n°9

Revue électronique des Technologies de l'Information
Electronic Journal of Information Technology

www.revue-eti.net

ISSN 1114-8802

Editorial du numéro 9

Editorial for the 9th edition

Mots-clés

e-TI, eTI, revue électronique, technologie de l'information, publication scientifique.

Keywords

e-TI, eTI, on line publication, e-journal, information technology, scientific publication.

Cette année s'est conclue par la migration du site Web de la revue vers une nouvelle plate-forme sous OJS *(Open Journal System)*. Cette solution, soutenue par de nombreuses universités, nous semble le gage d'une meilleure qualité de publication et d'une interface plus riche. Le bilinguisme est notamment géré, ce qui permet à e-TI de progresser dans cette direction. Nous avons également mené une réflexion éthique approfondie qui s'est exprimée dans notre politique de publication.

La rubrique **Etat de l'art** regroupe deux articles. L'article de Salim El Bazzi et al. propose des points de repère pour l'indexation automatique des textes arabes qui est une étape déterminante pour la fouille de textes. L'article souligne les spécificités de la langue arabe et analysent les réponses adaptées. Le second article est consacré à un thème d'actualité à savoir la sécurité de l'Internet des objets. Othmane El Mouaatamid et al. introduisent une classification des attaques et des contre-mesures possibles, organisées selon les couches OSI et l'objectif de sécurité à atteindre.

La rubrique **Recherche** comprend deux articles :

- L'article proposé par Davy Hélard et al. traite de la modélisation en temps continu dans le contexte de l'aide à la décision financière. Ils considèrent que la difficulté réside dans la nécessité d'aligner, selon une échelle commune, des modèles élaborés selon les échelles de temps différentes de chaque acteur. Leur solution a été implantée dans un contexte orienté services et illustrée par le cas de la société MGDIS.
- Le deuxième article se situe en amont de l'ingénierie dirigée par les modèles. Il s'agit de transformer les modèles de processus métiers en modèles de conception UML. Pour cela, Yassine Rhazali et al. enrichissent la modélisation des processus métiers à l'aide de DFD et de modèles d'activité UML. Ils génèrent une collection de modèles de niveau PIM tels le diagramme des cas d'utilisation, le diagramme de classe ou le diagramme d'états.

Dans la rubrique **Recherche et Développement**, Ali Sadiqui s'appuie sur son expérience à l'Office de la Formation professionnelle (OFPPT, Maroc) pour proposer un système d'aide à la mise en place souple de formations professionnelles adaptées aux clients demandeurs de formation. Destiné aux formateurx, ce système opérationnel organise les activités pédagogiques relatives à la formation, à l'évaluation et à la présence des étudiants.

Dans la rubrique **Fiche équipe**, c'est le Laboratoire IRF – SIC *(Image et Reconnaissance de Formes – Systèmes Intelligents et Communicants)* d'Agadir qui est à l'honneur. Il est présenté par Driss Mammass et Hassan DOUZI.

Pour terminer, Zineb Kacemi partage dans la rubrique **J'ai lu** ses réflexions concernant le livre de Najib Tounsi intitulé «Langage C».

Bonne lecture.

Mounia Fredj et Ounsa Roudiès
Rédactrices en chef de la 9ème édition

Les comités de la 9ème édition de la revue eTI

Committees for the 9th edition of the eTI Journal

Rédactrices en chef

- ROUDIES Ounsa — EMI, Univ. Mohammed V-Rabat, Maroc
- CHIADMI Dalila — EMI, Univ. Mohammed V-Rabat, Maroc

Rédactrice associée

- FREDJ Mounia — ENSIAS, Univ. Mohammed V-Rabat, Maroc

Comité de rédaction

- ABBAD Aïcha : suivi des soumissions
- BACHIRI Housseine, Univ. Ibn Tofail : relecture linguistique
- BENSAID Hicham, INPT : administration de la plate-forme
- IBNLKHAYAT Nozha : veille qualité et éthique
- KACEMI Zineb : administration de la plateforme
- MOKHTARI Abdelkrim, Univ. Ibn Tofail : qualification et relecture linguistique
- RAPOSO de BARBOSA Appoline : mise en page et infographie

Comité scientifique

- AHMED-NACER Mohamed — Univ. des S&T Houari Boumediene, Algérie
- ALIMAZIGHI Zahia — Univ. des S&T Houari Boumediene, Algérie
- AMGHAR Mustapha — GIE Galileo, Maroc
- AYACHI GHANNOUCHI Sonia — ESG-Sousse, Tunisie
- BAINA Karim — ENSIAS, Univ. Mohammed V-Rabat, Maroc
- BAINA Salah — ENSIAS, Univ. Mohammed V-Rabat, Maroc
- BAKKOURY Zohra — EMI, Univ. Mohammed V-Rabat, Maroc
- BELLAACHIA Abdelghani — George Washington University, USA
- BELOUADHA Fatima-Zahra — EMI, Univ. Mohammed V-Rabat, Maroc
- BENHLIMA Leila — EMI, Univ. Mohammed V-Rabat, Maroc
- BOUQATA Bouchra — General Electric Global research Center, USA
- BOUNABAT Bouchaïb — ENSIAS, Univ. Mohammed V-Rabat, Maroc
- CHIADMI Dalila — EMI, Univ. Mohammed V-Rabat, Maroc
- COHEN Atika — Univ. Libre de Bruxelles, Belgique
- COULETTE Bernard — IRIT, Univ. Toulouse Jean Jaurès. France
- DIOURI Ouafae — EMI, Univ. Mohammed V-Rabat, Maroc
- ELEULDJ Mohcine — EMI, Univ. Mohammed V-Rabat, Maroc
- EL MAGHRAOUI Kaoutar — Thomas J. Watson Research Center, IBM, USA
- EL MOHAJIR Mohammed — FSDM, Univ. Sidi Mohamed ben Abdellah, Maroc
- FREDJ Mounia — ENSIAS, Univ. Mohammed V-Rabat, Maroc
- FRONT Agnès — LIG, Univ. Grenoble Alpes, France
- GIRAUDIN Jean-Pierre — LIG, Univ. Grenoble Alpes, France
- KASSOU Ismail — ENSIAS, Univ. Mohammed V-Rabat, Maroc
- MEJRI Mohamed — Faculté des Sciences et de Génie, Univ. Laval. Canada
- MOULINE Salma — FSR, Univ. Mohammed V-Rabat, Maroc
- OULAD HAJ THAMI Rachid — ENSIAS, Univ. Mohammed V-Rabat, Maroc
- RIEU Dominique — LIG, Univ. Grenoble Alpes, France

Comité de lecture

Partenaires

Etat de l'art

Indexation automatique des textes arabes : état de l'art

Automatic indexing of Arabic documents: State of the art

Mohamed Salim El Bazzi
Laboratoire IRF-SIC, Université Ibn Zohr, Agadir, Maroc
elbazzi.mohamedsalim@edu.uiz.ac.ma

Taher Zaki
Laboratoire IRF-SIC, Université Ibn Zohr, Agadir, Maroc
t.zaki@uiz.ac.ma

Driss Mammass
Laboratoire IRF-SIC, Université Ibn Zohr, Agadir, Maroc
mammass@uiz.ac.ma

Abdelatif Ennaji
Laboratoire LITIS, Université de Rouen, Rouen, France
abdel.ennaji@univ-rouen.fr

Résumé

L'indexation des documents est une phase cruciale dans le processus de fouille de textes. Elle permet de représenter les documents par les descripteurs les plus pertinents vis-à-vis de leurs contenus. À ce propos, plusieurs approches sont proposées dans la littérature, notamment pour l'anglais, mais elles sont inexploitables par les documents en langue arabe en raison de ses caractéristiques spécifiques, de sa richesse morphologique et grammaticale et de son vocabulaire. Cet article dresse un état de l'art des méthodes d'indexation et de leurs apports à la langue arabe. Nous proposons une catégorisation des travaux selon les approches et les méthodes les plus utilisées en indexation automatique de documents textuels. Nous avons adopté une sélection qualitative des articles. Ainsi, avons-nous retenu les travaux constituant des contributions significatives au niveau de l'indexation et présentant des résultats considérables.

Abstract

Document indexing is a crucial step in the text mining process. It is used to represent documents by the most relevant descriptors of their contents. Several approaches are proposed in the literature, particularly for English, but they are unusable for Arabic documents, considering its specific characteristics and its morphological complexity, grammar and vocabulary. In this paper, we present a reading in the state of the art of indexation methods and their contribution to improve Arabic document's processing. We also propose a categorization of works according to the most used approaches and methods for indexing textual documents. We adopted a qualitative selection of papers and we retained papers approving notable indexation contributions and illustrating significant results.

Mots-clés

Fouille de textes, langue arabe, sémantique, méthode d'indexation, classification.

Keywords

Text mining, Arabic language, semantic, indexation methods, classification.

1. Introduction

La masse documentaire disponible sur internet et la numérisation des documents textuels ne cessent d'augmenter. Ce changement révolutionnaire présente de grands défis, et en même temps de grandes opportunités aux chercheurs pour exploiter les informations cachées en introduisant différentes approches.
La plupart des travaux effectués dans ce domaine ont été consacrés surtout aux langues occidentales, notamment l'anglais. En revanche, la langue arabe, étant une langue riche morphologiquement et fortement flexionnelle, a connu peu d'études au niveau de l'extraction des descripteurs. Ceci est dû au problème majeur de la complexité de son traitement automatique.
L'indexation des documents consiste à extraire les mots-clés qui représentent le mieux un document. Malgré le rôle primordial de cette phase dans la suite du processus de fouille et d'analyse des textes, peu sont les travaux recensés à ce niveau (Zaki, 2013). Cet article présente une lecture dans l'état de l'art des différentes méthodes d'extraction des descripteurs, ainsi que leurs applications et leurs compatibilités avec la langue arabe.
Le reste de cet article est organisé comme suit. La deuxième partie introduit le processus d'indexation. La troisième partie est dédiée à la présentation des différentes approches de sélection de descripteurs. Nous entamons une discussion dans la quatrième partie et enfin nous concluons par l'apport de ces approches au développement du traitement automatique de la langue arabe.

2. L'indexation des documents textuels

Indexer un document revient à élire ses descripteurs les plus représentatifs afin de générer la liste des termes d'indexation. C'est un moyen de retrouver l'ensemble des termes caractérisant un document. L'indexation des documents est une étape primordiale dans le processus de fouille de textes car elle détermine de quelle manière les connaissances contenues dans les documents sont représentées (Zaki, 2013) (Mountassir, 2012). Elle a lieu à chaque ajout d'un document dans le corpus étudié.
L'AFNOR définit l'indexation aussi comme le «Processus destiné à décrire et à caractériser, au moyen des termes ou indices d'un langage documentaire ou au moyen des éléments d'un langage naturel (libre), des données résultant de l'analyse du contenu d'un document (ressource, collection) ou d'une question, en vue d'en faciliter la recherche. On désigne également ainsi le résultat de cette opération».
Le processus d'indexation vise à faciliter le repérage de l'information dans un corpus documentaire. En conséquence, les approches d'indexation utilisées doivent faire face à deux problèmes majeurs :

- Le choix des termes représentatifs de chaque document. En effet, le choix de la forme des descripteurs, des méthodes de pondération et de sélection de termes définit le schéma sous lequel le document sera présenté.
- L'évaluation des index et de leur pouvoir de représentation. La liste des index retenus devrait couvrir tout le document et bien décrire son contenu.

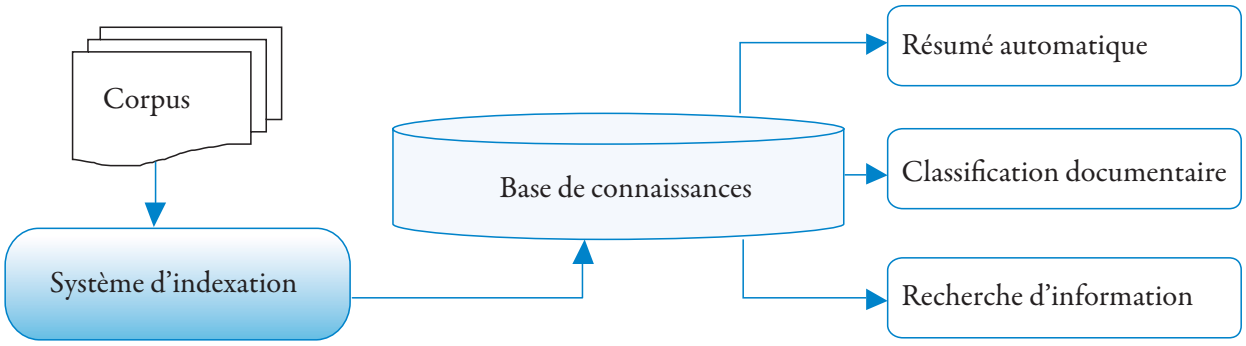

Figure 1. Processus de fouille de textes

2.1. Types d'indexation

L'indexation de toutes ses formes a pour but d'extraire les descripteurs les plus pertinents d'un document. Plus cette sélection est sophistiquée, plus les tâches ultérieures de fouille de textes exploitant le système d'indexation (classification, recherche d'information, etc.) s'avèrent précises. Il existe deux types fondamentaux d'indexation : l'indexation manuelle et l'indexation automatique.

En indexation manuelle, les descripteurs sont choisis par des experts ayant de bonnes connaissances linguistiques, le vocabulaire est alors contrôlé, une diversité de synonymes est utilisée et les aspects contextuels et sémantiques sont pris en considération. Malgré le fait que les spécialistes puissent se fonder sur les résumés des textes pour générer les index, l'indexation manuelle devient laborieuse dès que la masse documentaire s'accroît, et s'avère coûteuse en termes de temps de traitement des unités à indexer (Mallak, 2011).
L'indexation automatique, quant à elle, extrait des descripteurs automatiquement des textes en se fondant sur des règles d'analyse morphosyntaxique, des méthodes statistiques ou même sur des approches hybrides combinant les deux. Ce type d'indexation pallie les problèmes du coût de traitement, mais la conservation de la sémantique dans un document constitue alors un enjeu vital (Zaki, 2013).

2.2. Les formes de descripteurs

Les descripteurs sont les unités de texte qui représentent son contenu (Mallak, 2011). Il existe plusieurs catégories de descripteurs qui tentent d'une part de réduire la dimension du document, et d'autre part, de conserver son aspect sémantique. Nous en citons :

Les mots. Un mot est tout groupe de lettres formant un sens, compris entre deux séparateurs (espace, ponctuation, etc.). Le texte est alors segmenté en mots simples.

Les lemmes. Le processus de lemmatisation consiste à utiliser des règles grammaticales pour remplacer les verbes et les noms par leurs formes canoniques. Les lemmes peuvent ainsi correspondre à la forme des mots du dictionnaire.

Les racines *(stems)*. La racine est la plus petite unité lexicale qui permet de former un mot. Le processus de *stemming* extrait les racines de chaque mot du texte après l'élimination des préfixes, des infixes et des suffixes. Les mots partageant une racine commune sont tous associés à celle-ci.

Les concepts. Il s'agit des descripteurs issus d'une liste de lexique contrôlée, généralement des thésaurus ou des ontologies, et qui correspond aux notions principales contenues dans un document (Zaki, 2013).
Les multi-mots aussi appelés mots composés ou encore phrasèmes. Une séquence de mots décrit parfois plus précisément un document qu'un mot simple, en conservant mieux ses aspects contextuel et sémantique. Par exemple, l'expression «الأمم المتحدة» (Nations-Unies) est plus descriptive que «الأمم» (nations) et «المتحدة» (unies) prise séparément.

2.3. Préparation des documents

Avant d'entamer la phase d'indexation de documents arabes, une tâche très importante doit être accomplie, celle de nettoyage et de normalisation du texte (Bessou, Saadi *et al.*, 2007) (Mesleh, 2007a). Pour cette phase, les cinq étapes suivantes sont le plus souvent recommandées.

2.3.1. La segmentation (tokenization)

La segmentation est la production d'une séquence de segments séparés par des espaces ou des signes de ponctuation. La sortie est une liste de mots dépourvus de signe de ponctuation et de caractères spéciaux.

2.3.2. La suppression des mots vides

Les mots vides correspondent aux termes non porteurs d'information utile, qui figurent dans un texte. Il s'agit d'éliminer les mots dont l'occurrence est très fréquente et qui n'apportent aucune valeur ajoutée au processus d'indexation. Ces mots sont généralement des pronoms personnels, des articles ou des conjonctions, en l'occurrence : «أي», «أثناء», «إذ», etc. (Alajmi, Saad *et al.*, 2012).

2.3.3. Les conversions

La première conversion qui pourra être appliquée à un document arabe est l'élimination des signes diacritiques. Les signes diacritiques sont ajoutés au-dessus ou en dessous des lettres arabes afin de spécifier la prononciation du mot. Ce rôle phonologique influe aussi sur le sens de mot. En effet, deux mots peuvent être écrits de la même manière, mais différenciés par l'ajout de signes diacritiques différents. Par exemple, si le mot «عالم» est prononcé (عَالِم, *âalim*), il signifie «savant», et s'il est prononcé (عَالَم, *âalam*) il signifie «monde». Cette procédure vise à standardiser les documents du fait qu'il est rare de trouver un corpus entièrement accentué.
La deuxième conversion est celle des caractères qui a pour but de normaliser les lettres qui peuvent être écrites sous plusieurs formes. Ainsi les caractères «أ» «إ» et «آ» sont remplacés par «ا», de même «ة» est convertie en «ه» et «ي», «ئ» en «ى».

2.3.4. Le stemming

Le stemming consiste à extraire la racine d'un mot et à associer les mots liés morphologiquement à la même racine (Porter, 1980). Le nombre de termes est donc réduit, ce qui permet d'alléger le système. Cette technique dévoile un inconvénient majeur à savoir l'ambiguïté. Pour remédier à ce problème, la notion de *light stemming* est évoquée dans plusieurs travaux : elle consiste à éliminer juste les préfixes et les suffixes d'un mot donné, sans avoir à remonter à sa racine.

2.4. La représentation des documents

La représentation des documents est l'une des techniques qui sont utilisées pour réduire la complexité des documents et pour les rendre plus faciles à manipuler ; le document est alors transformé de sa version textuelle en une matrice [Document × Terme] (Figure 2). La représentation du document la plus utilisée est le modèle appelé vectoriel (VSM : *Vector Space Model*) dans lequel les documents sont représentés par des vecteurs de termes. Cette représentation a ses propres limites comme la grande dimension de représentation et la perte de corrélation entre les termes adjacents, ce qui entraîne la perte de la relation sémantique qui existe entre les termes d'un document. Pour surmonter ces problèmes, les méthodes de pondération sont utilisées pour attribuer des poids appropriés aux termes comme le représente la figure 2.

$$\begin{bmatrix} & \mathbf{T_1} & \mathbf{T_2} & \cdots & \mathbf{T_m} & \\ \mathbf{D_1} & p_{11} & p_{12} & \cdots & p_{1m} & \mathbf{C_a} \\ \mathbf{D_2} & p_{21} & p_{22} & \cdots & p_{2m} & \mathbf{C_b} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{D_n} & p_{n1} & p_{n2} & \cdots & p_{nm} & \mathbf{C_k} \end{bmatrix}$$

Figure 2. Matrice Document × Terme.

Chaque entrée représente un vecteur de termes où p_{nm} est le poids du terme $\mathbf{T_m}$ dans le document $\mathbf{D_n}$ et $\mathbf{C_i}$ est la classe attribuée au document $\mathbf{D_i}$.

2.5. La pondération

La pondération d'un terme d'indexation est l'association de valeurs numériques appelées poids à ce terme, de manière à représenter son pouvoir de discrimination pour chaque document de la collection. Cette caractérisation est liée au pouvoir informatif du terme pour le document donné. Le pouvoir de représentation d'un terme est parfois nommé l'informativité du terme. Cette notion fait référence à la quantité de sens qu'un mot porte.
Par exemple, la méthode TF-IDF (*Term Frequency – Inverse Document Frequency*) (Salton, Wong *et al.*, 1975) est l'une des méthodes les plus répandues dans le domaine de recherche documentaire (Trstenjak, Mikac *et al.*, 2013) (Zaki, Mammass *et al.*, 2010) (Hmeidi, Hawashin *et al.*, 2008) et elle est notamment très utilisée en modèle vectoriel.
TF représente le nombre d'occurrences d'un mot dans le document. IDF est la fréquence absolue inverse et égale à :

$$IDF_t = \log (N/n_t) \quad (1)$$

avec **N**, le nombre total de documents dans la collection et $\mathbf{n_t}$, le nombre de documents où le terme **t** apparaît.
Le poids d'un terme **t** dans le document d s'écrit généralement :

$$Poids_d(t) = TF_{dt} \times IDF_t \quad (2)$$

où $\mathbf{TF_{dt}}$ est la fréquence d'apparition du terme **t** dans le document **d** et $\mathbf{IDF_t}$ est la fréquence absolue inverse du terme **t** dans le corpus. Ainsi, le poids d'un terme augmente si celui-ci est fréquent dans le document et décroît si celui-ci est fréquent dans la collection.
Il existe d'autres façons de déterminer le poids d'un terme, en l'occurrence, la pondération booléenne, la fréquence de mots, l'entropie, etc. Cependant, les méthodes purement statistiques ont deux inconvénients majeurs. D'une part, il en résulte une énorme matrice creuse, ce qui pose un problème de grande dimension. D'autre part, elles ignorent la modélisation sémantique du document. De nouvelles méthodes que nous aborderons dans la suite sont apparues pour pallier ces limites.

2.6. La réduction de dimension

Après le prétraitement et l'indexation, une étape importante pour la classification de textes s'impose : il s'agit de réduire la dimension du texte (Mountassir, 2012). L'idée principale est de sélectionner un sous-ensemble de termes caractéristiques du document, et ce, en gardant les mots dotés des scores les plus élevés, en appliquant des mesures confirmant l'importance des termes sélectionnés. De nombreuses mesures d'évaluation des termes sont utilisées dans la littérature, nous en citons : le seuillage de fréquence (*Document Frequency Thresholding*), le gain d'information, la mesure de Chi-deux χ^2, *Odds Ratio* et l'information mutuelle.

2.7. La classification

La classification du texte est une partie importante du processus de fouille de textes (Figure 1). Elle consiste à fournir un ensemble de données d'apprentissage (documents étiquetés) au système de classification. La tâche est alors de déterminer un modèle de classification qui soit capable d'affecter la bonne classe à un nouveau document. Ces dernières années, la tâche de classification automatique de textes a été largement étudiée et les progrès semblent rapides dans ce domaine (Al-Mahmoud et Al-Razgan, 2015). Plusieurs méthodes de classification ont fait l'objet d'études comparatives et ont prouvé leur efficacité. A titre illustratif, nous citons : le classificateur bayésien, les arbres de décision, K-plus proche voisin (K-ppv), *Support Vector Machines* (SVM), et les réseaux de neurones (Alsaleem, 2011) (Bawaneh, Alkoffash *et al.*, 2008) (Alsalemi et Aziz, 2011) (El-Kourdi, Bensaid *et al.*, 2010).
La classification des documents textuels présente de nombreux défis et difficultés. Tout d'abord, il est difficile d'exprimer la sémantique de haut niveau et des concepts abstraits de la langue naturelle avec seulement quelques mots-clés, ce qui confirme le fait que l'efficience de l'étape d'indexation est primordiale et décisive.

2.8. Evaluation des systèmes d'indexation

L'évaluation expérimentale des classificateurs représente la dernière étape du processus d'indexation. Elle tente généralement d'évaluer l'efficacité d'un classificateur, à savoir sa capacité de prendre les décisions de catégorisation. Il existe à cet effet de nombreuses mesures, chacune mettant en évidence telle ou telle propriété du système. Nous avons retenu les mesures les plus utilisées suivantes : le rappel (3) qui est synonyme du taux de vraie acceptation, la précision (4) qui mesure le taux de bonnes réponses parmi les
réponses positives et la f-mesure (5) qui synthétise les deux premières. Considérons les nominations suivantes :

- **TP** *(True positive)* i.e. le nombre de documents correctement attribués à une catégorie,
- **FN** *(False Negative)* i.e. le nombre de documents incorrectement attribués à une catégorie,
- **FP** *(False positive)* i.e. le nombre de documents incorrectement rejetés affectés à une catégorie,
- **TN** *(True Negative)* i.e. le nombre de documents correctement rejetés attribués à une catégorie.

$$\textbf{Rappel} = \frac{TP}{TP + FN} \quad (3) \qquad \textbf{Précision} = \frac{TP}{TP+FP} \quad (4)$$

$$\textbf{F-mesure} = \frac{2 \times \text{Rappel} \times \text{Précision}}{(\text{Rappel} + \text{Précision})} \quad (5)$$

3. L'extraction des descripteurs

La fouille de textes repose sur un ensemble de techniques qui analysent de grandes quantités de données, extraient des relations qui sont inconnues au préalable, et fournissent des solutions pour aider à mieux représenter et exploiter ces données. L'indexation des documents, appelée aussi extraction des descripteurs, consiste à extraire les mots clés les plus pertinents dans un document qui décrivent mieux son contenu.

3.1. Problématique

Par rapport à d'autres langues, la langue arabe a une variation morphologique très riche et des caractéristiques syntaxiques extrêmement complexes, ce qui est l'une des principales raisons qui explique le manque de méthodes de recherche dans le domaine du traitement des textes arabes (El-Halees, 2007), (Samir, Ata *et al.*, 2005). L'indexation et la classification de textes sont des tâches importantes de ce traitement. Un processus typique de la classification de textes se compose des étapes suivantes : prétraitement, indexation, réduction de la dimension et classification (Wei, Gao *et al.*, 2010).

