007 and contains textual features and tags of webpages. This dataset is used to train a classifier for tag recommendation.

With this dataset the (phase 1) manual multi-classification and the (phase 2) feature extraction tasks are not necessary: features and tags are already associated with documents and a sub-dataset is predefined for the (phase 3) learning of the prediction model. Our prediction model is the set of α and β -rules generated by the method outlined in section III. The ontology is populated (phase 4), and some reasoners are used to perform the multi-classification task. During our evaluations different reasoners are used on different hardware (table 2). The results produced by the reasoner are not only a multilabel-classification of documents, but also a hierarchical reorganization of tags based on the equivalence rules.

α-rules	FaCT++	HermiT	Pellet
i7 4Go DDR3	50 s	n.e.m.[1]	n.e.m.
Xeon E3 24Go DDR3	-	8 h	18 h
$\alpha\beta$-rules	**FaCT++**	**HermiT**	**Pellet**
i7 4Go DDR3	n.e.m.	n.e.m.	n.e.m.
Xeon E3 24Go DDR3	n.e.m.	out[2]	out
Xeon E5 128Go DDR3	2 h / out[3]	out	out

TABLE II: Reasoner time computation comparison for the ontology populated with $\alpha\beta$-complex rules.

Evaluation	Precision	Recall	F1-Mesure
Proposal with A-type rules	30%	6%	10%
HOMER [16]	-	-	25%

TABLE III: Evaluation and comparison with a similar work of multi-classification.

Table III shows that the second type of rules is much more time and memory consuming. We have only one result to show. This result was produced by FaCT++, with the best machine and an ontology without any instance (i.e. document). In 2 hours the reasoner infers a hierarchical reorganization of tags based on the equivalence rules. Yet the ontology populated with documents and equivalent class rules seems very time consuming even for FaCT++. The ontology with B-type rules is not evaluated in the following steps due to the lack of results provided by reasoners.

Quality of the results are low and another approach [16] with this dataset also shows low value for the F-measure.

This precision-recall evaluation is only based on α rules, because of the difficulty for reasoners to provide results with β rules.

Our way to create rules (i.e. with an average of 10 terms for all rules) has the consequence the creation of some ruleless classes. With this method, for our 983 classes, only 427 have rules. There are 556 classes without labeling rule (obviously, theses classes should have had β rules). So there are classes that the predictive model can not affect. This impacts very negatively the Recall.

In our α rules, we consider that the presence of one of the selected terms for the rule is a sufficient clue. In fact, the terms selected for theses rules are in the majority of cases not enough frequent to be a sufficient evidence to qualify the class. So, the presence of only one of them is not enough for the decision making. It affects very negatively the Precision. The solution could be β rules. They allow to take a decision, based on a minimum of clues. β-rules are a small step to gain intelligence, but the impact on the computation time and memory used is very important.

The realization phase 6 shows interesting results as the semantic proximity between tags blogger, blogging, wp, wordpress is detected, for examples, but results that are not described further here..

VII. Discussion and Conclusion

This paper describes the process of using an HMC approach to enrich an already existing ontology to be used for automatic multi-classification of economic news articles. We decided to capture the prediction model into the taxonomic thesaurus part of the ontology, thus transforming it into a more semantically rich ontology. Based on the early experiments, it was observed that the logical axioms/rules suggested the existence of several subsumption relations that were not present in the taxonomic thesaurus, giving rise to Direct Acyclic Graphs, i.e. a class can have more than one super-class. While this observation is potentially relevant for the refinement of the taxonomic thesaurus and therefore for the classification, a deeper and finer analysis and expert-based experiments have to be performed to better understand the advantages, disadvantages and potential applications. Moreover, our preliminary tests have highlighted the complexity of reasoning on ontology, even with a relatively small ontology. As future work, two things are interesting to study in the context of this work. (i) The team will focus deeper on the supervision process, namely in reviewing results of the documentations, and managing it by interpreting it as feedback to the classification process and evolution of the taxonomy. (ii) The team will study solutions to distribute and split the reasoning task problem with the map-reduce approach.

Acknowledgment

This project is founded by the company Actualis SARL and the French agency ANRT.