Un ensemble de modèles de classification et des techniques d'apprentissage automatique ont été appliqués à la classification de textes arabes, comme l'illustre la liste suivante :

- les K plus proches voisins (Kanaan, Al-Shalabi *et al.*, 2006) (Syiam, Fayed *et al.*, 2006),
- le modèle bayésien (El Kourdi, Bensaid *et al.*, 2004),
- SVM (Gharib, Habib *et al.*, 2009) (Alsaleem, 2011) (Mesleh, 2007b), (Mesleh, 2007c),
- les réseaux de neurones (Harrag, El-Qawasmah *et al.*, 2011),
- le maximum d'entropie (El-Halees, 2007) (Sawaf, Zaplo *et al.*, 2001),
- l'algorithme de Rocchio (Syiam, Fayed *et al.*, 2006),
- le classificateur à base de distances (Duwairi, 2005) (Khreisat, 2006) (Duwairi, 2006),
- les classificateurs à base de connaissances WordNet (Benkhalifa, Mouradi *et al.*, 2001).

Cependant, la phase d'extraction des descripteurs n'a pas eu le même intérêt, malgré son rôle primordial et décisif en classification. (Al-Mahmoud et Al-Razgan, 2015) présentent une étude systématique des techniques de fouille de textes qui confirme le manque de travaux concernant l'extraction des caractéristiques des documents arabes. La figure 3 illustre la distribution des techniques de fouille de textes décrites dans cette étude qui a couvert plus de cent articles.

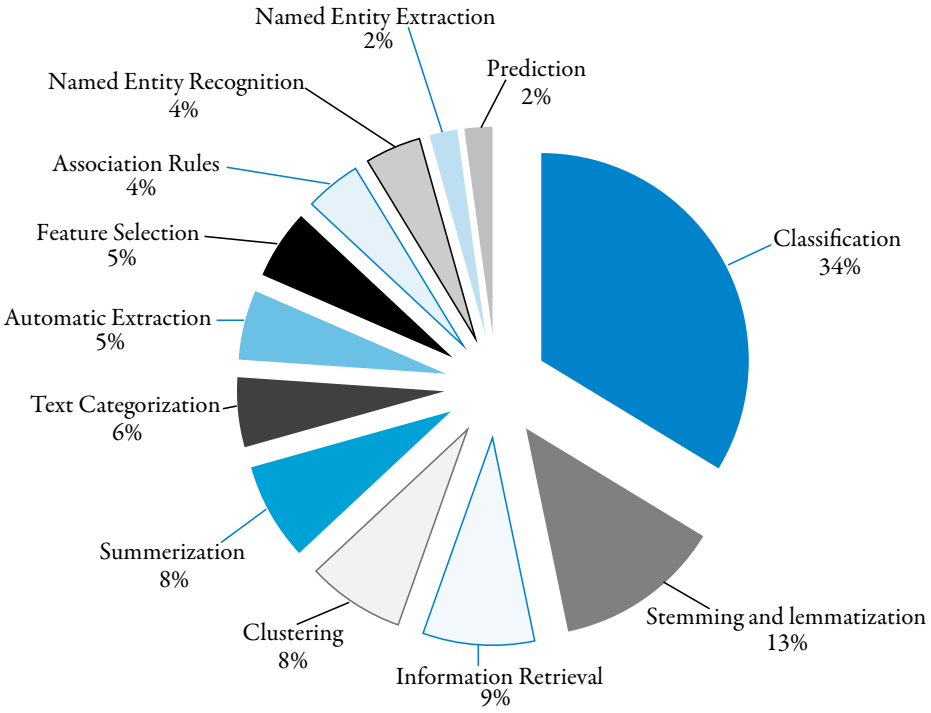

Figure 3 : Distribution des études en fouille de textes arabes selon (Al-Mahmoud et Al-Razgan, 2015).

Dans cet article, nous avons adopté une sélection qualitative des articles. Ainsi, avons-nous retenu les travaux montrant des contributions remarquables au niveau de l'extraction des descripteurs et présentant des résultats considérables.

3.2. Méthodes d'extraction des descripteurs

Dans cette section, nous souhaitons, d'une part, catégoriser les travaux selon les approches les plus utilisées pour synthétiser les avancées concernant l'extraction des descripteurs. D'autre part, nous introduisons quelques travaux susceptibles d'inspirer un système d'indexation de la langue arabe plus sophistiqué.

3.2.1. Approches linguistiques

Une approche linguistique consiste à apporter une analyse morphologique et syntaxique profonde du document traité, et ce, en se fondant sur les règles grammaticales et les relations entre les différentes unités textuelles, pour des fins de désambiguïsation sémantique, ou plus encore, d'indexation. Plusieurs tentatives de modélisation des règles linguistiques ont été proposées dans la littérature.

Al Molijy *et al.* (2012) utilisent l'analyse syntaxique des mots du document. L'algorithme proposé consiste à découper les mots en N-grammes (où N vaut 3, 4 ou 5), calculer leurs fréquences, ensuite retenir les 100 premiers mots les plus fréquents constituant ainsi un profil N-gramme du document. Cette méthode, utilisée aussi en processus de *stemming*, permet de réduire le nombre des mots représentant un document.

Mansour *et al.* (2008) procèdent à une analyse morphologique des mots du document pour extraire les index. D'une part, les auteurs proposent un processus d'extraction des stems. D'autre part, ils mettent en place un système de reconnaissance des noms et des verbes en se fondant sur les rimes et les règles grammaticales. Un poids est ensuite attribué à chaque stem en tenant compte de son occurrence et en introduisant une fonction indiquant comment le mot est étalé dans le document.
Aussi bien (Saadane, 2013) que (Bessou, Saadi *et al.*, 2007) proposent des systèmes d'extraction des connaissances, fondés sur une analyse linguistique profonde et faisant appel à une ontologie de domaine pour révéler le contenu sémantique. Les résultats de leurs travaux s'annoncent prometteurs, mais révèlent d'autres problématiques nécessitant des études minutieuses.
Quant à Hulth (2003), il présente une approche d'extraction de *chunk* nominaux (unité textuelle minimale ayant un sens, composée d'un mot ou plus) et de N-gramme. Dans ses expériences, Hulth intègre l'étiquetage en parties de discours, ce qui augmente significativement les performances du système.
Ces méthodes sont largement exploitées en fouille de textes arabes, grâce à la précision de leurs résultats et la fiabilité des algorithmes de reconnaissance syntaxique et sémantique. Le défi majeur de ces méthodes est de couvrir la diversité grammaticale et le vocabulaire de la langue arabe.

3.2.2. Approches numériques

Ces approches se fondent sur les techniques statistiques, la théorie des graphes ou les approches sémantiques prises séparément ainsi que sur leur combinaison.

3.2.2.1. Méthodes statistiques

Cette section du papier présente les travaux qui sont fondés sur des méthodes et des mesures purement statistiques pour l'extraction des mots clés. Ceci est le critère que nous avons adopté pour préparer ce regroupement de méthodes.
Nombreux sont les travaux qui ont adopté des approches statistiques pour l'extraction des mots clés, en étudiant le comportement des termes candidats dans un document, voire dans le corpus. Plus un terme candidat est jugé important dans le document, plus celui-ci est pertinent comme terme clé.
La méthode TF-IDF (5) fournit une bonne représentation du poids pour les corpus dont les documents sont de tailles homogènes, c'est-à-dire composés de documents de tailles similaires.
De nombreuses variantes de TF-IDF sont proposées dans la littérature, et elles ont fait objet d'un grand nombre de comparaisons. Okapi qui est une méthode alternative à la TF-IDF est très utilisée en recherche d'information. Elle prend mieux en compte la longueur des documents (Robertson, Walker *et al.*, 2000) (6) (7).

$$\text{Okapi (terme)} = \text{TF}_{\text{BM25}}\,\text{(terme)} \times log\left(\frac{N\text{-}DF(terme)+0{,}5}{DF(terme)+0{,}5}\right) \quad (6)$$

$$\text{avec :}\ \ \text{TF}_{\text{BM25}} = \frac{TF(terme)\times(k1+1)}{TF(terme)+k1\times\left(1-b+b\times\frac{DL}{DLmoyenne}\right)} \quad (7)$$

où **DL** représente la longueur du document traité et $\mathbf{DL}_{\mathbf{moyenne}}$ la longueur moyenne des documents de la collection. $\mathbf{K}_1$ et **b** sont des constantes fixées respectivement à 2 et 0,75 (Bougouin, 2013).
El-Khoribi et Ismael (2006) ont appliqué les stems comme caractéristiques de représentation. Ces caractéristiques sont ensuite représentées en tant que vecteurs de dimension égale au nombre de classes où la probabilité d'appartenance d'un stem est prise en considération. Ensuite une table de correspondance de stems est construite à partir des racines et des étiquettes des classes auxquelles elles appartiennent. Après, le modèle de Markov caché (HMM : Hidden Markov Model) est utilisé pour évaluer l'appartenance d'un nouveau document à une classe.
Khreisat (2006) a construit un système de classification de documents textuels arabes à l'aide de la technique statistique fréquentielle N-grammes et en utilisant une mesure de dissemblance appelée la distance de Manhattan, et l'opérateur de Dice comme mesure de similarité. La mesure de Dice a été utilisée à des fins de comparaison. Les résultats ont montré que la classification de textes en utilisant les N-grammes et la mesure de Dice surpasse la classification fondée sur les N-grammes et la mesure de Manhattan.
El-Halees (2007) a présenté des résultats prometteurs obtenus en utilisant des méthodes statistiques telles que l'entropie maximale sur une base d'articles en arabe et sans analyse morphologique.
(Mesleh, 2007a) a étudié l'usage du classificateur SVM avec six techniques de sélection des caractéristiques. Leurs expériences montrent que χ^2 s'impose par rapport aux autres
techniques.
Dans (Thabtah, Hadi *et al.*, 2008), les variantes du modèle vectoriel sont étudiées à l'aide de l'algorithme K-ppv. Ces variantes sont le coefficient Cosinus, le coefficient de Dice et le coefficient de Jacaard, en utilisant différentes méthodes de pondération des termes. Les résultats obtenus sur une base arabe ont montré que les performances

obtenues par Dice-TFIDF et Jaccard-TFIDF surpassent celles obtenues par Cosinus TFIDF, Cosinus à base FDIF, Cosinus-ITF, Cosinus à base log(1 + *TF*), Dice à base FDIF, Dice à base ITF, Dice à base log(1+*TF*), Jaccard à base FDIF, Jaccard à base ITF, et Jaccard à base log(1+*TF*).
C'est ACO *(Ant Colony Optimization)* qui est appliqué dans (Mesleh et Kanaan, 2008) comme mécanisme de réduction de l'espace des caractéristiques. La méthode χ^2 est utilisée comme une fonction de calcul de scores. Ils ont ensuite procédé à la classification des documents arabes en utilisant le classificateur SVM.
(Al-Shalabi et Obeidat, 2008) ont utilisé un K-ppv pour classer les documents arabes. Ils extraient en tant que caractéristiques des mots clés donnés par les unigrammes et les bigrammes, ensuite la mesure de TFIDF est appliquée en tant que procédé de sélection de ces caractéristiques.
(Al-Harbi, Almuhareb *et al.*, 2008) a testé le SVM et la C5.0 sur sept corpus arabes avec des descripteurs pondérés par χ^2. Les performances obtenues sont de 86% pour le SVM et de 92% pour C5.0.
(Bawaneh, Alkoffash *et al.*, 2008) a comparé les deux classificateurs K-ppv et NB. Le *light stemmer* a été utilisé comme caractéristique et la mesure TFIDF en tant que méthode de pondération des caractéristiques. Le classificateur K-ppv a été jugé plus performant.
(Thabtah, Eljinini *et al.*, 2009) a mis en place un système de catégorisation arabe en utilisant le classificateur bayésien naïf fondé sur les caractéristiques de pondération fournies par le test de χ^2 pour classer une simple base de données étiquetées. Les résultats expérimentaux montrent que la sélection des caractéristiques améliore souvent la précision de la classification en supprimant les termes vides ou rares.
Dans (Kanaan *et al.*, 2009), les documents en arabe sont classés avec l'algorithme espérance-maximisation (EM). La mesure TFIDF est appliquée en tant que méthode de pondération des éléments caractéristiques tandis que l'algorithme bayésien naïf est utilisé pour calculer les étiquettes des documents et que finalement on procède à la classification en utilisant l'algorithme EM.
(Zubi, 2009) a comparé les deux classificateurs K-ppv et NB appliqués à une base de 1562 documents. Ces derniers sont classés en 6 catégories et pondérés en utilisant la mesure TFIDF. L'expérience a montré que K-ppv est plus performant.
(Gharib *et al.*, 2009) ont appliqué quatre classificateurs, SVM, Bayésien naïf, K-ppv et la méthode de Rocchio, à une base de documents arabes, en utilisant le *stemming* comme méthode de représentation des caractéristiques et la mesure TFIDF comme méthode de pondération. Le classificateur de Rocchio fonctionne mieux lorsque l'espace des caractéristiques est petit, mais le SVM est plus performant quand l'espace devient de plus en plus grand.
Dans (Al-Shalabi, Kanaan *et al.*, 2010), l'algorithme de k plus proches voisins et les mots clés sont extraits selon leur pondération TFIDF dans les documents, en obtenant une micro-moyenne précision de 95%.
Dans leur étude comparative (Raheel et Dichy, 2010) ont montré l'influence du choix de type d'entités à manipuler sur les performances des classificateurs. Ils ont choisi comme descripteurs, les mots dans leur forme originale, les lemmes, les racines, et les n-grammes. Deux classificateurs ont été utilisés, le SVM et les réseaux bayésiens naïfs. Le SVM basé sur les 3-grammes a donné de meilleurs résultats de classification avec une F-mesure dépassant 92%.
(Al-Salemi et Aziz,2011) ont utilisé des techniques de sélection de caractéristiques telles que l'information mutuelle, la mesure statistique χ^2, le gain d'information, le coefficient ESG et Odds Ratio pour réduire la dimension de l'espace des caractéristiques en éliminant les éléments qui sont considérés comme non pertinents pour une catégorie étudiée.
D'autres modèles sont utilisés dans la littérature (Hasan et Ng, 2010) (Bougouin, 2013) (Mesleh, 2007a) tels que LSI *(Latent Semantic Indexing)* qui prend en considération la sémantique des termes pour la représentation des documents. Les documents sont représentés dans un espace réduit de termes d'indexation. (Hofmann, 1999) propose un modèle probabiliste de *Latent Semantic Indexing* (PLSI). Il émet l'hypothèse que les documents sont associés à un certain nombre de sens et que les termes correspondent à l'expression de ces sens.
En conclusion, ces méthodes, considérées comme simples à implémenter, sont efficientes et parfaitement tolérantes aux grandes masses documentaires. D'autre part, l'hypothèse considérant les mots comme étant des unités indépendantes engendre une perte d'information sémantique. Les index qui en résultent peuvent générer des problèmes de polysémie et dévier du contexte général du document.

3.2.2.2. Méthodes fondées sur les graphes

Ces méthodes proposent de représenter le texte sous forme de graphe. Généralement, les mots constituent les nœuds du graphe et les arcs représentent la relation entre les mots (relation sémantique, structurelle, etc.).
Mihalcea et Tarau (2004) proposent l'algorithme TextRank, une adaptation textuelle de l'algorithme PageRank (Page, Brin *et al.*, 1998). Il consiste à représenter les documents textuels sous forme de graphe où les nœuds peuvent représenter un mot ou un groupe de mots. Une pondération $\mathbf{w_{n1n2}}$ est associée à chaque arc liant deux nœuds $\mathbf{n_1}$ et $\mathbf{n_2}$, et représente la fréquence de cooccurrence des deux termes dans une fenêtre de **N** mots.
Le score du nœud $\mathbf{n_i}$, noté $\mathbf{S(n_i)}$, est initialisé par une valeur par défaut, et il est ensuite calculé d'une manière itérative jusqu'à convergence en utilisant la formule suivante :

$$S(n_i)=(1-d)+ d \times \sum_{n_j \in Adj(n_i)} \frac{w_{ji}}{\sum_{n_k \in Adj(n_j)} w_{jk}} S(n_i) \quad (8)$$

où **Adj(n_j)** représente les voisins de $\mathbf{n_i}$, et **d** est un facteur d'amortissement fixé à 0.85 (Page, Brin *et al.*, 1998). Intuitivement, un nœud recevra un score élevé si ses voisins ont des scores élevés. Finalement, après convergence, les **k%** termes ayant des scores élevés sont élus comme mots clés.

Dans leurs travaux, Mihalcea et Tarau utilisent l'étiquetage en parties de discours afin de réduire la liste des termes représentés par le graphe, et ce, en ne considérant que les noms et les adjectifs, ce qui améliore les performances du système. Cependant, l'orientation du graphe n'apporte pas d'amélioration à considérer, par rapport au graphe non orienté.

Des variantes de cet algorithme ont vu le jour, par exemple (Wan et Xiao, 2008b). Les auteurs proposent l'algorithme SingleRank qui définit trois différences majeures par rapport à TextRank. Premièrement, les arcs ont des poids correspondants au nombre de cooccurrence des deux termes connexes alors que les arcs ont le même poids pour TextRank. D'autre part, TextRank procède à un filtrage de termes contrairement à SingleRank qui n'effectue aucune discrimination. En outre, pour chaque phrasème candidat, le score est calculé en sommant les scores des termes formant ce phrasème, obtenus de la représentation graphe SingleRank. Les phrasèmes candidats ayant les plus grands scores sont considérés des termes clés.

ExpandRank (Wan et Xiao, 2008b) est une extension de TextRank qui consiste à exploiter le voisinage du document analysé. Pour un document **d**, les k-plus proches voisins sont trouvés à partir des documents de la collection, le graphe est généré ensuite à partir du document traité et ses k plus proches voisins. Ainsi, chaque document $\mathbf{d_0}$ est-il réuni avec ses documents voisins dk, formant un document plus large $\mathbf{d_{k+1}}$ qui servira à la construction du graphe. Les termes candidats correspondent aux nœuds et un arc relie deux nœuds si les termes candidats co-occurrent dans une fenêtre de N mots du document. Le poids de l'arc liant deux nœuds $\mathbf{v_i}$ et $\mathbf{v_j}$ est donné par :

$$w(v_i, v_j) = \sum_{dk \in D} sim(d_0, dk) \times freq_{dk}(v_i, v_j) \quad (9)$$

où :

- **sim($\mathbf{d_0}$, $\mathbf{d_k}$)** est la similarité cosinus entre $\mathbf{d_0}$, et $\mathbf{d_k}$,
- **freqdk**($\mathbf{v_i}$, $\mathbf{v_j}$) est la fréquence de cooccurrence des termes $\mathbf{v_i}$ et $\mathbf{v_j}$ dans dk.

Dès que le graphe est construit, le reste de la procédure est similaire à SingleRank.

MedRank est un algorithme proposé par (Herskovic et Jorge, 2011) pour réordonner les rangs des concepts extraits d'une base médicale. Ces concepts sont extraits par le programme MetaMap dans un premier temps. De nouveaux scores sont ensuite affectés aux concepts en utilisant l'algorithme TextRank. Les meilleurs résultats sont obtenus en utilisant l'approche MedRank.

	Référence	Techniques utilisées	Observation
Méthodes Statistiques	(El-Khoribi et Ismael, 2006)	probabilité d'appartenance, HMM	HMM est utilisé pour évaluer l'appartenance d'un nouveau document à une classe
	(Khreisat, 2006)	N-grammes, la distance de Manhattan, la mesure de Dice	La classification de textes en utilisant les N-grammes avec la mesure de Dice surpasse la classification en utilisant les N-grammes avec la mesure de Manhattan
	(El-Halees, 2007)	l'entropie maximale	Les résultats obtenus sont prometteurs.
	(Mesleh, 2007a)	χ^2, SVM.	Six techniques de sélection des caractéristiques introduites. χ^2 est la plus performante.
	(Thabtah, Hadi *et al.*, 2008)	K-ppv, Cosinus, Dice, Jaccard, TFIDF.	Les résultats obtenus sur une base arabe ont montré que les performances obtenues par Dice-TFIDF et Jaccard-TFIDF sont les plus élevées.
	(Mesleh et Kanaan, 2008)	Ant Colony Optimization (ACO), χ^2, SVM.	ACO est appliqué pour réduire l'espace de représentation des caractéristiques et la méthode χ^2 pour le calcul de scores.
	(Al-Shalabi et Obeidat 2008)	K-ppv, unigrammes, bigrammes, TFIDF	Les mots clés donnés par les unigrammes et les bigrammes, sont pondérés par la mesure de TFIDF.
	(Al-Harbi, Almuhareb *et al.*, 2008)	SVM, C5.0, χ^2.	Le test est effectué sur sept corpus arabes.

Méthodes Statistiques *(suite)*	(Bawaneh, Alkoffash *et al.*, 2008)	K-ppv, NB, TFIDF	Le light stemmer a été utilisé comme caractéristique.
	(Thabtah, Eljinini *et al.*, 2009)	NB, χ^2.	Les résultats expérimentaux montrent que la sélection des caractéristiques améliore souvent la précision de la classification.
	(Kanaan *et al.*, 2009)	espérance-maximisation, TFIDF, NB.	Les documents arabes sont classés avec l'algorithme espérance-maximisation (EM).
	(Zubi, 2009)	K-ppv, NB, TFIDF.	La classification se fait sur un corpus de 1562 documents appartenant à 6 catégories différentes. L'expérience a montré que K-ppv est plus performant.
	(Gharib *et al.*, 2009)	SVM, NB, K-ppv, Rocchio, TFIDF	Le classificateur de Rocchio fonctionne mieux lorsque l'espace des caractéristiques est petit mais le SVM est plus performant quand l'espace devient grand.
	(Al-Shalabi, Kanaan *et al.*, 2010)	K-ppv, TFIDF.	Implémentation classique pour la classification des textes arabes.
	(Raheel et Dichy, 2010)	SVM, NB, 3-grammes.	Dans leur étude comparative, les auteurs ont montré l'influence du choix de type d'entités à manipuler sur les performances des classificateurs.
	(Al-Salemi et Aziz,2011)	l'information mutuelle, χ^2, le gain d'information, le coefficient ESG et Odds Ratio.	Ces techniques sont utilisées pour réduire la dimension de l'espace des caractéristiques en éliminant les éléments qui sont considérés comme non pertinents pour une catégorie étudiée.
Méthodes fondées sur les graphes	(Mihalcea et Tarau, 2004)	TextRank.	Adaptation textuelle de l'algorithme PageRank. Les nœuds peuvent représenter un mot ou un groupe de mots et les arcs n'importe quelle relation reliant les mots.
	(Wan et Xiao, 2008b).	SingleRank.	Variantes de TextRank.
	(Wan et Xiao, 2008b)	ExpandRank, K-ppv.	Extension de TextRank qui consiste à exploiter le voisinage du document analysé.
	(Herskovic et Jorge, 2011)	MedRank, MetaMap,	MedRank est un algorithme pour réordonner les rangs des concepts extraits d'une base médicale en utilisant l'algorithme TextRank

Tableau 1 : Synthèse des approches numériques.

Les méthodes d'indexation à base de graphes semblent être mieux adaptées aux textes bruts pour leur efficacité à conserver l'aspect structurel. Cependant, la complexité de calcul des scores des nœuds générés à partir des textes constitue une limite majeure. Pourtant, les méthodes statistiques restent les plus utilisées pour leur simplicité en implémentation et leurs résultats efficients.

3.2.2.3. Approches sémantiques

Ces approches visent, d'une part, à lever l'ambiguïté sur le sens des mots et d'autre part, elles permettent de tisser les relations sémantiques entre ces mots. Les textes sont représentés par des concepts symbolisant le sens plutôt que des mots simples. Les relations sémantiques peuvent aussi être calculées par le biais des méthodes évaluant la quantité d'information entre les mots deux à deux, en l'occurrence, l'information mutuelle.

D'autres chercheurs sont allés jusqu'à l'exploration de l'information contextuelle. Le travail de Zargayouna et Salotti (2004), testé sur un corpus de documents semi-structurés en XML, considère le document comme un ensemble d'unités sémantiques (les balises) représentant chacune un contexte particulier d'occurrence des termes. Néanmoins, dans (Roche, 2011), l'auteur travaille sur la désambiguïsation des acronymes et définit le contexte comme des mots caractéristiques présents dans la page dans laquelle l'acronyme à définir est présent. Pareillement, Motasem et Joseph (2009) proposent une méthode d'exploration contextuelle afin de lever l'ambigüité de la séquence «alif-noun».
Cependant, Jamoussi (2009) propose une méthode d'extraction de mots clés en se fondant sur la représentation sémantique des termes. Il présente deux mesures fondées sur des distances sémantiques, la distance de Kullback-Leibler (DKL) et l'information mutuelle moyenne (IMM), pour calculer la quantité d'information entre deux mots ou deux classes de mots. Cette méthode est testée par rapport à une représentation vectorielle simple, avec trois classificateurs non supervisés : l'algorithme K-means, les cartes de Kohonen et le réseau bayésien AutoClass. Le Tableau 2 synthétise les résultats obtenus. Ces résultats expriment, en pourcentage, le taux de bonne classification, en précisant les intervalles de confiance. La performance des résultats met en évidence l'importance de l'utilisation des mesures sémantiques pour la fouille de textes.

Méthode Jamoussi		K-means	Kohonen	AutoClass
Représentation vectorielle simple	*DKL*	70,5 ± 2,2	75,5 ± 2,1	81,3 ± 1,9
	IMM	74,1 ± 2,1	77,1 ± 2,1	84,7 ± 1,8
Représentation matricielle mixte	*DKL*	72,5 ± 2,2	76,7 ± 2,1	86,3 ± 1,7
	IMM	76,4 ± 2,1	80,4 ± 1,9	89,3 ± 1,5

Tableau 2 : L'apport de l'approche matricielle mixte par rapport à la représentation vectorielle standard (Jamoussi, 2009)

En outre, une autre technique originale est exploitable pour les données textuelles en langue arabe, il s'agit du regroupement sémantique. (Liu, Peng *et al.*, 2009) proposent une méthode d'extraction de mots clés fondée sur le regroupement sémantique qui garantit une bonne couverture sémantique du document. La méthode extrait les termes candidats qui seront regroupés en classes après le calcul des liens sémantiques entre ces termes. Ce regroupement consiste à élaborer un ensemble de mots de référence pour chaque classe. Les mots de référence sont utilisés pour l'extraction des mots clés après le filtrage des termes candidats. Un mot clé doit contenir au moins un mot de référence.

3.2.3. Approches hybrides

L'adoption des approches hybrides pour la fouille de données textuelles est devenue courante. Plusieurs chercheurs essaient différentes combinaisons des méthodes linguistiques, numériques et sémantiques afin de révéler l'information cachée dans un document, et enrichir les liens contextuels qu'il contient (Bessou, Saadi *et al.*, 2007) (Jamoussi, 2009) (Saadane, 2013). Ces approches aboutissent souvent à des résultats meilleurs que ceux obtenus par l'utilisation des méthodes standards.
Par exemple, Tomokyo et Hurst (2003) proposent une méthode qui vérifie la grammaticalité (un mot clé doit être bien formé syntaxiquement) et l'informativité (le mot clé doit exprimer au moins une idée du contexte général du document) en utilisant la Kullback-Leibler divergence. Ainsi, pour un terme candidat, plus sa probabilité de passer du modèle uni-gramme généré à partir du document analysé au modèle N-gramme généré par le même document augmente, plus il respecte la propriété de grammaticalité. De même, plus sa probabilité de passer du modèle N-gramme généré à partir d'un corpus de référence vers un modèle N-gramme généré par le document traité augmente, plus le terme candidat est informatif.
Dans le but de réduire l'espace des caractéristiques (Harrag, El-Qawasmah *et al.*, 2011) compare trois techniques de prétraitement : *light stemming*, *root-based stemming* et *dictionary lookup stemming*. Ensuite, deux classifieurs ont été testés : les réseaux de neurones artificiels (ANN) et le SVM. Les performances données par SVM sont supérieures à celles de ANN avec le *light stemming*.
La méthode de (Motasem et Joseph, 2009) se fonde essentiellement sur l'analyse morphosyntaxique et l'exploitation des règles grammaticales pour la reconnaissance des mots adjacents à la séquence en question, pour découvrir son contexte. Néanmoins, d'après (Alwedyan *et al.*, 2011), leur propre classificateur multi-classes à base de règles d'associations fonctionne mieux que NB et SVM.
Quant au travail (Zaki, Mammass et al, 2014), les auteurs introduisent la notion de voisinage sémantique. Ils proposent un système hybride pour l'indexation contextuelle et sémantique des documents arabes, apportant une amélioration aux modèles classiques fondés sur les n-grammes et le modèle Okapi. Ils calculent la similarité entre les mots en utilisant une hybridation de mesures statistiques N-grammes, Okapi et une fonction noyau. Afin d'avoir

un indice de descripteur robuste, ils ont utilisé un graphe sémantique pour modéliser les connexions sémantiques entre les termes, en s'appuyant sur un dictionnaire auxiliaire pour augmenter la connectivité du graphe. Tout d'abord, le document est modélisé par un graphe. Ensuite, le graphe est renforcé par un dictionnaire de concepts. Les pondérations des mots sont ensuite calculées en utilisant une fonction à base radiale (ABR). Ceci a permis d'améliorer les performances du système d'indexation. Le k-ppv est utilisé pour la classification. Le rappel et la précision sont adoptés comme métriques d'évaluation. Le tableau 3 illustre les résultats de cette approche en la combinant avec des méthodes d'indexations très utilisées.

Méthode	Corpus	Précision	Rappel
TF IDF	*Sport*	0.83	0.73
	Politique	0.68	0.61
	Economie	0.56	0.71
TF IDF + ABR	*Sport*	0.94	0.77
	Politique	0.78	0.67
	Economie	0.59	0.71

Méthode	Corpus	Précision	Rappel
Okapi	*Sport*	0.82	0.75
	Politique	0.79	0.67
	Economie	0.73	0.65
Okapi + ABR	*Sport*	0.91	0.81
	Politique	0.81	0.70
	Economie	0.76	0.69

Tableau 3. Impact de la méthode ABR combinée avec TFIDF et Okapi (Zaki, Mammass et al, 2014)

D'autres hybridations sont introduites afin d'améliorer les résultats de classification des documents arabes. (Raheel, Dichy *et al.*, 2009) a combiné la méthode de Boosting et l'arbre de décision comme classificateur hybride. Ils ont utilisé les lemmes comme formes de caractéristiques et la TFIDF pour la pondération. Une comparaison de la méthode a été faite avec deux classificateurs, Bayésien naïf (NB) et SVM. Les résultats montrent que SVM et NB surpassent l'approche proposée.

Le modèle SVM a montré des succès remarquables dans la classification de textes (Joachims, 1998). À cet effet, dans (Shafiei, Wang *et al.*, 2007), une méthode hybride utilise le TSVM *(Transductive Support Vector Machines)* et le recuit simulé (SA : *simulated annealing*). Les auteurs ont choisi les deux mille meilleures caractéristiques à partir du test de χ^2 pour former les données de base et ils ont obtenu de meilleurs résultats de classification avec SA par rapport aux SVM et TSVM.

Dans (Mohamed et Watada, 2010), l'analyse sémantique latente (LSA) produit une évaluation de chaque terme dans un document, puis à l'aide du raisonnement probant (ER : *Evidential reasoning*) une catégorie selon la base documentaire est assignée au nouveau document. Des expériences ont été effectuées sur une combinaison de ER avec la LSA et de ER avec TFIDF qui ont montré que ER-LSA est plus performant que ER-TFIDF.

(Zaki, Mammass *et al.*, 2010) étendent le modèle vectoriel en combinant la TF-IDF avec la formule Okapi pour l'extraction des concepts pertinents qui représentent un document. Il propose une nouvelle mesure qui prend en considération la notion de voisinage sémantique en utilisant une mesure de similarité entre termes, et en combinant le calcul du TF-IDF-okapis avec une approche noyau (fonction à base radiale). Cette approche d'indexation permet de valoriser la notion de proximité sémantique. Les résultats expérimentaux confirment l'apport de la contribution.

Chantar *et al.* (Chantar et Corne, 2012) ont proposé une méthode de sélection de caractéristiques appelée Binary Particle Swarm Optimization and K-ppv (OSPB K-ppv). Trois algorithmes d'apprentissage sont utilisés: le SVM, Bayes Naïf et l'arbre de décision C4.5, afin de classer les documents en langue arabe. Les résultats obtenus par SVM ainsi que Naïve Bayes montrent que OSPB K-ppv fonctionne bien en tant que technique de sélection de caractéristiques.

Référence	Linguistique	Statistique	Graphe	Sémantique	Commentaire
(Tomokyo et Hurst 2003)	√	√	x	x	Introduction des notions de grammaticalité et informativité.
(Harrag, El-Qawasmah *et al.*, 2011)	√	√	x	√	Trois techniques de prétraitement sont comparées: light stemming, root-based stemming et dictionary lookup stemming.
(Motasem et Joseph, 2009)	√	x	x	√	Utilisation de l'analyse morphosyntaxique et l'exploitation des règles grammaticales pour la reconnaissance des mots.
(Alwedyan *et al.*, 2011)	√	√	x	x	Introduction de classificateur multi-classes à base de règles d'associations

(Zaki, Mammass et al, 2014)	x	√	√	√	utilisation des méthodes statistiques et des graphes pour l'indexation contextuelle et sémantique des documents arabes
(Raheel, Dichy *et al.*, 2009)	√	√	x	x	Combinaison de la méthode Boosting et l'arbre de décision comme classificateur hybride et TFIDF pour la pondération.
(Shafiei, Wang *et al.*, 2007)	x	√	x	x	La méthode hybride utilisée est le TSVM (Transductive Support Vector Machines) avec (SA : simulated annealing).
(Mohamed et Watada, 2010)	x	√	x	√	Utilisation de l'analyse sémantique latente (LSA) avec (ER : Evidential reasoning). Des expériences ont été effectuées sur une combinaison de ER avec la LSA et de ER avec TFIDF qui ont montré que ER-LSA est plus performant que ER-TFIDF.
(Zaki, Mammass *et al.*, 2010)	x	√	x	√	Combinant du calcul de TFIDF-Okapis avec une fonction à base radiale.
(Chantar et Corne, 2012)	x	√	x	x	Chantar *et al.* ont proposé une méthode de sélection de caractéristiques appelée Binary Particle Swarm Optimization and K-ppv

Tableau 4 : Synthèse des méthodes hybrides.

L'extraction automatique des descripteurs à partir de données textuelles afin de les exploiter implique la mise en place en place d'un moyen de réduction de calcul et d'accélération de traitement, en particulier pour de grandes quantités de données, tout en étant efficace. Ainsi, différentes approches et méthodologies pour la modélisation et la représentation de données textuelles ont été proposées. Nous discuterons dans la section suivante les avantages et les inconvénients de chaque approche.

4. Discussion

Les approches statistiques représentent le texte en sac-de-mots. Cette représentation est adaptée pour capturer les fréquences d'apparition d'un mot et ignore l'information structurelle et sémantique que contient le document.
La représentation fondée sur les graphes est plus adéquate pour modéliser la structure et la sémantique, et son utilisation pour l'extraction de mots clés a prouvé son efficacité (Mihalcea et tarau, 2004). Or, une limite principale de la représentation fondée sur les graphes est la complexité du graphe généré pour chaque document, et le calcul des scores qui augmente d'une façon exponentielle relativement à la taille du document.
Les approches linguistiques, quant à elles, exploitent des règles morphosyntaxiques pour extraire les termes. Ce genre de techniques offre de bons résultats dans des cas spécifiques, en désambigüisation de mots par exemple, mais s'avère moins compétitif pour les systèmes d'indexation, vu la complexité de la langue en question.
Les méthodes utilisant des ressources sémantiques externes (dictionnaire, ontologie ou autres) offrent une meilleure couverture sémantique du document. Sauf que la reconnaissance des unités sémantiques reste limitée au domaine décrit par la ressource utilisée. La génération automatique des dictionnaires à partir du corpus étudié s'avère être une piste prometteuse.
D'autre part, plusieurs chercheurs essaient différentes combinaisons des méthodes classiques, introduisant ainsi des méthodes hybrides de traitement de documents. Ces méthodes permettent non seulement d'augmenter le résultat d'indexation, mais elles deviennent aussi indispensables dans des cas de traitement particulier. Le tableau 5 synthétise les avantages et les limites de chacune de ces approches.
La fouille des documents en langue arabe est confrontée à un autre problème, celui des évaluations des méthodes sur les corpus. Dans la plupart des travaux sur les textes arabes, et en l'absence d'un corpus standard libre, les auteurs construisent leurs propres corpus. Ils choisissent le nombre des catégories et leurs thèmes. Pour chaque catégorie, les documents sont collectés manuellement. Les documents appartenant à plusieurs catégories sont souvent éliminés. Or, pour tester la précision des différentes méthodes, elles doivent être appliquées au même corpus. Plus encore, pour qu'une méthode prouve son efficacité, elle doit être appliquée à plusieurs corpus de taille et de thèmes différents. Un corpus standardisé encouragera donc les auteurs à introduire des nouvelles méthodes, comparer l'efficacité des approches d'une façon objective, et avoir une synthèse plus significative de l'état de l'art.

Approches		Avantages	Inconvénients
Approche linguistique		• Efficacité importante au niveau sémantique • Bon rapport pertinence / représentation	• Laborieuse en cas de grandes masses documentaires • Difficile de prendre en charge toute la complexité de la langue arabe • Relative à la langue
Approche Numérique	***Méthodes statistiques***	• Simples à déployer • Grande progression lors des dernières années • Résultats considérables du point de vue mathématique	• Les mots sont considérés indépendants (sac-de-mot) • Ignorent l'aspect sémantique
	Méthodes basées sur les graphes	• Modélisation sémantique, contextuelle et structurelle	• Graphes complexes dans le cas des textes longs • Temps de calcul élevé
Approche sémantique		• Représentation conceptuelle et sémantique riche • Espace de représentation réduit • Vocabulaire contrôlé	• Limitées au domaine décrit par la ressource sémantique utilisée • Pour la langue arabe, les ressources comme les ontologies et les thésaurus standardisés sont rares
Approches hybrides		• Compromis entre différentes approches	• Accroissement de la complexité par rapport aux systèmes classiques

Tableau 5 : Synthèse des approches et méthodes d'indexation

5. Conclusion

Dans cet article, nous avons présenté différentes techniques d'indexation automatique. Le choix d'une bonne méthode de représentation du document influencera significativement les étapes ultérieures d'analyse de documents. Cependant, plusieurs critères sont mis en question, notamment la réduction de dimension, la conservation du contexte et de la sémantique.

Les auteurs favorisent les méthodes statistiques pour la représentation des documents arabes, considérant la simplicité de leur traitement. Nous recensons peu de travaux sur les documents arabes qui s'intéressent à exploiter de nouvelles méthodes d'extraction de descripteurs, ceci est dû essentiellement à la complexité de la structure de cette langue.

Comme perspective à cette contribution, nous proposerons une nouvelle approche orientée contexte qui tire profit des méthodes classiques, tout en palliant à certaines limites de l'existant. Dans nos futurs travaux, nous souhaitons accorder plus d'importance à la phase d'indexation, et ce, en essayant d'améliorer les méthodes déjà existantes et tenter des hybridations entre différentes techniques. Notre objectif est de proposer un système d'indexation faisant face aux trois enjeux suivants : la sémantique, l'espace de représentation et la complexité de calcul.

6. Références

Al Molijy, A., Hmeidi, I. & Alsmadi, I. (2012) *Indexing of Arabic documents automatically based on lexical analysis*. International Journal on Natural Language Computing (IJNLC) Vol. 1, No.1.

Alajmi, A., Saad, E.M., Darwish, R.R. (2012) *Toward an ARABIC Stop-Words List Generation*. International Journal of Computer Applications (0975-8887) Volume 46- No.8.

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S. & Al-Rajeh, A. (2008). *Automatic Arabic Text Classification*. In Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data, JADT.

Al-Mahmoud, H., Al-Razgan, M. (2015). *Arabic Text Mining: A Systematic Review of the Published Literature 2002-2014*, International Conference on Cloud Computing (ICCC).

Alsaleem, S. (2011). *Automated Arabic Text Categorization Using SVM and NB*. International Arab Journal of e-Technology, vol. 2, no. 2.

Al-Salemi, B. & Aziz, M. J. A. (2011). *Statistical Bayesian Learning For Automatic Arabic Text Categorization*. Journal of Computer Science, vol. 7, no. 1, pages 39–45.

Al-Shalabi, R. & Obeidat, R. (2008). *Improving KNN Arabic Text Classification with N-Grams Based Document Indexing*. In Proceedings of the Sixth International Conference on Informatics and Systems, INFOS, pages 108–112.

Al-Shalabi, R. Kanaan, G. & Gharaibeh, M. (2006). *Arabic Text Categorization Using kNN Algorithm*. In Proceedings of The 4th International Multiconference on Computer Science and Information Technology, volume 4 of CSIT'2006.

Al-Shalabi, R. Kanaan, G. & Gharaibeh, M. (2010). *Arabic Text Categorization Using kNN Algorithm.* In Proceedings of the 6th International Conference on Advanced Information Management and Service, IMS. Institute of Electrical and Electronics Engineers (IEEE).

Alwedyan, J. Musa, W. H., Salam, M. & Mansour, H. Y. (2011). *Categorize arabic data sets using multi-class classification based on association rule approach*. In Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications, ISWSA'11, New York, NY, USA, ACM, pages 18 :1–18 :8.

Bawaneh, M. J. Alkoffash, M. S. & Al Rabea, A. I.. (2008). *Arabic Text Classification using K-NN and Naive Bayes*. Journal of Computer Science, vol. 4, pages 600–605.

Benkhalifa, M. Mouradi, A. & Bouyakhf, H. (2001). *Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization*. International Journal of Intelligent Systems, vol. 16, no. 8, pages 929–947.

Bessou, S., Saadi, A. et Touahria, M. (2007). *Un système d'indexation et de recherche des textes en arabe (SITRA).* 1er séminaire national sur le langage naturel et l'intelligence artificielle (LANIA), Université HAssiba ben Bouali, Département d'Informatique, Chlef (DZ), 20-21.

Bougouin, A. (2013). *État de l'art des méthodes d'extraction automatique de termes-clés.* TALN-RÉCITAL 2013, 17-21 Juin, Les Sables d'Olonne.

Chantar, H. K. & Corne D. W. (2012). *Arabic Text Categorization via Binary Particle Swarm Optimization and Support Vector Machines.* In The 5th International Conference on Bioinspired Optimization Methods and their Applications, BIOMA'2012.

Duwairi, R. M. (2005). *A Distance-based Classifier for Arabic Text Categorization.* In In Proceedings of The 2005 International Conference on Data Mining, DMIN'2005, CSREA Press, pages 187–192.

Duwairi, R. M. (2006). *Machine learning for Arabic text categorization: Research Articles*. Journal of American society for Information Science and Technology, vol. 57, no. 8, pages 1005–1010.

El Kourdi, M., Bensaid, A. & Rachidi, T. (2006). *Automatic Arabic document categorization based on the Naive Bayes algorithm*. In Proceedings of the Workshop on Computational Approaches to Arabic Script based Languages, SEMITIC '04, Stroudsburg, PA, USA. Association for Computational Linguistics, pages 51–58.

El-Halees, A. M. (2007). *Arabic Text Classification Using Maximum Entropy.* The Islamic University Journal (Series of Natural Studies and Engineering), vol. 15, no. 1, pages 157–167.

El-Khoribi, R. A. and. Ismael, M. A (2006). *An Intelligent System Based on Statistical Learning For Searching in Arabic Text.* ICGST International Journal on Artificial Intelligence and Machine Learning, AIML, vol. 6, pages 41–47.

El-Kourdi, M. Bensaid, A. & Rachidi, T. (2010). A*utomatic Arabic Document Categorization Based on the Naïve Bayes Algorithm*. In Proceedings of The 7th International Conference on Informatics and Systems, INFOS.

Gharib, T. F. Habib, M. B. & Fayed, Z. T. (2009). *Arabic Text Classification Using Support Vector Machines*. International Journal of Computers and Their Applications ISCA, vol. 16, no. 4, pages 192–199.

Harrag, F., El-Qawasmah, E. & Al-Salman, A. (2011). *Stemming as a Feature Reduction Technique for Arabic Text Categorization*. In Proceedings of The 10th International Symposium on Programming and Systems, ISPS, pages 128–133.

Hasan, K.S. & Ng, V. (2010). *Conundrums in unsupervised keyphrases extraction : Making sense of the state of the art.* Proceedings of the 23rd International Conference on Computatioonal Linguistics (COLING-10), Poster Volume.

Herskovic, J. R. & Jorge, R. (2011). *MEDRank: Using graph-based concept ranking to index biomedical texts.* International Journal of Medical Informatics volume 80 issue 6 pages : 431-441.

Hmeidi, I., Hawashin, B., El-Qawasmeh, E. (2008). *Performance of KNN and SVM classifiers on full word Arabic articles.* Advanced Engineering Informatics 22, pages 106–111.

Hofmann, T. (1999). *Probabilistic latent semantic indexing.* SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval Pages 50-57.

Hulth, A. (2003). *Improved automatic keyword ex- traction given more linguistic knowledge.* In Proceedings of EMNLP, pages 216–223.

Jamoussi, S. (2009). *Une nouvelle représentation vectorielle pour la classification sémantique*. TAL volume 50.

Joachims, T. (1998). *Text Categorization with Suport Vector Machines : Learning with Many Relevant Features*. In Proceedings of the 10th European Conference on Machine Learning, ECML'98, London, UK, UK. Springer-Verlag, pages 137–142.

Kanaan, G., Yaseen, M., Al-Shalabi, R., Al-Sarayreh, B. & Mustafa, A.. (2009). *Using EM for Text Classification on Arabic.* In Proceedings of the Second International Conference on Arabic Language Resources and Tools. The MEDAR Consortium.

Kanaan, G., Al-Shalabi, R. & AL-Akhras, A. (2006). *KNN Arabic Text Categorization Using IG Feature Selection.* In Proceedings of The 4th International Multiconference on Computer Science and Information Technology, volume 4 of CSIT'2006.

Khreisat, L. (2006). *Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study*. In Proceedings of The 2006 International Conference on Data Mining, DMIN '2006, CSREA Press, pages 78–82.

Li, H. Y. & Jain, K. A. (1998). *Classification of text documents*. The Computer Journal, vol. 41, no. 8, pages 537–546.

Liu, Z. , Peng, L. , Yabin, Z. & Maosong, S. (2009). *Clustering to find exemplar terms for keyphrase extraction*. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 257–266.

Mallak, I. (2011). *De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information.* Thèse pour l'obtention du grade de docteur. Université Paul Sabatier – Toulouse III. France.

Mansour, N., Haraty, R.A., Daher, W., Houri, M. (2008). *An auto-indexing method for Arabic text*. Information Processing and Management, volume: 44 issue: 4, pages: 1538-154.

Matsuo, Y et Ishizuka, M. (2004). *Keyword Extraction From a Single Document Using Word Co-Occurrence Statistical Information*. International Journal on Artificial Intelligence Tools volume 13 issue 1 pages: 157-169.

Mesleh, A. M. & Kanaan, G. (2008). *Support vector machine text classification system: Using Ant Colony Optimization based feature subset selection*. In proceeding of the International Conference on Computer Engineering & Systems, ICCES '2008, pages 143–148.

Mesleh, A. M. (2007b). *CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System.* Journal of Computer Science, vol. 3, no. 6, pages 430–435.

Mesleh, A. M. (2007c). *CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System.* In proceeding of the 2nd International Conference on Software and Data Technologies, (Knowledge Engineering), pages 235–240.

Mesleh, A. (2007a). *Support vector machines based Arabic language text classification system : feature selection comparative study.* In Proceedings of the 12th WSEAS International Conference on Applied Mathematics, MATHq07, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS), pages 11–16.

Mihalcea, R. & Tarau, P. (2004). *Textrank: Bring- ing order into texts*. In Proceedings of EMNLP, pages 404–411.

Mohamed, R. & Watada, J. (2010). *An Evidential Reasoning Based LSA Approach to Document Classification for Knowledge Acquisition.* In Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, IEEM'10. Institute of Electrical and Electronics Engineers (IEEE), pages 1092–1096.

Motasem, A. & Joseph, D. (2009). *Levée d'ambigüité par la méthode d'exploration contextuelle: la séquence 'alif-nûn* (ان) *en arabe*. In Ghenima, Malek, Ouksel, Aris et Sidhom, Sahbi (eds.), Systèmes d'Information et Intelligence Economique, 2ème Conférence Internationale, organisée par l'université de Nancy, France et l'université de la Manouba, École supérieure de commerce électronique (ESCE), Tunis, Tunis, Hammamet, IHE éditions, pages. 573-585.

Mountassir, A. (2012). *Sentiment Analysis: Classification supervisée de documents arabes.* Proceedings of 7th International Conference on Intelligent Systems : Theories and Applications. Mohammedia, Morocco.

Page, L. Brin, L. Motwanin R., & Winograd, T. (1998). *The pagerank citation ranking: Bringing order to the web.* Technical report, Stanford Digital Library Technologies Project, 1998.

Porter, M.F. (1980). *An algorithm for suffix stripping*. Program, Vol. 14 No.3, pp. 130-137.

Raheel, S. & Dichy, J. (2010). *An empirical study on the feature qs type effect on the automatic classification of arabic documents.* In Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing, CICLing'10 Berlin, Heidelberg, pages 673–686.

Raheel, S., Dichy, J. & Hassoun, M. (2009). *The Automatic Categorization of Arabic Documents by Boosting Decision Trees*. In Proceedings of the Fifth International Conference on Signal Image Technology and Internet Based Systems Washington, DC, USA. IEEE Computer Society, pages 294–301.

Robertson, S.E., Walker, S. & Beaulieu, M. (2000). *Experimentation as a way of life: Okapi at TREC.* Information Processing and Management, vol. 36, pages 95–108.

Roche, M. (2011). *Fouille de Textes : De l'extraction des descripteurs linguistiques à leur induction.* Thèse, université Montpellier II, France.

Saadane, H. (2013). *Une approche linguistique pour l'extraction des connaissances dans un texte arabe.* TALN-Récital, 17-21 juin, Les Sables d'Olonne.

Salton, G., Wong, A. & Yang, C. S. (1975). *A vector space model for automatic indexing. Commun*. ACM, vol. 18, no. 11, pages 613–620.

Samir, A. M., Ata, W. & Darwish, N. (2005). *A New Technique for Automatic Text categorization for Arabic Documents*. In Proceedings of the 5th Conference of the Internet and Information Technology in Modern Organizations, pages 13–15.

Sawaf, H., Zaplo, J. & Ney, H. (2001). *Statistical Classification Methods for Arabic News Articles*. In Arabic Natural Language Processing Workshop, ACL'2001, Retrieved from Arabic NLP Workshop at ACL/EACL 2001 website : `http ://www.elsnet.org/acl2001-arabic.html`, pages 78–82.

Schapire, R. E. & Singer, Y. (2000). *BoosTexter: A Boosting-based System for Text Categorization*. Machine Learning, vol. 39, no. 2/3, pages 135–168.

Schapire, R. E. Singer, Y. & Singhal, A. (1998). *Boosting and Rocchio applied to text filtering.* In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98, New York, NY, USA. ACM, pages 215–223.

Shafiei, M., Wang, S., Zhang, R., Milios, E., Tang, B., Tougas, J. & Spiteri, R. (2007). *Document Representation and Dimension Reduction for Text Clustering.* In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop, ICDEW q07, Washington, DC, USA. IEEE Computer Society, pages 770–779.

Stetina J., Kurohashi S. et Nagao. M. (1998). General word sense disambiguation method based on a full sentential context. Usage of WordNet in Natural Language Processing, Proceedings of COLING-ACL Workshop, Montreal, Canada, July 1998

Syiam, M. M., Fayed, Z. T. & Habib, M. B. (2006). *An Intelligent System For Arabic Text Categorization.* International Journal of Intelligent Computing and Information Sciences, vol. 6, no. 1, pages 1–19.

Thabtah, F., Eljinini, M., Zamzeer, M. & Hadi, W. (2009). *Naïve Bayesian based on Chi Square to Categorize Arabic Data*. In proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, IBIMA'2009, pages 930–935.