References

[1] *S. E. Middleton, N. R. Shadbolt, and D. C. De Roure, "Ontological user profiling in recommender systems,"* ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 54–88, Jan. 2004. [Online]. Available: http://doi.acm.org/10.1145/963770.963773*

[2] *W. IJntema, F. Goossen, F. Frasincar, and F. Hogenboom, "Ontology-based news recommendation," in* Proceedings of the 2010 EDBT/ICDT Workshops*, ser. EDBT '10. New York, NY, USA: ACM, 2010, pp. 16:1–16:6. [Online]. Available: http://doi.acm.org/10.1145/1754239.1754257*

[3] *P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: An open architecture for collaborative filtering of netnews," in* Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '94. New York, NY, USA: ACM, 1994, pp. 175–186. [Online]. Available: http://doi.acm.org/10.1145/192844.192905*

[4] *D. Billsus and M. J. Pazzani, "A personal news agent that talks, learns and explains," in* Proceedings of the Third Annual Conference on Autonomous Agents*, ser. AGENTS '99. New York, NY, USA: ACM, 1999, pp. 268–275. [Online]. Available: http://doi.acm.org/10.1145/301136.301208*

[5] *M. Balabanović and Y. Shoham, "Fab: Content-based, collaborative recommendation,"* Commun. ACM*, vol. 40, no. 3, pp. 66–72, Mar. 1997. [Online]. Available: http://doi.acm.org/10.1145/245108.245124*

[6] *P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends." in* Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer, 2011, pp. 73–105. [Online]. Available: http://dblp.uni-trier.de/db/reference/rsh/rsh2011.html#LopsGS11*

[7] *K. Rao and V. Talwar, "application domain and functional classification of recommender systemsa survey," vol. 28, no. 3, pp. 17–35, 2008.*

[8] *. F. Kondert and A. Schandl, T. and Blumauer, "do controlled vocabularies matter? surevey results," pp. 17–35, 2011.*

[9] *A. Isaac and E. Summers, "Skos simple knowledge organization system primer," World Wide Web Consortium, Working Draft WD-skos-primer-20080829, August 2008.*

[10] *D. Werner and C. Cruz, "Precision difference management using a common sub-vector to extend the extended {VSM} method,"* Procedia Computer Science*, vol. 18, no. 0, pp. 1179 – 1188, 2013, 2013 International Conference on Computational Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050913004274*

[11] *P. Cimiano,* Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.*

[12] *G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval,"* Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Aug. 1988. [Online]. Available: http://dx.doi.org/10.1016/0306-4573(88)90021-0*

[13] *K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the c-value/nc-value method," 2000.*

[14] *P. Pantel and D. Lin, "A statistical corpus-based term extractor," in* Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, ser. AI '01. London, UK, UK: Springer-Verlag, 2001, pp. 36–46. [Online]. Available: http://dl.acm.org/citation.cfm?id=647462.726284*

[15] *G. Tsoumakas and I. Katakis, "Multi-label classification: An overview,"* Int J Data Warehousing and Mining*, vol. 2007, pp. 1–13, 2007.*

[16] *. I. . . I. Tsoumakas, G. and Katakis, "effective and efficient multilabel classification in domains with large number of labels,"* ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD08)*, vol. 2008, pp. 30–44, 2008.*

[17] *C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification,"* Mach. Learn.*, vol. 73, no. 2, pp. 185–214, Nov. 2008. [Online]. Available: http://dx.doi.org/10.1007/s10994-008-5077-3*

[18] *N. Holden and A. Freitas, "Hierarchical classification of g-protein-coupled receptors with a pso/aco algorithm,"* IEEE Swarm Intelligence Symposium (SIS 06)*, vol. 2006, 2006.*

[19] *W. Bi and J. T. Kwok, "Multilabel classification on tree- and dag-structured hierarchies," in* ICML*, 2011, pp. 17–24.*

[20] *I. Johnson, J. Abcassis, B. Charnomordic, S. Destercke, and R. Thomopoulos, "Making ontology-based knowledge and decision trees interact: An approach to enrich knowledge and increase expert confidence in data-driven models." in* KSEM*, ser. Lecture Notes in Computer Science, Y. Bi and M.-A. Williams, Eds., vol. 6291. Springer, 2010, pp. 304–316. [Online]. Available: http://dblp.uni-trier.de/db/conf/ksem/ksem2010.html#JohnsonACDT10*

[21] *A.-E. Elsayed, S. R. El-Beltagy, M. Rafea, and O. Hegazy, "applying data mining for ontology building,"* 42nd Annual Conference On Statistics, Computer Science, and Operations Research*, vol. 2007, 2007.*

[22] *S. Vogrincic and Z. Bosnic, "Ontology-based multi-label classification of economic articles."* Comput. Sci. Inf. Syst.*, vol. 8, no. 1, pp. 101–119, 2011. [Online]. Available: http://dblp.uni-trier.de/db/journals/comsis/comsis8.html#VogrincicB11*

[23] *A. L. Garrido, O. Gomez, S. Ilarri, and E. Mena, "An experience developing a semantic annotation system in a media group." in* NLDB*, ser. Lecture Notes in Computer Science, G. Bouma, A. Ittoo, E. Mtais, and H. Wortmann, Eds., vol. 7337. Springer, 2012, pp. 333–338. [Online]. Available: http://dblp.uni-trier.de/db/conf/nldb/nldb2012.html#GarridoGIM12*