Thabtah, F., Hadi, W. & Al-shammare, G. (2008). *VSMs with K-Nearest Neighbour to Categorise Arabic Text Data*. In Proceedings of The World Congress on Engineering and Computer Science, WCECS '2008, pages 778–781.

Tomokiyo T. et Hurst M.. (2003). A language model approach to keyphrase extraction. In Proceedings of the ACL Workshop on Multiword Expressions.

Trstenjak, B., Mikac, S., Donko, D. (2013). *KNN with TF-IDF Based Framework for Text Categorization*. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation.

Wan, W et Xiao, J. (2008a). Collabrank: Towards a collaborative approach to single-document keyphrase extraction. In Proceedings of COLING, pages 969–976.

Wan, X et Xiao, J. (2008b). *Single document keyphrase extraction using neighborhood knowledge*. In Proceedings of AAAI, pages 855–860.

Wei, G., Gao, X.and Wu, S. (2010). *Study of Text Classification Methods for Data Sets With Huge Features*. In Proceedings of the 2nd International Conference on Industrial and Information Systems, volume 1, pages 433–436.

Wei,W. (2013). Regroupement sémantique de relations pour l'extraction d'information non supervisée. TALN-RÉCITAL, Les Sables d'Olonne.

Yang, Y. & Chute, G. C. (1994). *An example-based mapping method for text categorization and retrieval.* ACM Transactions on Information Systems, vol. 12, no. 3, pages 252–277.

Yongjing, L. (2007). *A Document Clustering and Ranking System for Exploring MEDLINE Citations*. Journal of the American Medical Informatics Association volume 14 issue 5, pages: 651-661.

Zaki, T. (2013). *Indexation par le contenu et archivage de fonds documentaires arabes*. Thèse pour l'obtention du grade de doctorat d'université, Université Ibn Zohr, Agadir, Maroc.

Zaki, T. Mammass, D. Ennaji A., Nicolas, S. (2014). *A kernel hybridization NGram-Okapi for indexing and classification of Arabic documents*. Journal of information and Computing Science. ISSN 1746-7659, England, UK. Vol. 9 No.2. pages:141-153.

Zaki, T. Mammass, D. Ennaji, A. (2010). *A semantic proximity based system of Arabic text indexation*. International Conference on Image and Signal Processing (ICISP).

Zargayouna, H. et Salotti, S. (2004). *Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML*. IC: 15es journées francophones d'ingénierie des connaissances.

Zubi, Z. S. (2009). *Using some web content mining techniques for Arabic text classification*. In Proceedings of the 8th WSEAS international conference on Data networks, communications, computers, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society, pages 73–84.

Internet of Things Security: Layered classification of attacks and possible Countermeasures

La sécurité de l'Internet des objets : classification en couches des attaques et contre-mesures possibles

Otmane El Mouaatamid
SIME Lab, ENSIAS, Rabat, Morocco
otmane_elmouaatamid@um5.ac.ma

Mohammed Lahmer
SIME Lab, EST My Ismail University, Meknes, Morocco
mohammed.lahmer@gmail.com

Mostafa Belkasmi
SIME Lab, ENSIAS, Rabat, Morocco
m.belkasmi@um5s.net.ma

Résumé

L'internet des objets (IdO) est un domaine actif de recherche. Assurer la sécurité des données échangées figure parmi ses grands défis. Cet article propose une nouvelle classification des attaques selon les couches OSI et l'objectif de sécurité à atteindre afin de développer de nouvelles techniques et procédures pour lutter contre ces attaques.

Abstract

Internet of Things is undoubtedly a well-known research area. In fact, ensuring security of data exchange is among the great challenges of the Internet of things. In this paper, we endeavour to introduce a new classification of attacks in compliance with the OSI layers and the objective of security that we seek to attained in order to develop novel techniques and processes to fight against these attacks.

Mots-clés

Internet des Objets, WSN, RFID, Sécurité, Attaques, Contre-mesures.

Keywords

Internet of Things, WSN, RFID, Security, Attacks, Countermeasures.

1. Introduction

The term Internet of Things was first coined by (Ashton, 1999) which is a technological revolution that represents the future of computing and communications. Its development depends on a dynamic technical innovation in a number of important fields, from wireless sensors to the nanotechnology based architecture (Akyildiz *et al.*, 2002); (Awerbuch and Scheideler, 2004); (Chaczko *et al.*, 2015). Today, we find this kind of technology in a wide range of potential applications, including smart city, control actuation and maintenance of complex systems in industry field, health, and transport. The IoT touches every facet of our lives. Security and privacy are two of the most crucial challenges that IoT is facing (FTC Sta¬ Report, 2015). Since sensor networks are highly vulnerable against attacks (Deng *et al.*, 2005), it is very important to have some mechanisms that can protect the network, devices, and users from all kinds of attack. It must be certain that the system is protected from any kind of attacks.

RFID (Radio Frequency Identification) and WSN (Wireless Sensor Networks) are two technologies used by IoT. Combining both RFID and WSN is of paramount importance as they can add additional services to each other. For instance, the identification of location can be performed using the RFID, whereas, WSN can be used to sense the objects surrounding environment. Many applications can get benefits from the integration of these two technologies, such as healthcare system and food chain tracking. But this combination can lead to multiple vulnerabilities that can jeopardize the benefit of these two technologies. Thus, security of Internet of Things is of paramount importance. The existing works in the literature focus only on the RFID or WSN. For instance, (Mitrokotsa *et al.*, 2010) give a classification of attacks for only RFID systems. Indeed, their classification is based only on the OSI layers whereas in our classification we classify the attacks based on both the security goals and attacks targeting each OSI layer. The authors (Sadeghi *et al.*, 2012) focus only on attacks that target the network layer in WSN.

In this paper, our contribution consists of classifying both WSN and RFID attacks and suggesting some countermeasures for these attacks. Our classification is based on both security requirements such as privacy, confidentiality, non-repudiation and the threats which seek a specific OSI layer. To the best of our knowledge, none of the existing papers address the attacks and countermeasures of both WSN and RFID according to the goal of security and the OSI layer. They tackle only one of them and they focus on attacks more than giving a detailed description of possible solutions.

The remainder of this paper is organized as follows. In the first section, we present an overview about Internet of things and their technologies, application areas, architecture and standardization. In the second section, we present features and goals of security for IoT. In the third section, we classify some WSN and RFID based on IoT attacks into three categories : Denial of service (DoS), Privacy, and Impersonation. Some attacks target both WSN and RFID. The last section suggests countermeasures for these attacks.

2. Internet of Things and Security

The CERP-IoT (Cluster of European Research projects on the Internet of Things) defines the Internet of things, such as a dynamic global network infrastructure with self-configuring capabilities based on standard and interoperable communication protocols where physical and virtual things have identities, physical attributes, and virtual personalities, use intelligent interfaces, and are seamlessly integrated into the information network (Sundmaeker *et al.*, 2010). This vision of the IoT will introduce a new dimension to the information and communication technologies. In addition to the two temporal and spatial dimensions that allow people to connect from anywhere at any time, we will have a new "object" dimension that will allow them to connect to anything. The IoT will cover a wide range of applications and almost touch all areas that we face every day. This will allow the emergence of smart spaces around a ubiquitous computing. These smart spaces include: cities, energy, transport, health, industry, and agriculture, etc. (Mitchell *et al.*, 2013).

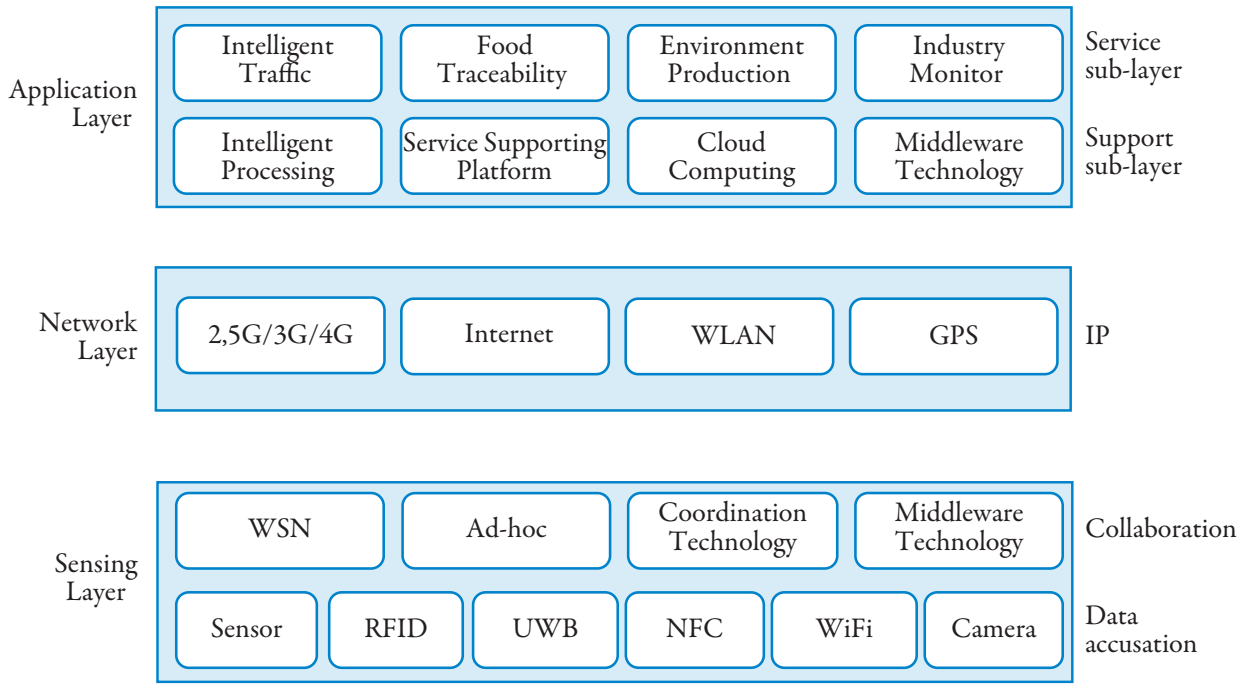

Figure 1. The IoT architecture model.

2.1. IoT Architecture

The IoT is characterized by a comprehensive perception, a reliable transmission and intelligent processing. Figure.1 shows the three-layer architecture of IoT : applications, network and sensing layer. The sensing realizes a comprehensive perception by collecting real-time dynamic data through various sensors (including tags) while the network layer is mainly responsible for the reliable data transmission, relaying data acquired from the sensing layer to the application layer. Using distributed computing technologies, including cloud computing, the application layer performs massive data processing and intelligent analysis for the purpose of intelligent control (Zheng *et al.*, 2011).

2.1.1. Internet of Things : Business scenarios

IoT systems are affecting all matters of our everyday life (Mitchell *et al.*, 2013). In fact, The IoT is making the whole world one place where everyone is interacting with one another. An example for such interaction is the physical entities that could exist in many different social environments (work, family, individual, leisure, etc.), which make determine clear boundaries difficult as shown in figure 2.

In order to show the impact of IoT, we present, thereafter, some scenarios where IoT technologies have a special relevance, taking into account that these scenarios frequently share the same applications, sensors, devices, and most certainly, users. These scenarios have been provided by the Internet of things Architecture (IoT-A) (Walewski *et al.*, 2011).

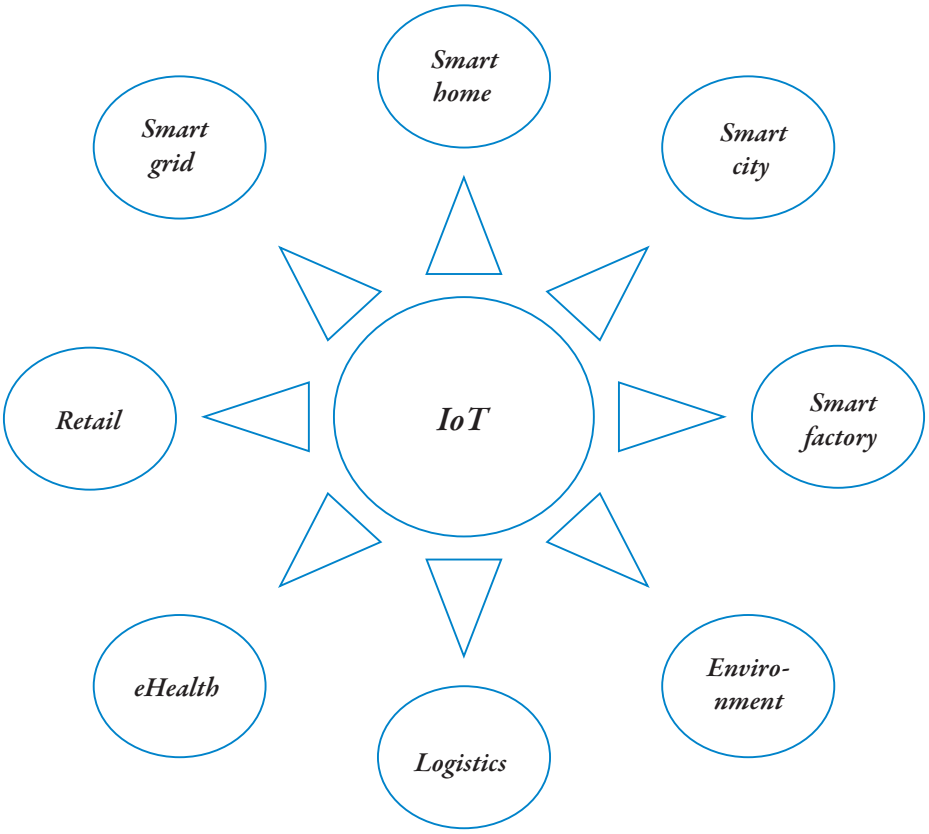

Figure 2. IoT business solutions

Smart home: Future smart homes will be conscious about what happens inside a building, mainly impacting three aspects: resource usage (water conservation and energy consumption), security, and comfort. The goal with all these is to achieve better levels of comfort while cutting overall expenditure. Moreover, smart homes address security issues by means of complex security systems to detect theft, fire, or unauthorized entries. The stakeholders involved in this scenario constitute a very heterogeneous group. There are different actors that will cooperate in the user's home, such as Internet companies, device manufacturers, telecommunications operators, media service providers, security companies, electric-utility companies, etc.

Smart city: A smart city can be defined as an urban community within which citizens, organizations and governing institutions deploy ICT to transform their locality in a significant way (Deakin, 2013). A smart city enables to implement a management infrastructure (water, energy, information and telecommunication, transport,

emergency services, public facilities, buildings, management and sorting waste, etc.). Likewise, a smart city is communicating, adaptable, sustainable, effective, eco-friendly, and ultimately automated to improve the quality of life of citizens.

Smart factory: The IoT is integrated into the objects of every day. It is the trend that is going to expand in the future. In this context, the IoT will allow companies to track all their products by means of RFID tags as they move through the global supply chain. As a result, companies will be able to reduce their Operational Expenditure (OPEX) and enhance their productivity. Hence, IoT will allow to automate procedures. As a consequence, the number of employees will be reduced. Workers will be replaced by complex robots, as efficient as humans. At the same time, these technologies will create new job opportunities for a big number of technicians to program and repair these machines.

Smart grids: Smart grids are fluid distribution materials networks (electricity, water, gas, oil ...) and / or information (telecommunications) that have been "augmented" (rendered intelligent) by computer systems, sensors, computer and electromechanical interfaces giving them a two-way exchange capacity and sometimes some capacity for autonomy in computing and materials flow management and information processing.

Environment: Smart environments are environments where sensors and actuators are integrated to react to events and to adapt to those present. For example, a smart home can adjust the temperature and lighting based on health, mood, and preferences of people and animals inside each piece.

Transportation/ Logistics: Intelligent Transportation Systems (ITS) are the applications of new information and communications technology in transport. They are called "intelligent" because their development is based on the functions usually associated with intelligence sensory, memory, communication, information processing and adaptive behavior. ITS are found in several areas of activity, such as in optimizing the use of transport infrastructure, in improving safety (including road safety) and security and in the development of services.

eHealth: Control and prevention are two of the main objectives of the future health care. Today, people already can have the opportunity to be followed and monitored by specialists, even if the two are not in the same location. Tracing people's health history is another aspect that makes IoT-assisted Health very versatile. Business applications could offer the possibility of a medical service, not only for patients, but also specialists who need information to perform their medical evaluation. In this field, IoT makes human interaction much more effective because doesn't only allow the localization, but also tracking and monitoring of patients. Providing information on the State of a patient makes the process more efficient, and also make people much more satisfied

Retail: IoT realizes the needs of customers and the needs of businesses. The comparison of a product price, or searching other products of the same quality at lower prices or shop promotions gives not only information to customers, but also businesses and affairs. Having this information in real time helps companies to improve their business and meet the needs of customers.

IoT is based on several technologies such as RFID, Near Field Communication (NFC), Sensors and Actuators Wireless Network (WSN), Machine-to-Machine communications (M2M), 3G/4G, IPv6 and 6LoWPAN. All of them play an important role in the development of IoT. In the remainder of our study we will be limited to RFID and WSN (Clauberg, 2004).

2.1.2. Wireless Sensor Networks (WSN)

WSN are structures of independent nodes whose wireless communication takes place over limited bandwidth and frequency. The nodes of wireless sensor networks are made up of following parts, such as Sensor, Microcontroller, Battery, radio Transceiver and Memory. Because of the limited communication range of each WSN sensor node, multi hop relay of information takes place between the source and the base station. The communication networks are dynamically formed by the use of wireless radio transceivers that facilitates data transmission between nodes.

2.1.3. Radio Frequency Identification (RFID)

In situation to the IoT, RFID is a method for storing and retrieving data remotely using markers called (RFID tag) or (RFID transponder). RFID system activated by a transfer of electromagnetic energy consists of the following two components: RFID tags and RFID readers.

RFID tags (Transponders): Radio-frequency identification system uses tags, or labels attached to the objects to be identified. Two-way radio transmitter-receivers called interrogators or readers send a signal to the tag and read its response. The RFID tag is also made up of memory units, which houses a unique identifier known as Electronic Product Code (EPC). As described in [14], the classification of the RFID tags types are active and passive :

- ***Active tag***: An active tag has an on-board battery and periodically transmits its ID signal.
- ***Passive tag***: A passive tag is cheaper and smaller because it has no battery; instead, the tag uses the radio energy transmitted by the reader.

RFID readers (Transceivers): A radio frequency identification reader (RFID reader) is a device used to gather information from a RFID tag, which is used to track individual objects (Zharinov *et al.*, 2014).

2.2. The Security Features

As wireless networks become ubiquitous and their security becomes an important design of a secure solution that should meet some basic and significant requirements. We primarily focus on security requirements, and then we address the main security issues in order to ensure the deployment of a secure IoT.

2.2.1. Security concepts

The term security subsumes a wide range of different concepts (Borgohain *et al.*, 2015); (Garcia-Morchon *et al.*, 2013); (Veríssimo and Rodrigues, 2001). In the first place, it refers to the basic provision of security services including:

- ***Authentication***: The process of determining whether someone or something is, in fact, who or what it is declared to be. We distinguish two kind of attacks related to authentication namely, impersonation attack where an attacker pretends to be another entity, and Sybil attack where the attacker uses different identities at the same time.
- ***Authorization***: The process of giving someone permission to do or have something.
- ***Integrity***: Set of means and techniques to restrict the modification of data to authorized persons. Attacks related to data integrity are message alteration attack and message fabrication attack.
- ***Confidentiality***: Concept to ensure that information can only be read by authorized persons. Attacks on confidentiality consist of accessing illegally to confidential data.
- ***Non-repudiation***: Set of means and techniques to prove the involvement of an entity in a data exchange. Attacks on non-repudiation consist of a denial of participation in all or part of communications.
- ***Availability***: the objective is to guarantee the survivability of network services against Denial-of-Service attacks. The attack aiming at an aggregator can make some part of the network losses its availability because the aggregator is responsible to provide the measurement of that network part.
- ***Privacy***: The objective of this security requirement is to prevent private information from being leaked to malicious entities. Attacks on privacy are related to illegally gathering sensitive information about entities (e.g., eavesdropping).

2.2.2. Security concerns in IoT

Privacy for IoT: As much of the information in an IoT system may be personal data, there is a requirement to support anonymity and restrictive handling of personal information.
There are a number of areas where advances are required (Weber, 2010); (Mattern and Floerkemeier, 2010).

- Cryptographic techniques that enable protected data to be stored, processed and shared, without the information content being accessible to other parties.
- Techniques to support Privacy by Design concepts, including data minimization, identification, authentication and anonymity.

And there are a number of privacy implications arising from the ubiquity and pervasiveness of IoT devices where further research is required, including:

- Preserving location privacy, where location can be inferred from things associated with people.
- Prevention of personal information inference, that individuals would wish to keep private, through the observation of IoT related exchanges.
- Keeping information as local as possible using decentralized computing and key management.

3. Classification of Attacks on IoT

3.1. Types of attacks

We can classify generally five types of security attacks, namely Physical attacks, Side channel attacks, Cryptanalysis attacks, Software attacks and Network Attacks (Babar *et al.*, 2011).

Physical attacks: These types of attacks tamper with the hardware components and are relatively harder to perform because they requires an expensive material. Some examples are de-packaging of chip, layout reconstruction, micro-probing, particle beam techniques, etc.

Side channel attacks: These attacks are based on a side channel Information that can be retrieved from the encryption device that is neither the plaintext to be encrypted nor the cipher text resulting from the encryption process. Encryption devices produce timing information that is easily measurable, radiation of various sorts, power consumption statistics, and more. Side channel attacks makes use of some or all of this information to recover the key the device is using. It is based on the fact that logic operations have physical characteristics that

depend on the input data. Examples of side channel attacks are timing attacks, power analysis attacks, fault analysis attacks, electromagnetic attacks and environmental attacks.

Cryptanalysis attacks: These attacks are focused on the ciphertext and they try to break the encryption, i.e. find the encryption key to obtain the plaintext. Examples of cryptanalysis attacks include ciphertext only attack, known-plaintext attack, chosen-plaintext attack, man-in-the-middle attack, etc.

Software attacks: Software attacks are the major source of security vulnerabilities in any system. Software attacks exploit implementation vulnerabilities in the system through its own communication interface. This kind of attack includes exploiting buffer overflows and using Trojan horse programs, worms or viruses to deliberately inject malicious code into the system. Jamming attack is the one of the ruinous invasion which blocks the channel by introducing larger amount of noise packets in a network. Jamming is the biggest threat to IoT where a network consists of small nodes with limited energy and computing resources. So it is very difficult to adopt the conventional anti jamming methods to implement over IoT technologies.

Network Attacks: Wireless communications systems are vulnerable to network security attacks due to the broadcast nature of the transmission medium. Basically attacks are classified as active and passive attacks. Examples of passive attacks include monitor and eavesdropping, Traffic analysis, camouflage adversaries, etc. Examples of active attacks include denial of service attacks, node subversion, node malfunction, node capture, node outage, message corruption, false node, and routing attacks, etc.

3.2. Classification of attacks on WSN and RFID

In this section, we classify attacks of WSN and RFID based on the layer that each attack is taking place, giving special characteristics (Figure 3 and 4). We discriminate attacks that are deployed in the physical-link layer, the network-transport layer and the application layer, as well as multilayer attacks, which affect more than one layer and in the last we suggest new classification in which, attacks are sorted based on the target of the attacker. For example, many attacks are designed for destroying the signal while some others are targeting the privacy issues. Based on this view, we classify attacks in three main categories: Denial of Service (DoS), Privacy, and Impersonation as shown in (Table 1).

3.2.1. Layered classification of attacks on the WSN

As summarized in Figure 3 and mentioned in the previous paragraph, there are several varieties of possible attacks in WSN that we have classified depending on which layer the attack happens.

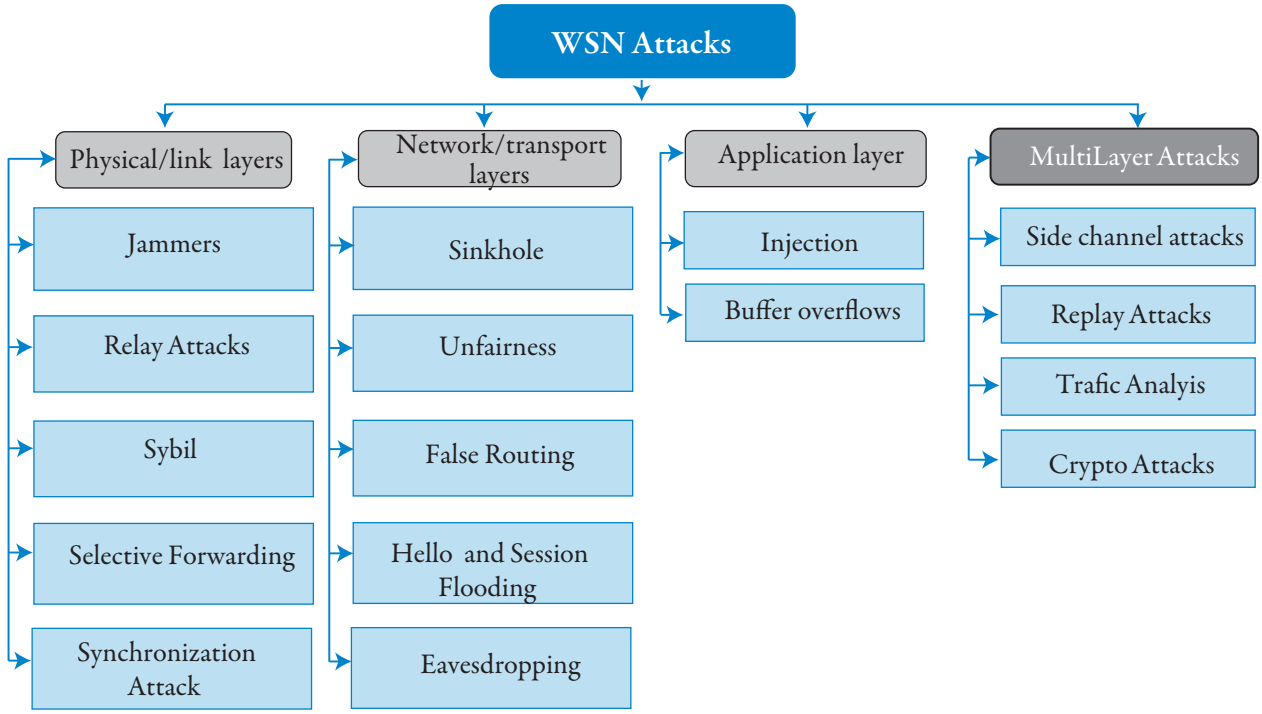

Figure 3. Layered classification of WSN attacks.

Hence, this classification has allowed us to easily locate each attack and then tackle the security issues according to the actions performed by the attacker. The attacker could be either an active attacker by performing an action that could jeopardize the benefit of the WSN, or a passive attacker whose objective is to eavesdrop the network. In this context, numerous techniques and tools have been developed to deal with WSN security attacks. The most existing attacks and vulnerabilities in WSN will be detailed later, whereas, in the last section, we will suggest some countermeasures against these attacks (Karlof and Wagner, 2003).

3.2.2. *Layered classification of attacks on the RFID*

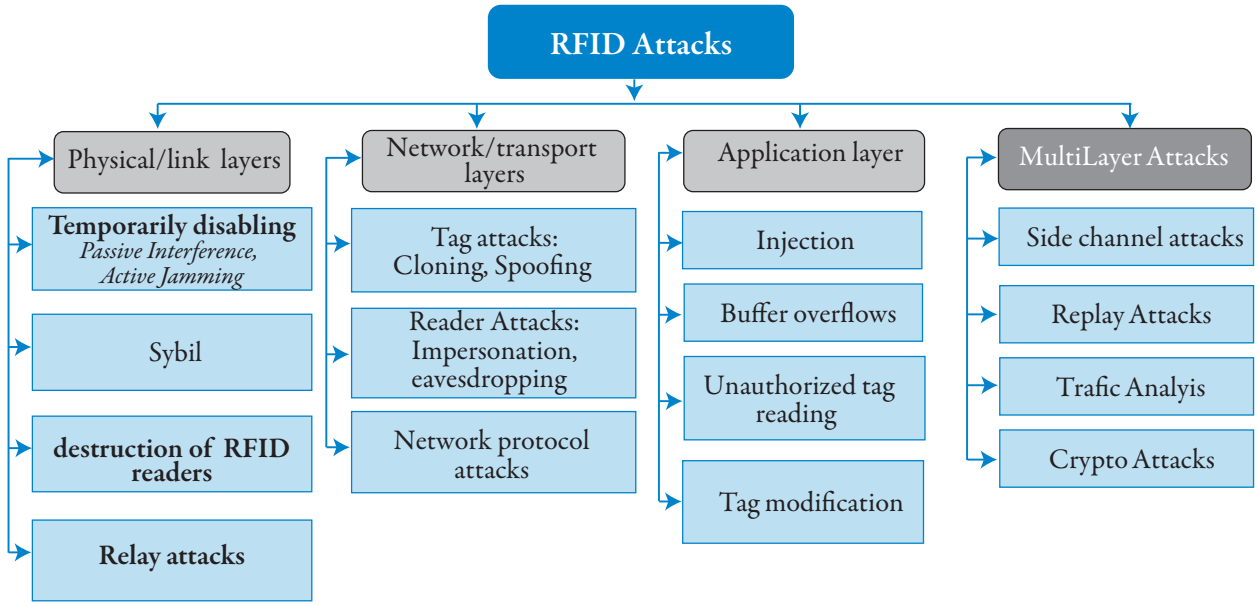

Figure 4. Layered classification of RFID attacks.

Despite the facilities it offers, the wireless medium used in RFID network has some drawbacks that leave it vulnerable to different types of attacks that target this type of transmission medium. We classified these attacks based on the layer where each attack could be performed. The Figure 4 represents a classification of RFID network attacks. As mentioned above, we discriminate attacks that could be deployed to physical layer, network-transport layer and the application layer, as well as multilayer attacks, which affect more than one layer. According to the functionalities and features of each layer, an attacker chooses a specific attack to carry out. Among these attacks we point out the relay attacks, destruction of RFID readers, Sybil attack and the temporarily disabling passive interference, active jamming… as security risks that could be faced on the physical/link layer. Regarding the threats associated to the network/transport layer we find the tag attacks such as cloning and spoofing, the reader attacks like impersonation, eavesdropping and the network protocol attacks. As to application layer several attacks can be considered such as injection, buffer overflows, unauthorized tag reading (Rieback *et al.*, 2006).

3.2.3. *Goal based Classification of WSN and RFID*

We can classify generally three types of security requirements in WSN and RFID according to the Goal (Table 1).

3.2.4. *Denial of Service (DoS)*

As mentioned in Table 1, there are four different ways of denying a service in the WSNs and RFID systems. There is a group of attacks called Jammers which try to reshape signal or change few bits of the packet by making interferences during communication. There is a group of attacks called network congestion which try to make network congested. There is a group of attacks called packet dropping, the goal of these attacks is dropping or discarding the packets, and there are attacks called network consumption, these attacks are specifically designed for draining the nodes energy (Ghildiyal et al, 2014)

		Technologies / Attacks	WSN	RFID
Denial of Service	*Jammers*	Physical Layer Jammers	✓	
		Link Layer Jamming	✓	
	Network Congestion	Unfairness	✓	
		Spoofing	✓	
		Wormhole	✓	✓
		Unauthorized Tag Reading		
		Sinkhole	✓	
		False Routing	✓	
		Unauthorizedtag reading		✓
	Packet Dropping	Selective Forwarding	✓	
		Synchronization	✓	✓
	Energy Consumption	Hello Packet And Session Flooding	✓	
Privacy	*Data oriented*	Eavesdropping	✓	✓
		Skimming		✓
		Substitution		✓
		Clonage		✓
		Replay	✓	✓
	Context oriented	Traffic Analysis	✓	✓
		Tempering Attacks	✓	✓
		Tag modification		✓
Impersonation	*Physical node*	Physical Layer Identification	✓	✓
		Spoofing	✓	
	Virtual node	Sybil	✓	✓

Table 1. Goal based classification of WSN and RFID

Jammers

Jammers are one of the oldest and famous attacks in WSN. Jamming can happen in the deferent OSI layers. Note that jamming in each layer means targeting the specific packets related to that layer, e.g. ACK packets in layer two.

Physical Layer Jammers: The main target of these attacks is radio signal which is jammed with Radio Frequency (RF) transmitter. Because in RFID system the communication media is shared between the Tags /Reader and in WSN the communication is shared between the nodes, adversaries have a great chance to interfere and deny the service. There are three techniques for physical layer jamming (Xu et al, 2006), (Constant Jamming) where attacker sends nonstop random bits, (Deceptive Jamming) whose main target is to send continuous stream of regular packets. Attacker can also send the jamming signal in a random periodic format to save the energy of the jamming device (Random jamming). All The three techniques mentioned above are considered as active jamming, because, the jammer can be detected.

Link Layer Jamming: Link layer jammers are complicated and energy inefficient compared to physical layer jammers. The target of this attack is data packets whereas in physical layer the target is just any packet. As described by (Law et al, 2005) this attack in link layer is harder to detect. Note that, the link layer jamming might also focus on the controlling signal such as ACK message. These specific jammers are called collision makers. Link layer jammer tries to jam the data packets. Since different types of MAC protocol exist, the jammer should jam based on the type of the MAC protocol in WSN. The challenge for link layer jamming is in predicting the arrival of the data packets. Jamming for different MAC protocols has been proposed by (Law et al, 2005). This attack is known in WSN environment but in RFID system we just talk about an active jammer in physical layer.

Network Congestion

The main objective of this group attacks is to create the delay in delivering the data. These groups of attacks include all the attacks that are based on the way RFID systems and WSN are communicating and the way those data are transferred between the entities of an RFID network (tags, readers) or between nodes of a WSN. These attacks pose a major threat to networks in which the data freshness is playing an important role. Below, we review the main congestion makers.

Unfairness: unfairness is a repeated collision based on attack. It can also be referred to as exhaustion based attacks by (Ghildiyal et al, 2014). This type of attacks is famous in WSN, but unknown on the RFID system.

Wormhole: The wormhole attack is independent of Link layer protocols as it is considered dangerous and the attacker do not need to understand the link layer Protocol or be able to decode encrypted packets. Wormhole could be performed at the bits level or at the physical layer. Wormhole is a low latency connection (tunneling) between two adversary nodes, geographically located in different parts of WSN (Tayebi et al, 2013).

Sinkhole: It is a special kind of the group network congestion attacks (Tayebi et al, 2013). A compromised node tries to draw all or as much traffic as possible from a particular array, by giving itself more attraction to the surrounding nodes with respect to the routing metric.

False Routing: The main objective of this attack is that the adversary node tries to make and propagate false routing information. There are different ways to perform this attack. This attack is famous in network layer as described (Ghildiyal et al, 2014) there are four ways to implement this attack: Overflowing routing table with Nonexistence routes, poisoning either routing table or routing cache (this way only applicable for on-demand routing protocols), and finally rushing attack.

Dropping

The goal of this attack is dropping or discarding the packet, the attacker can use two ways (packet forwarding) or (De-synchronization).

Selective Forwarding: The goal of this attack is to select some packets, forward it and drop the rest of this packet. There is another type of this attack, the attacker can drop all the packets and do not forward any of them. This type is called BlackHole.

Synchronization Attack: In this study, we classified attacks based on the OSI Model layers, these attacks can be used for link layer and transport layer of OSI model. The technique of Synchronization attacks for Listen-Sleep Slotted MAC protocols has been suggested by (Lu et al, 2008). In this attack, the adversary node tries to extend its listening slot and propagates the extended listening slot to the other nodes.

Consumption

In WSN the problem of energy consumption is much known, all groups of DoS attacks make node to eat up its battery power.

Hello Packet and Session Flooding: Some routing protocols are using hello packets for establishing the neighborhood relationship or connection request. An adversary can constantly send a hello packet by using a high power radio transmitter. The nodes which receive the hello packet believe that the adversary node is their neighbor, even though the adversary node is located far away. Attacker can also send the session request to the victim nodes until they get exhausted or they reach their limit for maximum number of connections.

3.2.5. Privacy Attacks

The main goal of Privacy attacks is finding the information about devices or about persons. The privacy attacks are considered dangerous and they are classified into two groups, Data oriented and Context oriented, because the attackers are only interested in the information (Li et al, 2009).

Data oriented

Eavesdropping: The eavesdropping is a potentially dangerous attack because it allows an attacker to retrieve such confidential information exchanged between a reader and a card measuring the RF field emitted by the reader. Its development and implementation are quite simple since antenna connected to an oscilloscope can be used

to collect the exchanged binary data (Thevenon *et al.*, 2011). While the communication distance between a reader and a card is close to ten centimeters, a spy is able to recover the signal sent by a player over 20 meters (200 times the operating distance).

Substitution, clonage, and replay attack: The three attacks are grouped in the same group as their main characteristics are the same. All these attacks require data recovery on another card without contact. These attacks are often preceded by an eavesdropping attack or skimming attack which allows the attacker to retrieve the data stored in the memory of a transponder. This data can then be recorded onto a blank transponder to get a copy of the card previously attacked. Writing data on a blank card is quite simple since we can find on the internet all the equipment used to program any card, using a microprocessor.

Context oriented

Traffic Analysis: The attacker listens to the packets and tries to find the sensed data or ask to be sent. For example, in health application, an attacker may try to access the patient's confidential medical information.

Tempering Attacks: As described by (Sharifi *et al.*, 2013) the tampering attacks are well surveyed and classified into three difficult levels called easy, medium, and hard. The higher the difficulty level is, the most accessed and controlled over the victim hardware component. On the other hand, the easier attacks need less facility.

Tag modification: The most RFID tags use a writable memory. As a consequence, modifying or deleting valuable information could be performed easily by an attacker and it depends on the used standard and the READ/WRITE protection employed.

3.2.6. Impersonation

The last group is impersonation attack; the attackers want to impersonate themselves either as a physical node or a number of virtual nodes. This type of attacks might not sound very harmful for WSN, but destructive for RFID system and can be combined with DoS attacks and make them more destructive.

Physical Layer Identification: The main goal of the attacker is to impersonate a target device by generating packets or signals that contain factors which are sensitive to the fingerprint device. Therefore, the attacker can impersonate himself as one of targeted devices. This attack is used to detect the wireless devices in a network. It is based on the distinctive physical layer characteristics of a single device which are mainly due to manufacturing imperfection. The attacker uses two main techniques for device identifications. Transient based technique which is based on the unique features during the transient phase when radio is turning On (Danev and Capkun, 2009). The other technique is Modulation based techniques in which, modulation imperfection of wireless transceivers is the focusing point (Zeng *et al.*, 2010).

Sybil: In Sybil attacks, the attacker makes multiple identities and replicates a single node. These identities could be fabricated or stolen identities. Fabricated identities are fake identities which are randomly generated by the attacker. For example, if a node ID is represented by 32 bits, an attacker can randomly create 32 bits identities. In some networks where new nodes are not allowed to join, the attacker can steal the identities of legitimate nodes and use them for Sybil attack (Newsome *et al.*, 2004).

4. Common attacks countermeasures

This section is an overview of existing countermeasures to enhance security of IoT communication technologies. We identified countermeasures for WSN/RFID combined attacks based on different OSI layers, as shown in Table 2.

Attacks	Countermeasures
Jamming	Regulated transmitted power, Direct-Sequence Spread Spectrum, Direct-Sequence Spread Spectrum, and Hybrid FHSS/DSSS.
Wormhole	Physical monitoring of Field devices and regular monitoring of network using Source Routing. Monitoring system may use packet leach techniques.
Replay	Timestamps, one-time passwords, and challenge response cryptography
Traffic Analysis	Sending of dummy packet in quite hours: and regular monitoring WSN network
Eavesdropping	Session Keys protect NPDU from Eavesdropper
Sybil	Trusted Certification, Resource Testing, Recurring Fees, Privilege Attenuation, Economic Incentives, Location/Position Verification, Received Signal Strength Indicator (RSSI)–based scheme and Random Key Predistribution.

Table 2. Common attacks and countermeasures

4.1. Countermeasure against Jamming

4.1.1. Regulated transmitted power

By using low transmitted power, the discovery probability from an attacker decreases (an attacker must locate first the target before transmitting jamming signal). Higher transmitted power implies higher resistance against jamming because a stronger jamming signal is needed to overcome the original signal (Zhang and Kitsos, 2009).

4.1.2. Frequency-Hopping Spread Spectrum

Frequency hopping spread spectrum (FHSS) is a way of transmitting radio signals by fast switching a carrier amid many frequency channels, benefitting from the use of a shared algorithm known both to the transmitter and the receiver. FHSS brings forward many advantages in WSN and RFID systems, e.g (Mpitziopoulos and Gavalas, 2009).

- It reduces unauthorized interception and jamming of radio transmission between Tag and Reader in RFID and the nodes in WSN.
- It deals effectually with the multipath effect. One of the main drawbacks of frequency-hopping is that the overall bandwidth required is much wider than that required to transmit the same data using a single carrier frequency. However, transmission in each frequency lasts for a very limited period of time so the frequency is not occupied for long.

4.1.3. Direct-Sequence Spread Spectrum

Direct-Sequence Spread Spectrum (DSSS) transmissions are performed by multiplying the data (RF carrier) being transmitted and a pseudo-noise (PN) digital signal. This PN digital signal is a pseudorandom sequence of one and one values, at a frequency (chip rate) much higher than that of the original signal. This process causes the RF signal to be replaced with a very wide bandwidth signal with the spectral equivalent of a noise signal; however, this noise can be filtered out at the receiving end to recover the original data, through multiplying the incoming RF signal with the same PN modulated carrier. The first three of the above-mentioned FHSS advantages also apply into DSSS. Furthermore, the processing applied to the original signal by DSSS makes it difficult to the attacker to descramble the transmitted RF carrier and recover the original signal (Fang *et al.*, 2016).

4.1.4. Hybrid FHSS/DSSS

In WSN the Hybrid FHSS/DSSS communication between nodes represents the hoped anti-jamming measure. In general terms, direct-sequence systems achieve their processing gains through interference attenuation using a wider bandwidth for signal transmission, while FHSS through interference avoidance. Thus Hybrid FHSS/DSSS develop the solidity to combat the near/far problem, which arises in DSSS communications schemes. Another welcome feature is the capability to adapt to a diversity of channel problems (Mpitziopoulos and Gavalas, 2009).

4.2. Wormhole Countermeasure

A wormhole attack is considered dangerous as it is independent of MAC layer protocols and immune to cryptographic techniques. Strictly speaking, the attacker does not need to understand the MAC protocol or be able to decode encrypted packets to be able to replay them. Different papers in literature have developed countermeasures for wormhole attacks. The authors (Maheshwari et al, 2007) discussed them in two approaches. The first one is related to that Bound Distance or Time, and the second is based in graph theoretic and geometric.

4.3. Replay Countermeasure

In order to defend against replay attacks some simple countermeasures exist such as the use of timestamps, one-time passwords and challenge response cryptography. Nevertheless, these schemes are inconvenient and with doubtful efficiency considering the vulnerabilities to which challenge response protocols are susceptible to. Another approach is the use of RF shielding on readers in order to limit the directionality of radio signals and subsequently the appearance of a ghost. Another approach is based on the distance between the information requestor and the information owner. Implied that the signal-to-noise ratio of the reader signal in an RFID system can reveal even roughly the distance between a reader and a tag. This information could definitely be used in order to make a discrimination between authorized and unauthorized readers or tags and subsequently mitigate replay attacks (Mitrokotsa *et al.*, 2010).

4.4. Traffic Analysis Countermeasure

The way to defend against traffic analysis is to control the packet sending rate of every node in the network in such a way that every node sends packets with the same rate (Deng *et al.*, 2006). There is another way to defend traffic analysis is to ensure that the external appearance of a packet changes as it moves forward through a multi-hop sensor network. To do this, a cluster key is established among each set of neighboring nodes. The packet destination address, packet type, and packet contents are encrypted by a node, using its cluster key (Deng *et al.*, 2006). As a packet moves forward, each node first decrypts the packet and then re-encrypts it, using the cluster key. The current senders address remains in plaintext so that the receiver can choose the correct cluster key to decrypt the packet.

4.5. Countermeasure against Eavesdropping

Communications between WSN nodes and RFID (Tags and Readers) are vulnerable to the eavesdropping because very few nodes and passive tags are using the cryptographic protections. However, due to the short reading range of passive tags (Zhang and Kitsos, 2009), the eavesdroppers need to be the physical proximity of RFID tags, which is a sporadic activity. In order to protect against eavesdropping, data cryptography can prevent these security issues. Presently, sensor networks are supplied exclusively through symmetric key cryptography. The entire network is under risk if only one of its nodes has to be compromised by using symmetric cryptography. It means that the shared secret among those nodes is exposed. Another approach is to use a shared key between two nodes in the whole network. Then, it removes the network wide key. The disadvantage is additional nodes which cannot be added after the deployment process. In a sensor network with n nodes, each node needs to store (n-1) keys.

4.6. Countermeasure against Sybil attacks

There are different methods proposed against Sybil attacks but still there is no general solution to the Sybil attack. A number of approaches for various combinations of environments and attacks have been proposed (Levine *et al.*, 2006).
The most prominent techniques to resist Sybil attacks are as under.

- ***Trusted Certification***: is by far the very often cited solution to subdual Sybil attacks. It involves the presence of a trusted Certifying Authority (CA) that validates the one is to one correspondence between nodes on the network and its associated identity.
- ***Resource Testing***: is the most habitually implemented solution against Sybil attacks, despite it is ineffective for most systems.
- ***Recurring Fees*** or ***(Recurring Costs)*** is a variation method of resource examining where resource tests are conducted after certain specific time intervals to impose a specific "cost" on the attacker that is incurred for every identity that he controls. Using recurring costs or fees per identity is more effective to inhibit Sybil attacks than a one-time resource test.
- ***Privilege Attenuation***: is a technique to mitigate Sybil attack limited to Social Network System (SNS) as an application domain, this technique frequently used in (SNS) despite its disadvantages is only applied to monotonic policies. Significant run-time and storage overhead for generalized extensions of the idea (Fong, 2011) .
- ***Economic Incentives***: is a general technique used to mitigate Sybil attack, but this method is not efficient because it may encourage Sybil attackers that have no interest in subverting the application protocols, but that are interested in being paid to reveal their presence (Margolin and Levine, 2007).
- ***Location/Position Verification***: this technique is only limited to ad hoc networks. Methods employing this technique make use of the fact that any identities that are projected by any single physical device must be in the same location. Locations are verified using specific methods such as triangulation (Tangpong, 2010). So for an attacker with a single physical device, all Sybil identities will be in the same place or will appear to move together
- ***Received Signal Strength Indicator (RSSI)–based scheme***: is a technique used to mitigate Sybil attack. It does not deal with existing Sybil nodes in the network, Location calculations are costly. It is limited to Sensor Networks (Balachandran and Sanyal, 2012) .
- ***Random Key Redistribution***: is a technique limited in wireless sensor network but we can use it in other systems like RFID (Newsome *et al.*, 2004).

5. Conclusion

The Internet of things technologies are exposed to different types of attacks. An attacker can attack for different objectives. Attacks are categorized based on attacking goals and different OSI layers. In this paper, the most important attacks on WSN and RFID are identified, discussed, and presented in a systematic form to allow their comparison and trace the future research activities in this field. The use of conventional cryptography in the Internet of things is limited or even impossible. For that, our research will be oriented towards alternative solutions less costly and complex, including the use of codes in Cryptography. As known, the performance of an algorithm in IoT is of paramount importance. To this end, code based cryptography is most suitable for IoT as it has a very fast and efficient encryption procedure (Persichetti, 2012). In addition, there are no known vulnerabilities on this solution so it should be more secure even against quantum computer.

6. References

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). *A survey on sensor networks*. IEEE communications magazine, 40(8), 102-114.

Ashton, K. (2009). *That 'internet of things' thing*. RFiD Journal, 22(7), 97-114.

Awerbuch, B., & Scheideler, C. (2004). *Group spreading: A protocol for provably secure distributed name service*. In International Colloquium on Automata, Languages, and Programming . Springer Berlin Heidelberg. (pp. 183-195).

Babar, S., Stango, A., Prasad, N., Sen, J., & Prasad, R. (2011). *Proposed embedded security framework for internet of things (iot). In Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE)*, 2011 2nd International Conference on (pp. 1-5). IEEE.

Balachandran, N., & Sanyal, S. (2012). *A review of techniques to mitigate sybil attacks*. arXiv preprint arXiv:1207.2617.

Borgohain, T., Kumar, U., & Sanyal, S. (2015). *Survey of security and privacy issues of internet of things*. arXiv preprint arXiv:1501.02211.

Chaczko, Z., Jacak, W., & Łuba, T. (2015). *Computational Intelligence and Efficiency in Engineering Systems* (Vol. 595). G. Borowik (Ed.). Springer.

Clauberg, R. (2004). *RFID and sensor networks*. In Proc. RFID Workshop, St. Gallen, Switzerland (pp. 1-6).

Danev, B., & Capkun, S. (2009). Transient-based identification of wireless sensor nodes. In Proceedings of the 2009 International Conference on Information Processing in Sensor Networks (pp. 25-36). IEEE Computer Society.

Deakin, M. (2013). *Smart cities: governing, modelling and analysing the transition*. Routledge.

Deng, J., Han, R., & Mishra, S. (2005). *Countermeasures against traffic analysis attacks in wireless sensor networks.* In First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM'05) . IEEE. (pp. 113-126).

Deng, J., Han, R., & Mishra, S. (2006). *Decorrelating wireless sensor network traffic to inhibit traffic analysis attacks*. Pervasive and Mobile Computing, 2(2), 159-186.

Fang, S., Liu, Y., & Ning, P. (2016). *Wireless communications under broadband reactive jamming attacks*. IEEE Transactions on Dependable and Secure Computing, 13(3), 394-408.

Fong, P. W. (2011). *Preventing Sybil attacks by privilege attenuation: A design principle for social network systems*. In 2011 IEEE Symposium on Security and Privacy (pp. 263-278). IEEE.

FTC Sta¬ Report. (2015). *Internet of things: Privacy & security in a connected world*. Washington, DC: Federal Trade Commission.

Garcia-Morchon, O., Kumar, S., Struik, R., Keoh, S., & Hummen, R. (2013). *Security Considerations in the IP-based Internet of Things*.

Ghildiyal, S., Mishra, A. K., Gupta, A., & Garg, N. (2014). *Analysis of Denial of Service (DoS) Attacks in Wireless Sensor Networks*. IJRET: International Journal of Research in Engineering and Technology, 2319-1163.

Karlof, C., & Wagner, D. (2003). *Secure routing in wireless sensor networks: Attacks and countermeasures*. Ad hoc networks, 1(2), 293-315.

Law, Y. W., Hartel, P., den Hartog, J., & Havinga, P. (2005, January). *Link-layer jamming attacks on S-MAC*. In Wireless Sensor Networks, 2005. Proceeedings of the Second European Workshop on (pp. 217-225). IEEE.

Levine, B. N., Shields, C., & Margolin, N. B. (2006). *A survey of solutions to the sybil attack.* University of Massachusetts Amherst, Amherst, MA, 7.

Li, N., Zhang, N., Das, S. K., & Thuraisingham, B. (2009). *Privacy preservation in wireless sensor networks: A state-of-the-art survey.* Ad Hoc Networks, 7(8), 1501-1514.

Lu, X., Spear, M., Levitt, K., Matloff, N. S., & Wu, S. F. (2008). *A synchronization attack and defense in energy-efficient listen-sleep slotted MAC protocols.* In 2008 Second International Conference on Emerging Security

Information, Systems and Technologies (pp. 403-411). IEEE.
Maheshwari, R., Gao, J., & Das, S. R. (2007). *Detecting wormhole attacks in wireless networks using connectivity information*. In IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications (pp. 107-115). IEEE.
Margolin, N. B., & Levine, B. N. (2007). *Informant: Detecting sybils using incentives.* In International Conference on Financial Cryptography and Data Security .Springer Berlin Heidelberg. (pp. 192-207).
Mattern, F., & Floerkemeier, C. (2010). *From the Internet of Computers to the Internet of Things.* In from active data management to event-based systems and more .Springer Berlin Heidelberg. (pp. 242-259).
Mitchell, S., Villa, N., Stewart-Weeks, M., & Lange, A. (2013). *The Internet of everything for cities: connecting people, process, data and things to improve the livability of cities and communities.*
Mitrokotsa, A., Rieback, M. R., & Tanenbaum, A. S. (2010). *Classification of RFID attacks.* Gen, 15693, 14443.
Mpitziopoulos, A., & Gavalas, D. (2009). *An effective defensive node against jamming attacks in sensor networks*. Security and Communication Networks,2(2), 145-163.
Newsome, J., Shi, E., Song, D., & Perrig, A. (2004). *The sybil attack in sensor networks: analysis & defenses.* In Proceedings of the 3rd international symposium on Information processing in sensor networks (pp. 259-268). ACM.
Persichetti, E. (2012). *Improving the efficiency of code-based cryptography.* (Doctoral dissertation, Department of Mathematics, University of Auckland).
Rieback, M. R., Simpson, P. N., Crispo, B., & Tanenbaum, A. S. (2006). *RFID malware: Design principles and examples*. Pervasive and mobile computing, 2(4), 405-426.
Sharifi, A., Khosravi, M., & Shah, A. (2013). *Security Attacks and Solutions On Ubiquitous Computing Networks.* International Journal of Engineering and Innovative Technology (IJEIT), 3(4).
Sadeghi, M., Khosravi, F., Atefi, K., & Barati, M. (2012). *Security analysis of routing protocols in wireless sensor networks.* International Journal of Computer Science Issues, 9(1), 465-472.
Sundmaeker, H., Guillemin, P., Friess, P., & Woelfflé, S. (2010). *Vision and challenges for realising the Internet of Things.* Cluster of European Research Projects on the Internet of Things, European Commision.
Tangpong, A. (2010). *Managing Sybil Identities in Distributed Networks.* (Doctoral dissertation, The Pennsylvania State University).
Tayebi, A., Berber, S., & Swain, A. (2013). *Wireless Sensor Network attacks: An overview and critical analysis.* In Sensing Technology (ICST), 2013 Seventh International Conference on (pp. 97-102). IEEE.
Thevenon, P. H., Savry, O., Malherbi-Martins, R., & Tedjini, S. (2011). *Attacks on the HF physical layer of contactless and RFID systems.* INTECH Open Access Publisher.
Veríssimo, P., & Rodrigues, L. (2001). *Fundamental security concepts.* InDistributed Systems for System Architects .Springer US. (pp. 377-393).
Walewski, J. W., Bauer, M., Bui, N., Giacomin, P., Gruschka, N., Haller, S., ... & Magerkurth, C. (2011). *Project Deliverable D1. 2–Initial Architectural Reference Model for IoT*.
Weber, R. H. (2010). *Internet of Things–New security and privacy challenges*. Computer law & security review, 26(1), 23-30.
Xu, W., Ma, K., Trappe, W., & Zhang, Y. (2006). *Jamming sensor networks: attack and defense strategies*. IEEE network, 20(3), 41-47.
Zeng, K., Govindan, K., & Mohapatra, P. (2010). *Non-cryptographic authentication and identification in wireless networks*. network security, 1, 3.
Zhang, Y., & Kitsos, P. (2009). *Security in RFID and sensor networks.* Auerbach Publications.
Zharinov, R., Trifonova, U., & Gorin, A. (2014). *Using RFID Techniques for a Universal Identification Device.* In Proceedings of the 13th Conference of Open Innovations Association FRUCT and 2nd Seminar on e-Tourism for Karelia and Oulu Region (pp. 244-248).
Zheng, L., Zhang, H., Han, W., Zhou, X., He, J., Zhang, Z., ... & Wang, J. (2011). *Technologies, applications, and governance in the internet of things.* Internet of Things-Global technological and societal trends. From smart environments and spaces to green ICT.

Recherche

Modélisation en temps continu pour les systèmes d'aide à la décision appliqués à la programmation physico-financière

Continuous-time Modeling for Decision Support Systems applied to Budgetary Planning

Davy Hélard
IRISA – Université de Bretagne-Sud & MGDIS, Vannes, France
davy.helard@gmail.com

Jean-Philippe Gouigoux
MGDIS, Vannes, France
gouigoux-jp@mgdis.fr

Flavio Oquendo
IRISA – Université de Bretagne-Sud, France
flavio.oquendo@irisa.fr

Résumé

Dans le cadre de l'informatique décisionnelle, la programmation physico-financière doit permettre à des acteurs d'une collectivité provenant de différents domaines de faire converger leurs problématiques vers un objectif commun lors des dialogues de gestion. Cette programmation permet ainsi à ces acteurs d'étudier les conditions de réalisation et de mettre en place un suivi de l'avancement de cet objectif.
L'une des principales difficultés de la modélisation d'une programmation physico-financière est que chaque acteur exprime ses problématiques dans des échelles de temps différentes. Leur mise en lien dans un modèle représentant une réalité commune pose donc des problèmes au mode de représentation discret traditionnellement utilisé par les outils d'analyse financière, basés sur la logique des tableurs.
Dans cet article, nous proposons une nouvelle approche fondée sur les modèles en temps continu afin de permettre aux acteurs de regrouper leurs visions dans un modèle unique, tout en se plaçant sur des échelles de temps différentes.
Cette approche innovante a été implantée au sein de la société MGDIS à l'aide d'une architecture orientée service.

Abstract

In the scope of Business Intelligence, planning aims to support multiple actors in their process of converging different views and issues from different domains to get a shared business planning model. It is in particular the case of business planning in local governments.
A major difficulty in business planning is that each actor states her/his views and issues with a different time scale. Integrating them into a unique model that represents a common state of reality becomes very costly and awkward to manage when basing the construction of these models on discrete modeling techniques used by current tools of business planning.
This article proposes a novel solution, beyond the state-of-the-art, for addressing these issues : it conceives a novel meta model based on a continuous time calculus. Through the developed approach, it allows multiple actors to integrate the different business logics of their planning domain in a shared model as well as to observe it from different time scales.
This approach was implemented within a real industrial set in MGDIS following a service oriented architecture.

Mots-clés

Systèmes d'aide à la décision, modélisation en temps continu, programmation physico-financière.

Keywords

Decision support system, continuous-in-time model, budgetary planning.

1. Introduction

La gestion des finances des collectivités locales en lien avec leur action de terrain, et plus précisément la programmation physico-financière pose de nouveaux défis aux systèmes d'information et de décision, notamment pendant les dialogues de gestion.
La gestion des finances d'une entité publique est une activité complexe et peu standardisée. Cette complexité se reflète dans la grande hétérogénéité des méthodes proposées par les cabinets de conseil en gestion des finances des collectivités publiques. Chaque consultant utilise ses propres classeurs et ses feuilles de calcul reflétant son expérience et l'état financier de ses clients, et il n'y a que très peu de standardisation. La comptabilité (c'est-à-dire la représentation écrite des flux financiers) est normalisée, mais la façon dont sont effectivement gérées les finances (et notamment du point de vue prospectif et programmatique) ne l'est pas.
Les causes racines de cette complexité sont une difficulté réelle à prévoir l'évolution d'une situation financière soumise à de nombreuses contraintes endogènes et exogènes, une sémantique complexe, l'intervention de différents acteurs (décideurs, financiers, opérationnels internes et externes, etc.) avec des rôles et des objectifs très différents, voire parfois opposés, et enfin une difficulté de conceptualisation de l'axe temporel.
Ceci est notamment le cas de la programmation physico-financière consistant à décrire un projet de développement et à le confronter à une prévision de financement. L'objectif est ainsi de pouvoir réaliser au mieux les missions de la collectivité locale, en maîtrisant les équilibres financiers entre charges et ressources, avec pour finalité un axe politiquement fort sur la gestion du risque financier (respect des contraintes prudentielles). La récente crise des financements des entités publiques a eu pour conséquence que l'Etat exige de plus en plus de ces dernières qu'elles se projettent dans un futur économiquement incertain en analysant les impacts des aléas financiers.
Au-delà de la gestion de risque, il s'agit bien sûr également d'assurer une bonne gestion de l'argent public. Les collectivités publiques, en plus de répondre à des réglementations prudentielles, sont aussi engagées dans des approches de rationalisation des dépenses. Cette rationalisation a pour objectif de servir au mieux les usagers, tout en maintenant un endettement limité et des marges de manœuvre financières. L'actualité récente montre que ces problématiques sont de plus en plus connues du grand public. Cette exposition et son impact électoral potentiel vont de fait rendre encore plus critique la maîtrise des finances publiques.
En effet, l'actualité focalise souvent son analyse des finances des collectivités publiques sur le contrôle de celles-ci, mais le versant «positif» de la finance est la mise en œuvre de projets au service des citoyens.
Différentes approches de programmation financière permettent de remplacer les approches ad hoc matérialisées dans des feuilles de calcul à l'aide de tableurs par une approche fondée sur un modèle générique de programmation physico-financière. Ces approches ont le même fondement : la modélisation discrète. Parfaitement adaptées aux approches «simples» de type élaboration d'une PPI *(Programmation Pluriannuelle des Investissements)*, portées par la Direction des finances en réponse à une obligation réglementaire ou une volonté stratégique, ces solutions trouvent leur limite dès que l'on souhaite étendre leur usage vers le pilotage financier «au fil de l'eau» en intégrant la complexité du dialogue de gestion entre financiers et opérationnels.
Cet article analyse ainsi les limites de la modélisation discrète pour les systèmes d'aide à la décision appliqués à la programmation physico-financière et présente une nouvelle approche fondée sur la modélisation continue permettant de résoudre ces limitations tout en augmentant, de manière significative, le pouvoir d'expression des modèles génériques pour la programmation physico-financière. Les modèles ainsi spécifiés ouvrent de nouvelles perspectives pour le suivi entre le prévisionnel et le réel.
Dans les deux prochaines sections, nous présentons l'approche et la solution que nous avons développées selon l'approche de modélisation continue pour la programmation physico-financière. Ensuite, dans la quatrième section, nous analysons l'état de l'art des techniques d'aide à la décision dans le domaine de la programmation physico-financière. La cinquième section décrit les activités de validation à l'aide d'un prototype en grandeur réelle. Finalement, dans la sixième section, nous soulignons les implications et les limites de notre solution, avant de conclure notre propos et de présenter des perspectives de recherche.

2. L'approche de solution : la nécessité de penser continu

Dans le cadre des collectivités locales, la planification physico-financière est réalisée par un dialogue de gestion entre plusieurs acteurs. Bien que ces acteurs aient un objectif commun, chacun d'eux possède une problématique propre qui leur impose une vision du temps à des échelles différentes. Étant donné que ces problématiques sont liées entre elles, un dialogue de gestion fluide nécessite un modèle capable d'unifier ces problématiques tout en s'adaptant aux besoins de chaque acteur en termes d'échelles de temps.
Nous montrerons, dans la suite de cet article, comment la modélisation continue rend possible la création d'un modèle unifié qui peut être évalué sur plusieurs échelles de temps et ainsi permettre un dialogue fluide entre les acteurs de la programmation physico-financière.

2.1. Une voie de solution pour la problématique posée

La modélisation continue consiste à représenter les valeurs par des fonctions à temps continu plutôt que de les représenter par des valeurs discrètes.
Pour prendre un exemple simple : imaginons la gestion d'une cantine au niveau d'une commune. Au lieu de faire porter une valeur «fruits & légumes» avec 100 000 € dans la colonne pour l'année 2015, il s'agit désormais de raisonner autour d'une fonction mathématique décrivant la répartition de la densité de dépenses dans le temps, et dont l'intégration sur l'année 2015 retombera sur cette valeur de 100 000 €. Dans ce cas particulier, une fonction continue sur 10 mois avec une densité de 10 000 € par mois, et également continue sur les deux mois d'été avec une densité nulle peut être choisie. Bien sûr, des modélisations plus précises pourraient être envisagées, mais cela dépend du niveau de précision souhaite ultérieurement.
Évidemment, dans d'autres cas de dépenses, la fonction de répartition ne serait pas du tout la même. Par exemple, la représentation en termes de fonction la plus logique pour les salaires est une fonction composée de pics, représentés par des distributions de Dirac, portant la densité complète de salaire sur des instants précis de l'axe temporel qui sont typiquement le dernier jour ouvré du mois, à minuit.

2.2. L'apport de la modélisation continue

Bien qu'elle soit conceptuellement moins facile pour les analystes financiers, cette approche présente l'avantage de lever les problématiques de représentation incompatibles dans les différentes périodes temporelles utilisées par les différents acteurs du dialogue de gestion. En effet, la valeur sur une périodicité donnée étant systématiquement retrouvée par l'intégration d'une seule fonction mathématique représentant la réalité, les nombres obtenus sont automatiquement cohérents sur toutes les échelles de temps.
Cette approche a également pour avantage de forcer les analystes financiers, modélisateurs, à se poser les bonnes questions de la répartition effective des mouvements financiers. Pour reprendre les exemples précédemment cités, dans une approche discrète, rien ne contraint les modélisateurs à se poser la question de la répartition continue ou ponctuelle des dépenses. Or, ceci pose un réel problème métier lorsqu'il s'agit de gérer la trésorerie, au jour le jour. Un financier, à son niveau métier, peut se permettre de considérer les salaires et les dépenses pour les fruits et légumes au même niveau, à savoir une masse mensuelle, voire annuelle. Mais il n'empêche que pour le gestionnaire d'une cantine qui paie au jour le jour ses fournisseurs, mais en fin de mois les salaires, la différence est fondamentale car c'est ce qui pilote sa trésorerie, sa relation à la banque et au final son budget propre.
Nous revenons donc aux difficultés d'interaction que nous décrivions plus haut : chacun s'occupant de son point de vue, si le modèle n'est pas commun, il est très difficile de réconcilier les valeurs et de faire comprendre à chacun les besoins de l'autre. Le fait de mettre en place un modèle mathématique commun, au-delà de l'apport technique, a également une portée très forte du point de vue conceptuel, car il modélise mieux une réalité partagée que de simples chiffres dans des cases. De plus, chacun peut le considérer de son propre point de vue, sans faire perdre de la richesse à un autre. Les états fonctionnent alors comme des plans de coupe ou des vues particulières d'un modèle unique, plutôt que comme des représentations de portions de modèles incohérents et qu'il faut rassembler après coup, au prix de circonvolutions intellectuelles et de discussions compliquées entre acteurs.

3. La solution : la modélisation continue

Fondée sur cette approche de modélisation en temps continu, nous montrons comment construire un modèle indépendamment de la périodicité choisie lors de l'évaluation de ces modèles.
Un modèle pour la programmation physico-financière est exprimé à l'aide des concepts de variables et d'équations. Dans la suite, nous présentons un extrait de la solution afin d'illustrer les principes qui l'ont guidée, en comparant à chaque fois le modèle résultant d'une modélisation discrète avec celui résultant d'une modélisation continue.

3.1. La représentation des variables

Les données d'un modèle sont représentées sous la forme de variables.

3.1.1. La solution en modélisation discrète

La modélisation discrète représente une variable par un nombre fixe de valeurs placées dans le temps selon une certaine périodicité. L'évaluation des variables est possible uniquement pour la périodicité choisie lors de la modélisation ou pour des périodicités de taille supérieure et parfaitement alignées sur les limites de la périodicité choisie, ce qui est très restrictif.

3.1.2. La solution en modélisation en temps continu

La modélisation en temps continu représente les variables continues sous forme de fonctions en temps continu. Deux types de variables sont utilisés pour la modélisation continue dans notre solution. La différence se traduit dans la façon de lire les valeurs sur la fonction représentant la variable. Dans le cas d'une variable de type «mesure», une valeur est évaluée pour un intervalle de temps, par exemple le nombre de tickets de cantine vendus entre le 1er mai 2015 à minuit et le 1er juin 2015 à minuit. La valeur correspond à la surface de l'aire sous la courbe (figure 1).

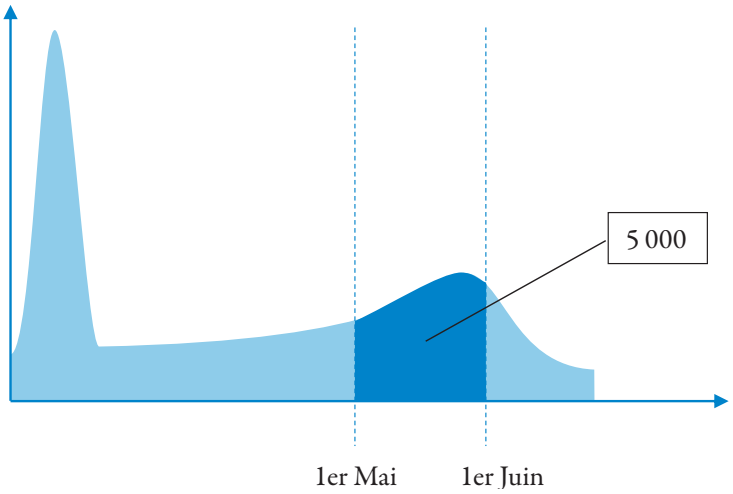

Figure 1. L'interprétation de la valeur d'une mesure

Pour ce type de variable, l'évaluation pour un intervalle de temps est donnée par la somme des évaluations des partitions de cet intervalle.
Dans le cas d'une variable de type «champ», une valeur est évaluée pour un instant précis, par exemple, le prix des billets le 1er mai 2015 à minuit (figure 2).

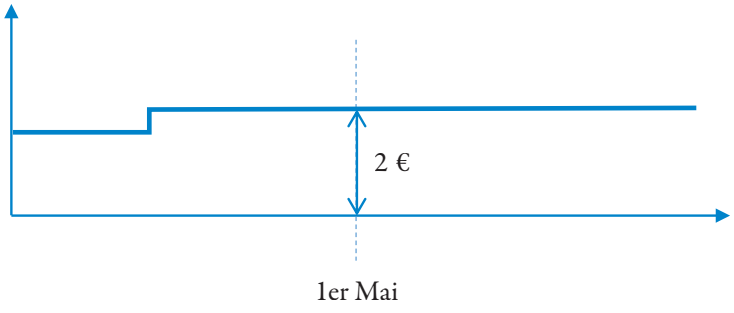

Figure 2. L'interprétation de la valeur d'un champ

3.1.3. La comparaison des solutions en modélisation discrète vs continue

L'utilisation de fonctions pour représenter les variables permet en principe aux modèles en temps continu d'être évalués pour n'importe quelle périodicité, contrairement aux modèles discrets qui sont liés à une périodicité par la nature même des variables.
Cependant, un modèle ne se limite pas aux variables. Celles-ci doivent être mises en relation par des équations afin de pouvoir estimer des variables à partir des variables connues. Ces équations doivent conserver l'avantage offert par la nature des variables de la modélisation continue.

3.2. La modélisation des équations

Les équations ont pour rôle de mettre en relation les différentes variables modélisées, établissant ainsi les différentes dépendances entre ces variables.

3.2.1. L'utilisation de différentes périodicités pour l'évaluation

Le passé est évalué en fonction de valeurs connues alors que le futur est estimé en fonction du passé et des décisions prises concernant le futur souhaité. Le passé et le futur sont deux cas qui nécessitent une modélisation particulière. Un suivi est réalisé au fur et à mesure que le présent avance. Progressant en même temps que des données sûres sont produites, les estimations établies pour le futur sont remplacées par la saisie de ce qui est réellement arrivé. Le suivi est réalisé régulièrement sur des périodes assez courtes, afin de gérer au plus tôt les imprévus.
À court terme, le futur est donc estimé sur des périodes fines alors que des périodes plus larges permettent d'avoir une vision plus globale à moyen terme. Par exemple, on peut envisager de réaliser les estimations du futur de l'année courante mensuellement et des années suivantes annuellement.

3.2.1.1. La solution en modélisation discrète

La modélisation discrète des équations se fonde sur un système d'indices, à la façon des suites numériques, pour faire référence aux périodes de temps. Il est ainsi difficile de prendre en compte un changement de taille de période dans l'écriture des équations. Afin de rester dans un cas classique, deux modèles discrets, fondés sur deux périodicités différentes, sont utilisés ensemble : l'un réalise les estimations à court terme et l'autre les estimations à long terme. Par exemple, le suivi de l'année en cours est réalisé mois par mois, tout en réalisant les estimations à long terme année par année. La figure 3 montre ces deux échelles de temps.

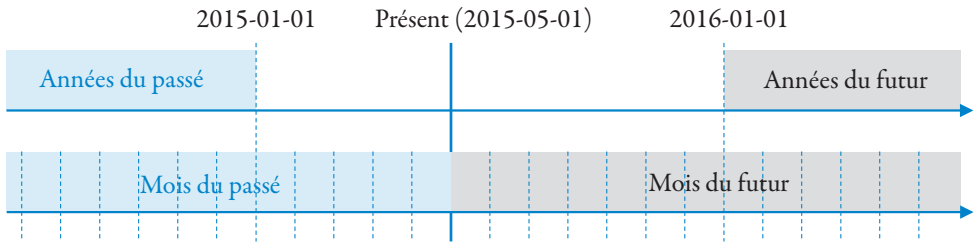

Figure 3. Une date qui sépare les périodes mensuelles ne permet pas nécessairement de séparer les périodes annuelles

Bien qu'en utilisant ces deux modèles il soit possible d'utiliser les périodicités adaptées au cas présent, le choix de celles-ci a dû être réalisé lors de la modélisation. De plus, la modélisation doit être déterminée selon les périodicités de l'évaluation alors que ce n'est pas la façon la plus naturelle de procéder.

3.2.1.2. La solution en modélisation en temps continu

En modélisation continue, comme la répartition des valeurs est formalisée, le choix des périodes d'évaluation est indépendant de la modélisation. Ainsi, est-il possible d'évaluer mois par mois pour l'année courante, et année par année pour les années suivantes, sans que cela nécessite une gestion particulière dans le modèle. En effet, la même formule est évaluée dans les deux cas.

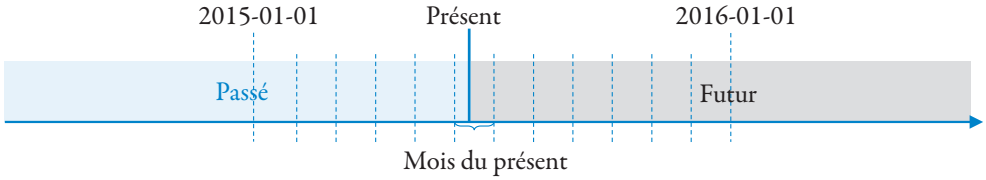

Figure 4. Le découpage du temps par année ou par mois

3.2.2. L'exemple de l'estimation du coût des denrées alimentaires

Afin d'illustrer la modélisation d'estimation à court et long terme, considérons l'évolution du coût d'un repas d'une cantine. En première approximation, l'inflation fait augmenter le coût des aliments d'une année sur l'autre selon une croissance exponentielle. Les équations des modèles permettent d'exprimer cette croissance.

3.2.2.1. La solution en modélisation discrète

Dans le cas d'une modélisation discrète, un modèle sert pour les prévisions à long terme année par année et un autre modèle aux prévisions à court terme mois par mois.
Pour le calcul avec une périodicité annuelle, la moyenne sur une année du coût en denrées d'un repas est calculée en multipliant simplement le coût moyen de l'année précédente par le coefficient d'inflation.

```
Présent et Futur :
"Coût d'un repas"[n] := 1.02 * «Coût d'un repas"[n - 1]
```

Figure 5. La formule de la variable «Coût d'un repas» pour une périodicité annuelle en modélisation discrète

Pour réaliser le suivi de l'année en cours, le modèle doit prendre en compte les variations des prix dues aux saisons. La formule modélise le fait que chaque mois gagne 2% d'une année sur l'autre. Il est inutile d'utiliser la racine douzième du facteur d'évolution pour lisser l'augmentation, car, dans le cas présent, l'inflation au cours de l'année est déjà présente dans les données saisies pour les années passées.

```
Présent et Futur :
"Coût d'un repas"[n] := 1.02 * «Coût d'un repas"[n - 12]
```

Figure 6. La formule de la variable «Coût d'un repas» pour une périodicité mensuelle en modélisation discrète

3.2.2.2. La solution en modélisation en temps continu

Afin d'estimer les coûts des années futures, il est important de comprendre ce qui détermine son évolution. En décomposant l'évolution du prix en plusieurs facteurs, il sera possible de jouer sur ceux-ci pour gagner en précision si besoin.
De nombreux facteurs sont responsables de l'évolution du coût des denrées. Les deux facteurs qui semblent avoir le plus d'importance ici sont l'inflation et les variations dues aux saisons.

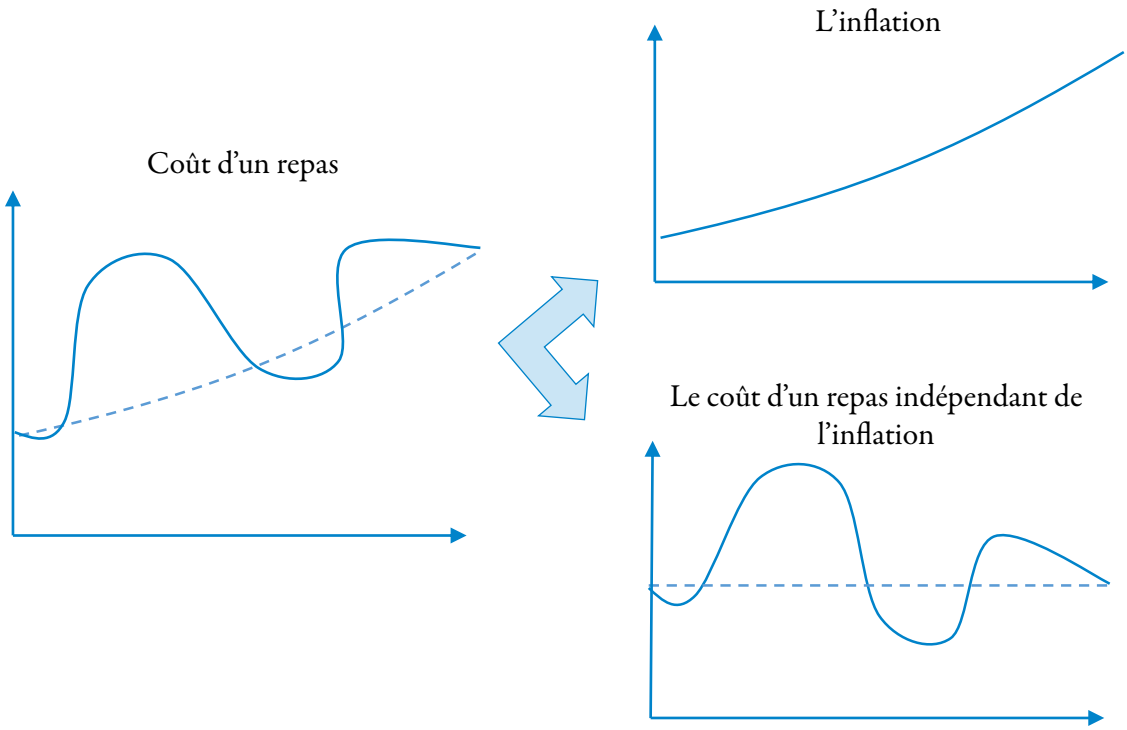

Figure 7. La décomposition des variations du coût des denrées en plusieurs facteurs

La répartition du coût des denrées prend déjà en compte l'inflation. Afin d'en extraire une répartition indépendante de cette croissance, la variable est divisée par l'exponentielle qui représente cette croissance. La fonction «Exp» de la formule en figure 8 décrit une fonction exponentielle qui vaut «1» pour le début de l'année courante et croît de 2% par an.

```
"Coût d'un repas"(t) / Exp(BeginningOf(Year(1), «now"()), 1.02, Year(1)))
```

Figure 8. La formule du coût indépendant de l'inflation ramené au 1er janvier de l'année du présent

Une fois la répartition au cours de l'année déterminée, celle-ci est répliquée sur les années du futur. L'estimation réalisée avec la formule ci-dessous ne prend pas en compte l'inflation. La modélisation se base sur une périodicité annuelle, car c'est la périodicité naturelle de l'évènement modélisé (typiquement liée au jeu des saisons et de la variation de production associée). Cette périodicité n'empêche toutefois pas une évaluation sur d'autres périodes, car, comme cela est visible sur la figure 9, l'évolution du coût est conservée par les calculs modélisés.

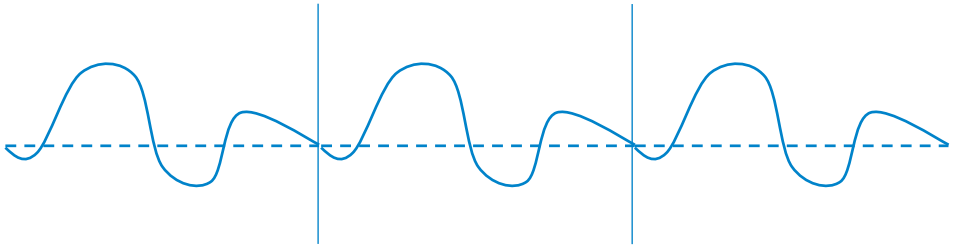

Figure 9. La courbe de répétition du coût d'un repas indépendant de l'inflation

L'opération «Repeat» (figure 10) permet de répéter les valeurs d'un intervalle en boucle. Cette opération peut être utilisée afin de représenter une situation cyclique.

```
Repeat(date d'origine : valeur scalaire, taille de l'intervalle : valeur scalaire,
 : valeur continue) : valeur continue
```

Figure 10. La signature de l'opération «Repeat»

La figure 11 présente la formule qui réplique cycliquement les variations du coût d'un repas au cours d'une année indépendamment de l'inflation.

```
If(t <BeginningOf(Year(1), «now"(),
 0,
 Repeat
(
BeginningOf(Year(1), «now"() - Year(1)),
Year(1),
"Coût d'un repas"(t) / Exp(BeginningOf(Year(1), «now"()), 1.02, Year(1)))
 )
)
```

Figure 11. La formule de répétition du coût d'un repas indépendant de l'inflation

Afin de prendre en compte l'inflation dans l'estimation du futur, les répartitions dupliquées sont multipliées par une fonction exponentielle qui représente l'inflation.

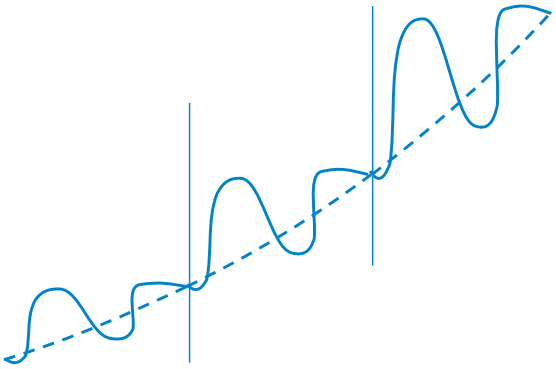

Figure 12. La courbe de l'estimation du coût des denrées d'un repas

La figure 13 présente la formule obtenue après avoir pris en compte l'inflation.

```
"Coût d'un repas"(t) :=
If(t <BeginningOf(Year(1),"now"(),
 0,
 Repeat
 (
BeginningOf(Year(1), «now"() - Year(1)),
Year(1),
"Coût d'un repas"(t) / Exp(BeginningOf(Year(1), «now"()), 1.02, Year(1)))
 )
* Exp(BeginningOf(Year(1), «now"()), 1.02, Year(1))
)
```

Figure 13. La formule de la variable «Coût d'un repas» en modélisation continue

D'une année à l'autre, la variation du coût due aux saisons n'est pas tout à fait identique. Des événements non déterminés peuvent induire du bruit dans cette variation. Ce bruit peut être éliminé, car il n'est pas pertinent de le conserver d'une année sur l'autre dans le cadre d'une activité de simulation.
Une moyenne sur plusieurs années permet de réduire ce bruit, mais ce n'est pas la méthode la plus efficace. L'idéal serait de déterminer les facteurs nécessaires au calcul en utilisant une approche de traitement du signal. Nous n'aborderons toutefois pas ce sujet dans le cadre du présent article.

3.2.2.3. Comparaison des solutions en modélisation discrète versus continue

A cause de la dépendance de la modélisation discrète à la périodicité d'évaluation, deux cas doivent être différenciés afin de limiter les calculs inutiles. La modélisation en temps continue permet de s'abstraire de ce problème, car la périodicité est choisie lors de l'évaluation.

3.3. Bilan de la comparaison entre les modélisations discrète et continue

Ces extraits de la solution développée montrent l'avantage de la modélisation continue sur la modélisation discrète. En effet, la modélisation discrète a un défaut majeur : un modèle discret doit être réalisé en tenant compte des contraintes liées à son évaluation. Lorsque ces contraintes sont multiples et paraissent incompatibles d'un point de vue discret, la modélisation discrète n'est plus suffisante. Comme nous l'avons vu au travers des exemples précédents, la modélisation continue permet de s'abstraire des préoccupations propres à l'évaluation et ainsi d'adapter l'évaluation d'un même modèle aux contraintes désirées. Un modèle unique peut ainsi être évalué pour chacune des périodicités nécessaires aux acteurs de la programmation physico-financière.

4. État de l'art

Afin d'établir l'état de l'art sur les techniques d'aide à la décision dans le domaine de la programmation physico-financière, nous avons utilisé une méthode de revue de la littérature scientifique nommée cartographie systématique. Cette méthode permet de recueillir et d'analyser les articles portant sur un domaine particulier à l'aide de questions de recherche précises et d'un protocole défini en amont des recherches bibliographiques réalisées sur des bases documentaires significatives.
La cartographie systématique permet ainsi de pallier la méthode traditionnelle souvent retrouvée dans les articles, dite revue narrative, qui consiste en un rappel des connaissances portant sur un sujet précis, recueillies de façon ad hoc par les auteurs.La revue systématique consiste à :

- rassembler, évaluer et synthétiser toutes les études pertinentes qui abordent un problème donné, en l'occurrence la programmation physico-financière ;
- limiter l'introduction d'erreurs aléatoires ou de biais.

L'élaboration d'une revue systématique est fondée sur un protocole détaillé préalable et est composée des cinq étapes présentées :

- Formulation des questions de recherche, comprenant la détermination des objectifs et des critères d'inclusion et d'exclusion des études ;
- Recherche et sélection des études pertinentes ;
- Évaluation de la qualité des études retenues ;

- Extraction des données pertinentes et analyse de ces données ;
- Interprétation des résultats obtenus à partir des questions de recherche.

Notre revue systématique propose ainsi une cartographie des articles de la littérature qui s'intéressent à l'aide à la décision dans le domaine de la planification financière, y compris la programmation physico-financière.

4.1 Formulation des questions de recherche

Afin de présenter l'état de l'art pour la problématique traitée dans cet article, notre revue systématique et d'élaborer une cartographie systématique, répondra aux questions de recherche suivantes :

- Quelles sont les approches utilisées concernant la modélisation de la planification financière pour les systèmes d'aide à la décision ?
- Comment la modélisation en temps continu a-t-elle été appliquée aux systèmes d'aide à la décision pour la planification financière ?
- Quels sont les domaines d'application de la modélisation de la planification financière pour les systèmes d'aide à la décision ?

Afin de répondre à ces questions, les bases documentaires utilisées pour mener cette cartographie systématique sont les bases anglophones majeures en informatique décisionnelle (Tableau 1).

Source	URL
ScienceDirect	`http://www.sciencedirect.com`
IEEE Xplore	`http://ieeexplore.ieee.org`
SpringerLink	`http://link.springer.com`
ACM Digital Library	`http://dl.acm.org`

Tableau 1. Les sources utilisées pour la recherche de publications

Sur ces bases, les requêtes ont été définies à l'aide de mots clés :

- «decision», «Business Intelligence» et «BI» pour l'aide à la décision ;
- «budgetary planning», «business planning», «budget planning», «budgetary control», «business control», «budget control», «financial strategy» pour la planification financière.

```
(
«budgetary planning» OR «business planning» OR «budget planning» OR «budgetary control"
OR «business control» OR «budget control» OR «financial strategy"
)
AND
(
decision OR «Business Intelligence» OR «BI"
)
```

Figure 14. La requête pour la recherche de publications

Suite aux résultats obtenus par l'exécution de ces requêtes, les critères d'inclusion et d'exclusion suivants ont été appliqués.

Les critères d'inclusion :

- La publication présente une approche pour réaliser un système d'aide à la décision appliqué à la planification financière d'un domaine.

Les critères d'exclusion :

- La publication a pour sujet la planification financière, mais ne présente pas d'approche pour réaliser un système d'aide à la décision.
- La publication présente une approche pour réaliser un système d'aide à la décision, mais n'a pas pour sujet la planification financière.
- La publication ne présente pas d'approche pour réaliser un système d'aide à la décision et n'a pas de planification financière comme domaine d'application.

4.2 Recherche et sélection des études pertinentes

La requête définie a été exécutée sur chacune des bases documentaires donnant les résultats suivants le 30 juin 2015. Ces recherches ont été appliquées aux domaines de «Computer Science», «Decision Sciences» et «Business Information Systems».

Source	Nombre de résultats
ScienceDirect (SD)	57
IEEE Xplore (I3E)	37
SpringerLink (SL)	9
ACM Digital Library (ACM)	11

Tableau 2. Le nombre de résultats obtenus à partir des sources

Suite à l'application des critères d'inclusion et d'exclusion, les études primaires retenues sont, par ordre chronologique décroissant, les suivantes (Tableau 3).

Type	Année	Auteurs	Source
JA	2014	Cha *et al.*	ACM
BS	2013	Baker *et al.*	SD
JA	2012	Sato *et al.*	SD
JA	2010	Lourdes Borrajo *et al.*	SD
CP	2009	Eisenreich *et al.*	ACM
JA	2008	Wang *et al.*	SD
CP	2007	Suggs *et al.*	I3E
JA	2007	Wong *et al.*	SD
JA	2006	Xie *et al.*	SD
JA	2006	Bermúdez *et al.*	SD
JA	2005	Wen *et al.*	SD
CP	2004	Wang *et al.*	I3E
CP	2003	Mayo *et al.*	I3E
JA	2001	Rees *et al.*	SD
JA	1998	Choi *et al.*	SD
JA	1997	Kim *et al.*	SD
JA	1997	Cho *et al.*	SD
JA	1994	Sethi *et al.*	SD
JA	1993	Baugh *et al.*	SD
JA	1988	Hruschka *et al.*	SD
JA	1983	Broeckx *et al.*	SD
JA	1980	Lin *et al.*	SD

Tableau 3. Liste des études primaires

4.3 Évaluation de la qualité des études retenues

Chaque étude primaire retenue présente l'une des deux qualités suivantes : (i) études fondées sur des données probantes ; (ii) études fondées sur des résultats novateurs.

4.4 Extraction des données pertinentes et analyse de ces données

Les études primaires ont été classées par domaine d'application. La quasi-totalité des articles s'intéresse au cas de la programmation financière en entreprise.

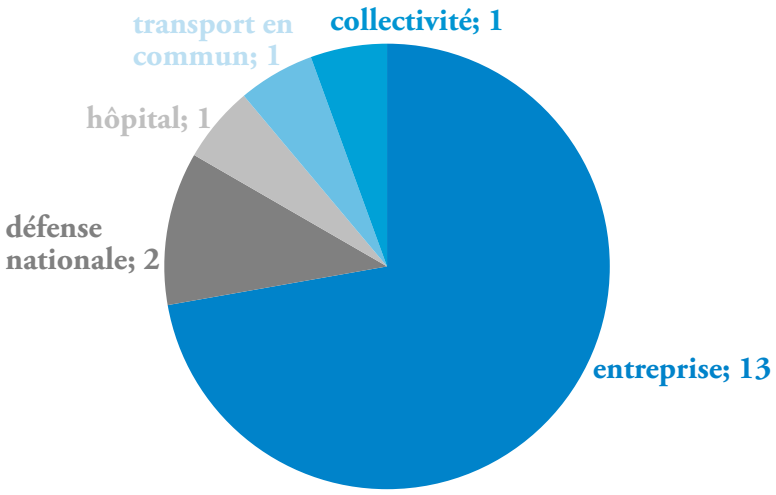

Figure 15. Le diagramme des études primaires par domaine d'application

Plus précisément, le tableau 4 liste les articles par domaine d'application.

Domaine d'application	Année	Auteurs	Source
Entreprise	2010	Lourdes Borrajo *et al.*	SD
	2007	Suggs *et al.*	I3E
	2005	Wen *et al.*	SD
	2004	Wang *et al.*	I3E
	1993	Baugh *et al.*	SD
	2006	Bermúdez *et al.*	SD
	1988	Hruschka *et al.*	SD
	2009	Eisenreich *et al.*	ACM
	2007	Wong *et al.*	SD
	2012	Sato *et al.*	SD
	1997	Kim *et al.*	SD
	1980	Lin *et al.*	SD
	2014	Cha *et al.*	ACM
	1997	Cho *et al.*	SD
Défense nationale	1998	Choi *et al.*	SD
	2006	Xie *et al.*	SD
Hôpital	2003	Mayo *et al.*	I3E
Transport en commun	2008	Wang *et al.*	SD
Collectivité	2008	Wang *et al.*	SD

Tableau 4. La liste des études primaires par domaine d'application

La plupart des articles se fondent sur des méthodes mathématiques afin d'optimiser des systèmes à contraintes multiples.

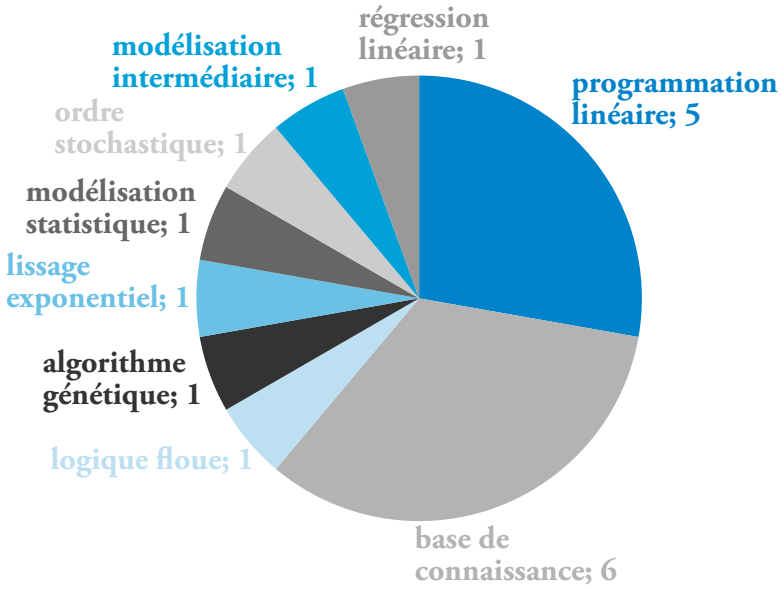

Figure 16. Le diagramme des résultats obtenus par approche

Les approches de solution proposées sont identifiées dans le tableau 5.

Approche proposée	Année	Auteurs	Source
Base de connaissances	2010	Lourdes Borrajo *et al.*	SD
	2008	Wang *et al.*	SD
	2007	Suggs *et al.*	I3E
	2005	Wen *et al.*	SD
	2004	Wang *et al.*	I3E
	1993	Baugh *et al.*	SD
Programmation linéaire	2012	Sato *et al.*	SD
	1998	Choi *et al.*	SD
	1997	Kim *et al.*	SD
	1997	Cho *et al.*	SD
	1980	Lin *et al.*	SD
Logique floue	1988	Hruschka *et al.*	SD
Lissage exponentiel	2006	Bermúdez *et al.*	SD
Algorithme génétique	2001	Rees *et al.*	SD
Régression linéaire	2014	Cha *et al.*	ACM
Dynamique des systèmes	2003	Mayo *et al.*	I3E
Modélisation intermédiaire	2009	Eisenreich *et al.*	ACM
Modélisation statistique	2006	Xie *et al.*	SD
Ordre stochastique	2007	Wong *et al.*	SD

Tableau 5. Les résultats obtenus par approche

4.5 Interprétation des résultats

La première question de recherche concerne les approches utilisées dans la modélisation de la planification financière pour les systèmes d'aide à la décision. Les principales approches proposées sont les bases de connaissance, la programmation linéaire, la logique floue et les algorithmes génétiques. Elles ont pour objectif d'optimiser un système sous plusieurs contraintes. Cette problématique est orthogonale à la problématique traitée dans cet article qui s'intéresse davantage à la modélisation notamment lorsque plusieurs domaines sont concernés.

La deuxième question de recherche a pour objectif d'analyser comment la modélisation en temps continu a été appliquée aux systèmes d'aide à la décision pour la planification financière afin de positionner l'approche présentée dans cet article en fonction des travaux existants. Notre cartographie systématique n'a trouvé aucun article proposant une approche utilisant la modélisation en temps continu.
La troisième question de recherche concerne les domaines d'application de la modélisation de la planification financière pour les systèmes d'aide à la décision. Comme cela est visible sur le tableau 4, la plupart des articles s'intéressent aux entreprises. Toutefois, quelques articles traitent de défense nationale, d'hôpitaux ou des transports en commun.
Cette étude corrobore la veille technologique réalisée par MGDIS avant de lancer une recherche coopérative à ce sujet. En effet, plusieurs méthodes d'optimisation présentées par les articles identifiés par la cartographie systématique ont été expérimentées par l'entreprise. Ces expérimentations ont également montré la nécessité de développer une solution nouvelle, au-delà de l'état de l'art.
La modélisation discrète est le dénominateur commun dans les différentes approches de solution présentées dans la littérature ainsi que dans les approches expérimentées. Ainsi, en conclusion, «dépasser les limitations de la modélisation discrète» a été identifié comme étant le principal verrou technologique.
Notre approche s'attaque à ce verrou technologique en développant une solution novatrice fondée sur la modélisation en temps continu dédiée à la programmation physico-financière.

5 La validation

Notre solution a été implémentée et validée sur un cas d'étude ainsi que sur des cas réels extraits des clients de la société MGDIS. Au-delà, elle a été largement validée vis-à-vis des besoins exprimés par les experts de la modélisation de programmation physico-financière.
Cette approche innovante a été implémentée au sein de la société MGDIS à l'aide d'une architecture orientée service. Du point de vue de la valorisation, notre approche pour la modélisation continue a abouti à la réalisation d'un prototype «preuve de concept»démontrant la faisabilité de l'application de la modélisation continue pour les clients du progiciel SOFI de la société MGDIS. Notre solution permettra ainsi de développer une nouvelle version de SOFI intégrant un moteur de calcul en temps continu.
Il est intéressant de souligner que l'un des objectifs de valorisation a été de rendre «transparent» le mode de modélisation temporelle : l'utilisateur ne sait donc pas, s'il reste sur des périodicités de rendus standards, que le moteur de calcul a changé.

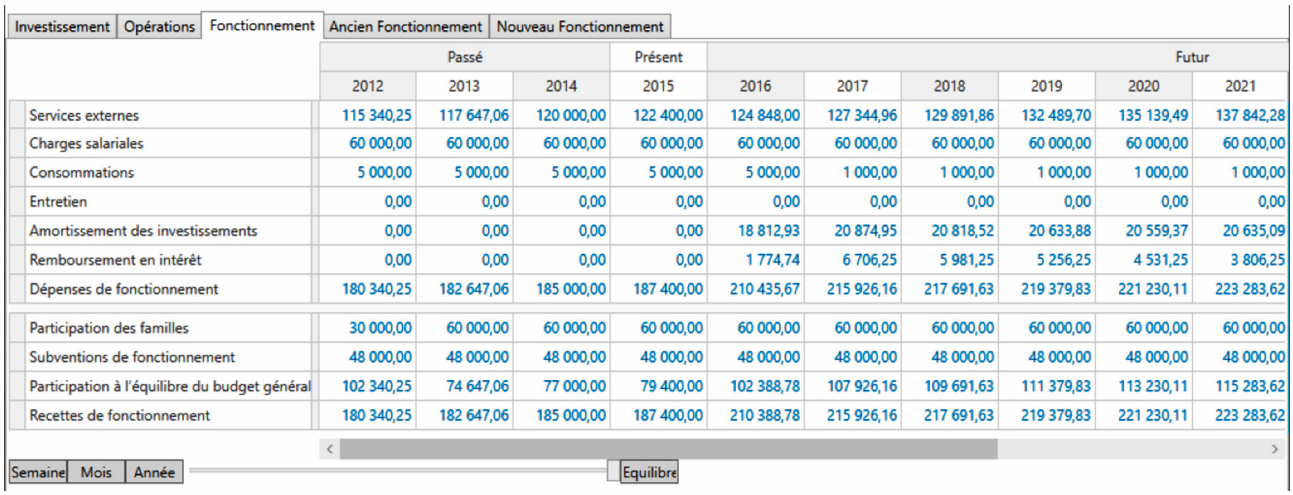

Figure 17. Une capture d'écran de l'interface graphique du prototype SOFI-Continu avec une périodicité annuelle

A l'inverse, si l'utilisateur souhaite se positionner en étude discrète d'un modèle sur des périodicités inconnues auparavant, les résultats sont recalculés comme illustrés dans la figure 18.

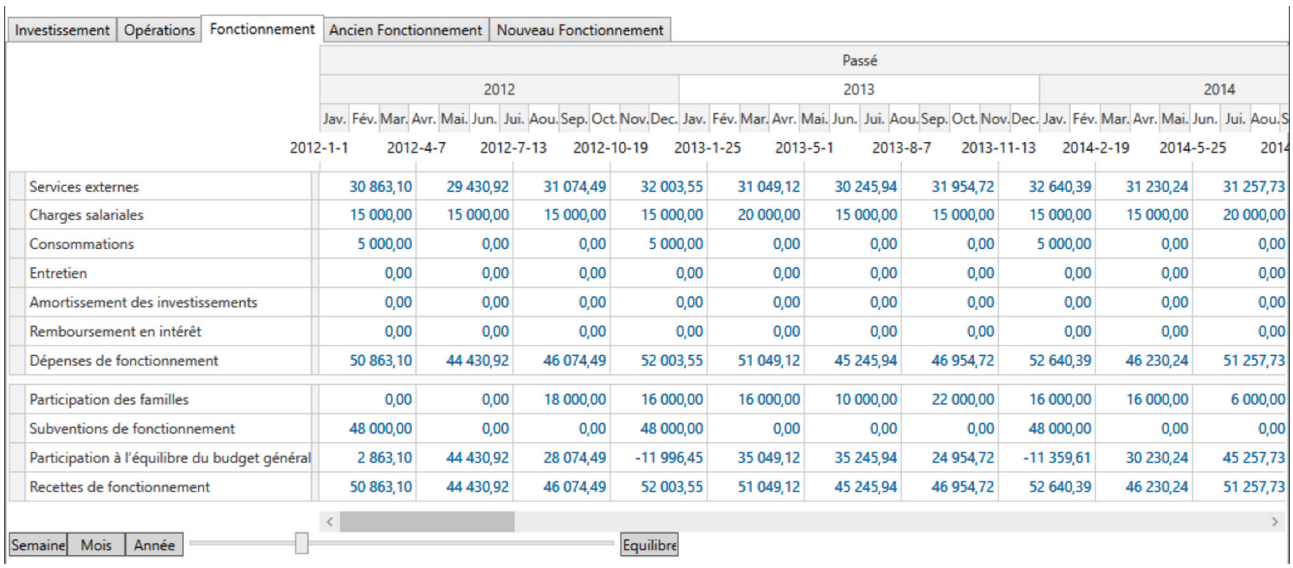

Figure 18. Une capture d'écran de l'interface graphique du prototype SOFI-Continu avec une périodicité quelconque

Ainsi, une analyse sur des demi-trimestres, par exemple, est dorénavant possible. De même, la vision du modèle du point de vue de la trésorerie, avec une valeur pour chaque jour est désormais disponible, et ce sur le même modèle mathématique que les gestionnaires financiers travaillant sur des années, ce qui est une richesse fonctionnelle considérable. Auparavant, la discrétisation au jour d'un modèle s'étalant sur vingt ans était impossible pour des raisons de performance. Le prototype ne voit pas ses performances varier significativement en fonction du niveau de détail dans lequel on observe le modèle. Il s'agit sans conteste d'un point très positif sur le résultat obtenu, car tel était le premier objectif fixé par MGDIS.

6 Conclusion et perspectives

Cet article a présenté une nouvelle approche pour la réalisation d'un système d'aide à la décision dans le domaine de la programmation physico-financière qui s'appuie sur une modélisation mathématique avec des fonctions continues. Revenons aux questions de recherche qui ont été posées. La première est «Comment formaliser une modélisation de planification financière applicable à des domaines variés ?».

Comme cela a été vu, la solution développée reprend les fondements posés par un méta modèle de calcul discret existant, en proposant pour chaque concept discret un raffinement en modélisation continue. Le méta modèle de calcul discret existant permettait déjà de définir la logique propre à chaque domaine d'application de la planification financière. Cependant, leur coexistence sur un modèle unique exigeait la reformulation de tous les modèles (ce qui est impraticable dans le cadre des dialogues de gestion). Le méta modèle proposé permet de résoudre ce problème avec une modélisation continue.

La seconde question de recherche traitée est «Comment évaluer un modèle d'application de la programmation physico-financière sur plusieurs échelles de temps ?».

La comparaison entre l'approche de modélisation continue et celle de modélisation discrète a permis de montrer comment le méta modèle proposé permet de décrire les valeurs et les formules de calcul en s'abstrayant de la périodicité choisie pour l'évaluation de celui-ci. L'implémentation du prototype de service de calcul a permis de valider que les valeurs obtenues pour l'évaluation d'un unique modèle sur plusieurs échelles de temps sont correctes.

La troisième et dernière question de recherche traitée est «Comment permettre une intégration du suivi de la programmation physico-financière au système d'information ?».

Pour permettre une intégration du suivi de la programmation physico-financière à un système d'information de gestion automatisé, le système logiciel a été développé avec une approche d'urbanisation dans laquelle l'architecture orientée service permet aux services de modélisation et de calcul de s'intégrer au bus de services d'entreprise. Un important bénéfice additionnel de cette approche est qu'elle permet la parallélisation, augmentant ainsi la capacité de traitement des requêtes.

Différentes perspectives sont envisagées pour la suite de ces travaux, dont la principale porte sur la détermination en temps réel de la robustesse du modèle financier. Elle vise ainsi la mise en œuvre d'un accompagnement du gestionnaire financier sur la robustesse de sa simulation. Lors de la recherche de valeurs correctes pour le modèle physico-financier, les gestionnaires recherchent avant tout des solutions qui permettent de réaliser le plus de projets possibles tout en maintenant les ratios prudentiels d'endettement à un niveau souhaitable.

Cette activité est complexe et nécessite une très grande expertise, mais elle pourrait être efficacement assistée par un moteur d'optimisation. Toutefois, une fois un ensemble de valeurs correctes trouvé, le gestionnaire financier n'est pas au bout de son travail, car il doit encore évaluer la résistance de son modèle à des aléas. Que se passe-t-il par exemple en cas d'aléa sur les travaux programmés ? Le modèle choisi comme idéal vole-t-il en éclats, avec une dégradation brusque de tous les indicateurs ou rend-il possible de compenser les dépenses par un autre poste d'emprunt ou de recette sans mettre en difficulté l'entité financière ?
C'est tout le rôle du calcul de robustesse du modèle que d'assister le financier en lui préparant des calculs de simulation de tel ou tel aléa. Nous pourrions par exemple imaginer que chaque valeur simulée pour le modèle soit accompagnée d'un résultat en temps réel sur l'impact d'une baisse de 5% de la population ou d'une baisse drastique des dotations aux collectivités par l'État. Disposer de ce genre de mesures en temps réel aiderait fortement les gestionnaires à choisir une solution peut-être un peu moins prometteuse sur le nombre de projets financés, mais avec une bien plus grande résistance aux évènements extérieurs.

7. Références

Baati, Lassaad*et al. Approche de modélisation DEVS à structure hiérarchique et dynamique* LSIS UMR-CNRS 6168, 2007.

Baker, David, et Wendy Evans, éd. *1 - Digital economics: introduction and overview.* In A Handbook of Digital Library Economics, 1-21. Chandos Publishing, 2013.

Baugh, P, A Gillies, et P Jastrzebski. *Combining knowledge-based and database technology in a tool for business planning.* Information and Software Technology 35, no 3 (mars 1993): 131-37. doi:10.1016/0950-5849(93)90050-D.

Bermúdez, J.D., J.V. Segura, et E. Vercher. *A decision support system methodology for forecasting of time series based on soft computing.* The Fuzzy Approach to Statistical Analysis 51, no 1 (1 novembre 2006): 177-91. doi:10.1016/j.csda.2006.02.010.

Broeckx, F. *Simulation in business planning and decision making: Thomas H. Naylor (Ed.) Volume 9, Number 1 in: Simulation Proceedings Series, Simulation Councils Inc., La Jolla, 1981, ix + 131 pages, $30.00.* European Journal of Operational Research 12, no 3 (mars 1983): 317. doi:10.1016/0377-2217(83)90206-0.

Cha, Sang K., Kunsoo Park, Changbin Song, Kihong Kim, CheolRyu, et Sunho Lee. *Interval Disaggregate: A New Operator for Business Planning.* Proc. VLDB Endow. 7, no 13 (août 2014): 1381-92. doi:10.14778/2733004.2733011.

Choi, Sang Hyun, ByungSeokAhn, Chang Hee Han, SoungHie Kim, et Jae Kyeong Kim. *Knowledge-based decision system for goal directed military resource planning.* Computers & Industrial Engineering 35, no 1-2 (octobre 1998): 299-302. doi:10.1016/S0360-8352(98)00079-5.

Cho, KwunIk, et SoungHie Kim. *An improved interactive hybrid method for the linear multi-objective knapsack problem.* Computers & Operations Research 24, no 11 (novembre 1997): 991-1003. doi:10.1016/S0305-0548(97)00021-X.

Eisenreich, Katrin. *Towards an Algebraic Foundation for Business Planning.* In Proceedings of the 2009 EDBT/ICDT Workshops, 161-69. EDBT/ICDT '09. New York, NY, USA: ACM, 2009. doi:10.1145/1698790.1698817.

Hruschka, Harald. *Use of fuzzy relations in rule-based decision support systems for business planning problems.* European Journal of Operational Research 34, no 3 (mars 1988): 326-35. doi:10.1016/0377-2217(88)90153-1.

Kim, Soung-Hie, Byeong-SeokAhn, et Sang-Hyun Choi. *An efficient force planning system using multi-objective linear goal programming.* Computers & Operations Research 24, no 6 (juin 1997): 569-80. doi:10.1016/S0305-0548(96)00040-8.

Lin, W Thomas. *An accounting control system structured on multiple objective planning models.* Omega 8, no 3 (1980): 375-82. doi:10.1016/0305-0483(80)90065-1.

Lourdes Borrajo, M., Juan M. Corchado, Emilio S. Corchado, María A. Pellicer, et Javier Bajo. *Multi-agent neural business control system.* Special Issue on Modelling Uncertainty 180, no 6 (15 mars 2010): 911-27. doi:10.1016/j.ins.2009.11.028.

Mayo, D.D., W.J. Dalton, et M.J. Callaghan. *Steering strategic decisions at London underground: evaluating management options with system dynamics.* In Simulation Conference, 2003. Proceedings of the 2003 Winter, 2:1578-84 vol.2, 2003. doi:10.1109/WSC.2003.1261605.

Rees, Jackie, et Reza Barkhi. *The problem of highly constrained tasks in group decision support systems.* European Journal of Operational Research 135, no 1 (16 novembre 2001): 220-29. doi:10.1016/S0377-2217(00)00323-4.

Sato, Yuji. *Optimal budget planning for investment in safety measures of a chemical company.* Sixteenth internationalworkingseminaronproductioneconomics, Innsbruck, 2010 140, no 2 (décembre 2012): 579-85. doi:10.1016/j.ijpe.2012.05.030.

Sethi, S.P., M. Taksar, et Q. Zhang. *Hierarchical decomposition of production and capacity investment decisions in stochastic manufacturing systems.* International Transactions in Operational Research 1, no 4 (octobre 1994): 435-51. doi:10.1016/0969-6016(94)90006-X.

Suggs, R., et B. Lewis. *Enterprise simulation - a practical application in business planning.* In Simulation Conference, 2007 Winter, 205-9, 2007. doi:10.1109/WSC.2007.4419602.

Wang, Fen, G. Forgionne, et Lidan Ha. *Reestimation of e-business planning model in real business world.* In e-Commerce Technology, 2004. CEC 2004. Proceedings. IEEE International Conference on, 317-20, 2004. doi:10.1109/ICECT.2004.1319750.

Wang, Huey-Jiun, Chien-Wei Chiou, et Yi-Kai Juan. *Decision support model based on case-based reasoning approach for estimating the restoration budget of historical buildings.* Expert Systems with Applications 35, no 4 (novembre 2008): 1601-10. doi:10.1016/j.eswa.2007.08.095.

Wen, W., W.K. Wang, et C.H. Wang. *A knowledge-based intelligent decision support system for national defense budget planning.* Expert Systems with Applications 28, no 1 (janvier 2005): 55-66. doi:10.1016/j.eswa.2004.08.010.

Wong, Wing-Keung. *Stochastic dominance and mean–variance measures of profit and loss for business planning and investment.* European Journal of Operational Research 182, no 2 (16 octobre 2007): 829-43. doi:10.1016/j.ejor.2006.09.032.

Xie, Haifeng, Thierry Chaussalet, Sam Toffa, et Peter Crowther. *A software tool to aid long-term care budget planning at local authority level.* International Council on Medical and Care Compunetics (ICMCC) 75, no 9 (septembre 2006): 664-70. doi:10.1016/j.ijmedinf.2006.04.009.

Model transformation from CIM to PIM in MDA: from business models defined in DFD to design models defined in UML

Transformation de modèles CIM à PIM selon MDA : des modèles métiers définis avec DFD aux modèles de conception UML

Yassine Rhazali
MISC Laboratory, Faculty of Science Kenitra, Ibn Tofail University, Kenitra, Morocco
dr.yassine.rhazali@gmail.com

Youssef Hadi
MISC Laboratory, Faculty of Science Kenitra, Ibn Tofail University, Kenitra, Morocco
hadiyoussef@gmail.com

Abdelaziz Mouloudi
MISC Laboratory, Faculty of Science Kenitra, Ibn Tofail University, Kenitra, Morocco
mouloudi_aziz@hotmail.com

Résumé

Cette recherche représente une méthodologie qui contrôle la transformation de modèles du niveau CIM au niveau du PIM en respectant l'approche MDA. Notre méthodologie est fondée sur l'établissement d'un bon niveau CIM, grâce à des règles bien choisies, afin de faciliter la transformation au niveau PIM. Cependant, nous créons un niveau de PIM riche grâce à un modèle du diagramme de cas d'utilisation, modèle de diagramme d'états, modèle de diagramme de classes et modèle de diagramme de package. Ensuite, nous établissons des règles de transformation pour assurer une transformation semi-automatique depuis le niveau CIM vers le niveau du PIM. Notre approche est conforme à l'approche MDA en prenant en considération la dimension métier au niveau CIM, puisque, nous présentons ce niveau par DFD et par le diagramme d'activité de l'UML. Cependant, nous modélisons le niveau PIM par les diagrammes d'UML, parce que UML est recommandé par MDA à ce niveau.

Abstract

This research represents a methodology that controls model transformation from CIM level to PIM level in accordance with MDA approach. Our approach is founded on establishing a good CIM level, through well-selected rules, to facilitate transformation to PIM level. However, we create a rich PIM level through use case diagram model, state diagram model, class diagram model and package diagram models. Then, we establish transformation rules to ensure a semi-automatic transformation from CIM level to PIM level. Our Approach conforms to MDA approach by taking into consideration the business dimension in CIM level, since we present this level through DFD and UML activity diagram. However, we model PIM by UML diagrams, because UML is recommended by MDA on this level.

Mots-clés

Transformation des modèles, MDA, CIM, PIM, PSM, processus métier.

Keywords

Model transformation, MDA, CIM, PIM, PSM, business process.

1. Introduction

Model Driven Engineering (MDE) (Schmidt, 2006) is an alternative approach, which aims at the development of information systems, based on the creation of source models and transforming them to multiple levels of abstraction until we automatically get a code. Its objective is to automate the process of software development that the specialists follow manually. MDE is a general approach that can be seen as a family of approaches, where Model Driven Architecture (MDA) supported by OMG is presented as the most interesting and the most common variant (OMG-MDA, 2014). MDA has the same principle as MDE, but it provides its own characteristics, defined on three levels of abstraction, defines some requirements to be respected, and also requires the use of some standards.
The first level of MDA is the Computation Independent Model (CIM) presented as model used by business managers and business analysts in order to describe the business process. The second level is the Platform Independent Model (PIM), which allows defining the models used by analysts and software designers to achieve an independent analysis and the conception of the developed software. The third level is the Platform Specific Model (PSM) which is considered model of code used by software developers. These models are believed to contain all the information needed to operate an execution platform used by software developers. The code is not a model of MDA, but it is the final result of the MDA process.
Transformations between the different levels of MDA begin with the transformations from CIM to PIM that aim to partially build PIM models from CIM models. The goal is to rewrite the information contained in the CIM models into PIM models, which would ensure that business information is conveyed and respected throughout the MDA process. Then, the transformation of PIM models to PSM models adds PIM technical information related to a target platform.
In practice, the automatic transformation starts from the second PIM level. However, our ultimate goal is to make the CIM level a productive one and also a basis for building PIM level through an automatic processing. The purpose is that the business models would not be limited to documents of communication between business experts and software designers.
In this paper, we propose a solution to induce the automation of the transformation of CIM level to the PIM level by studying how to use the current standards of business modeling effectively so as to achieve focused CIM models to simplify the transformation into PIM. Then, we define a set of rules to automate the transformation into a PIM level.
Our approach uses the DFD and the UML (OMG-UML, 2011) activity diagrams which represent standards of business model to define the CIM level. Then the rich business models of well-concentrated information help us to achieve models of PIM level. We divide the models of PIM level according to the three classical modeling views (Roques, 2004) including functional, static, and dynamic views. According to (Blanc, 2005), UML is recommended by MDA on the PIM level. Indeed, the first model of the PIM level is the use case diagram that defines the functionality of the information system, then the dynamic view is presented through the state diagram. Next, the class diagram model allows modeling static view through representing system classes and their relationships independently of a programming language in particular. Finally, all classes are structured in packages that are transformed from the CIM level.
Our previous works were based on BPMN and/or Activity diagram to model business process. However, the data flow diagram (DFD) is a simple modeling standard, because it is composed of a limited number of notations, and it is used by many researchers in their works to model the business process on CIM level like in (Kardoš and Drozdová, 2010). Indeed the searchers, who do not master design with complex language like BPMN and activity diagram, can design the CIM level, in our approach, through a simple language as DFD. For that, in this approach we benefit from the simplicity of the DFD to model the business process.
The rest of this paper is organized as follows. Section 2 presents our approach and describes the rules for constructing models of CIM level and the rules for transformation from the CIM level into the PIM level. In Section 3, we illustrate our proposal in a case study showing the construction of the CIM level and the transformation into the PIM level. Section 4 analyzes the related works of the CIM transformation into PIM. Then, we represent evaluation criteria and meta-models. Next, we analyze and discuss the transformation from CIM into PIM, by studying related works and our method. Finally, in Section 7, we conclude by determining the outcome of our work and describing future works.

2. Proposed method of transformation from CIM into PIM

Our proposal considers the business dimension on the CIM presentation level through the use of real high-level business models to preserve the business knowledge during the transformation into the PIM level in order to achieve a quality information system.

In this approach, we do not use DFD and UML 2 activity diagram just to present an approach that uses both standard modeling business processes, but we try to take advantage of each one to achieve a rich and concentrated level, which simplifies our transformation into the PIM level.
MDA recommends the use of UML on the PIM level. The use case diagram model presents the information system functionalities while class diagram shows the structure of the information system. Then, we organize all classes in packages themselves transformed from the CIM level.
Consequently, the input models in our transformation approach are DFD and activity diagram and the outputs models are use case diagram, state diagram, class diagram and package diagram.

Figure 1. Input models and output models

All PIM models in our approach are realized through an automatic transformation of CIM level, via well-defined and concentrated transformation rules; indeed, we have defined several rules in the beginning and then we filtered these rules by eliminating less powerful rules, by merging the identical rules, and by treating exceptional cases.
In our approach, CIM level is described by the business process while PIM level is represented by the three perspectives: functional view, dynamic view and static view. Then, based on the source meta-model and target meta-model, we define a set of transformation rules to move from CIM level to PIM level.
Below, we present the rules of construction of CIM level and the rules of transformation into the PIM level.

2.1 Construction rules of CIM Level

The rules for constructing the model of DFD (figure 2):

- Define means and not complex processes, i.e. each process must not contain other processes. In fact, each process must be comprised of about 4 to 12 tasks.
- Merge two processes into one if a sub-process consists of less than 4 tasks, or represents a complementary operation to another process.
- Coloring manual processes with another color; for example, we used gray.
- Verify that the model describes the most common business processes.
- Identify most actors who interact and who collaborate in the achievement of business processes since we talk about an enterprise process.

In this model, we show the processes and their relationship for modeling a general business process. This model identifies all business actors in order to show a real business process. The representation of multiple external triggers facilitates the transformation from CIM into PIM. Indeed, when moving from DFD model to the use case diagram model, the external trigger became actors. Nevertheless, we can specify average processes; for example in hotel accommodation, the customer must present the processes «choose room», «start reservation» and «present information», but the process «start reservation» holds less than four tasks; for this, we merge «choose room» and «start reservation» into a single process called «choose rooms for reservation».
The rules of construction of activity diagram model (cf. figure 2):

- Detail individually each process in a model as several actions (this latter constitute the fundamental unit in the activity diagram).
- In activity diagram model do not represent the manual tasks of DFD model.
- Present connections in this model.
- Enrich this model with the most exceptional ways through the gateways (decision node, merge node…). For example, in the case of the payment of an order, after we enter the password of our credit card, in the nominal case, the amount of order will be subtracted from our amount. But it should be noted that there is a case where the password is incorrect and another case in which the amount in account is insufficient to complete the transaction.
- Add an object node containing object state at the output of each action.

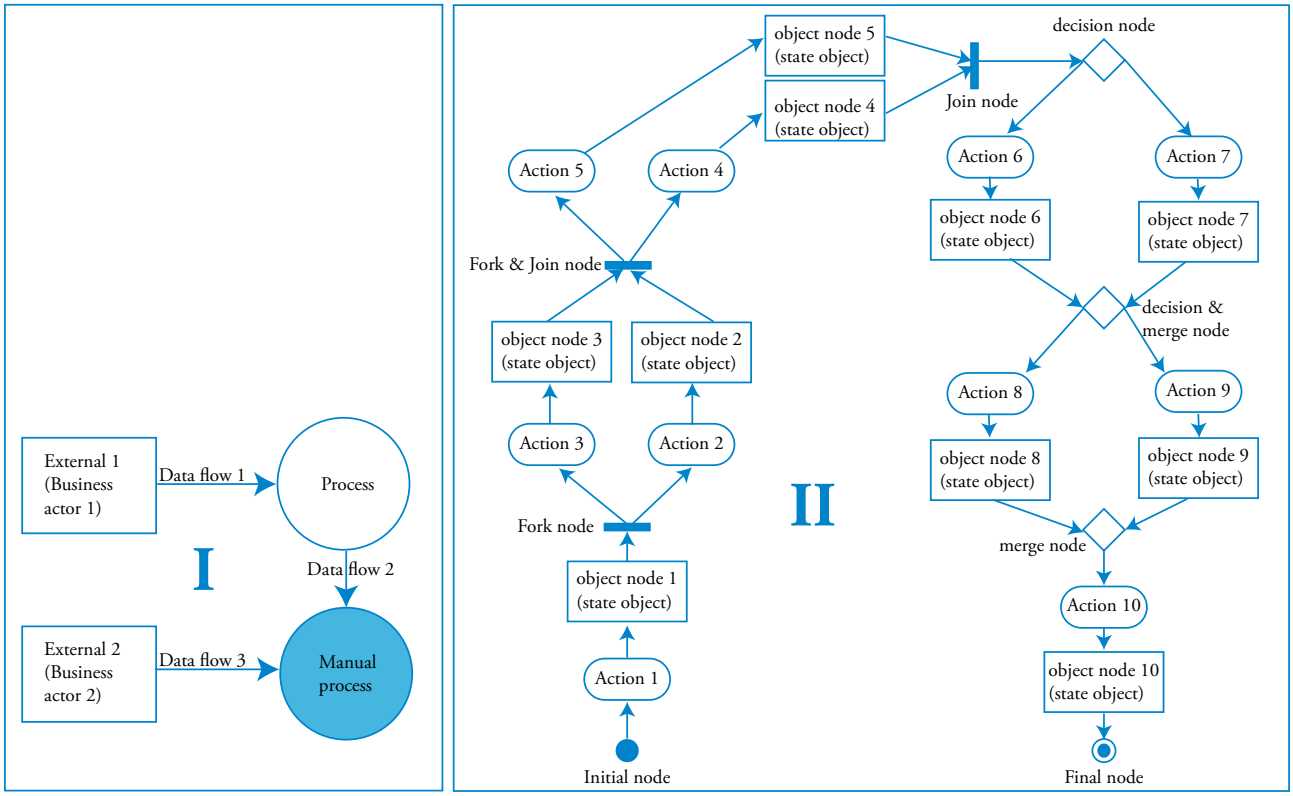

Figure 2. I: Generic model proposed of Data Flow Diagram II: Generic model proposed of activity diagram.

In this model, we represent the different actions contained in each process. For example in hotel accommodation, the process «choose room for reservation» is described by multiple actions. In the output of every task we show an object node with its state.

2.2 Transformation rules from CIM into PIM

The rules of passage from the CIM level models to model of use case diagram (Figure 3):

- Every action of the model activity diagram that corresponds to a functionality of the system is transformed into the use case.
- The collaborator, who realizes the process of the model of DFD, becomes an actor of use cases that correspond to the actions of this process.
- If there is a «decision node» between two actions, the corresponding use cases are connected by a relationship «extend». For example, Decision Point is between Action 1 and Action 2, and between Action 1 and Action 3, so the relationship «extends» will be presented between UC1 and UC2 then between UC1 and UC3.
- If there is just a control flow between two actions, the corresponding use cases are connected by a relationship «include».
- Do not transform the control flow returning back.
- Each process of DFD model is transformed into a package, which includes the use cases corresponding to the actions of this process.

For example in hotel accommodation, the process «choose room for reservation» is transformed into a package of use case diagram. Then the external «customer» transformed to actor, and the actions that detail the process «choose room for reservation» become use cases. Next, the gateway xor, which connects two actions, is transformed into a relation «extend», and the sequence flow, which connects two tasks, transforms into a relation «include». Nevertheless, we do not show the sequence flows that recur back.

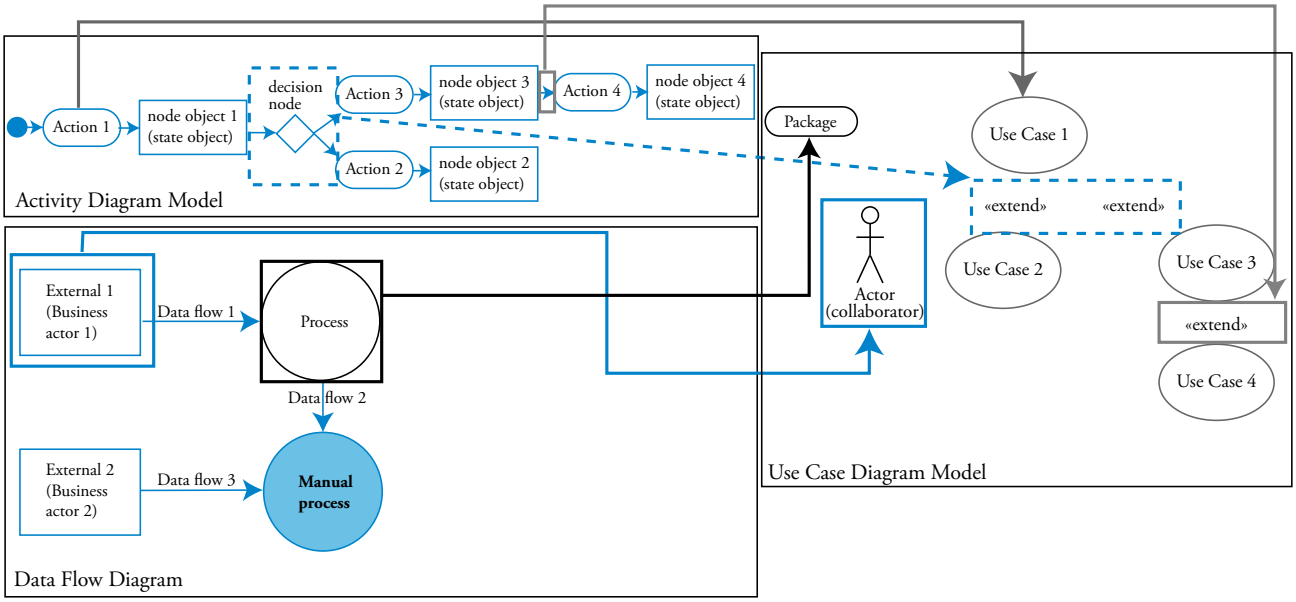

Figure 3. Schema of passage from DFD model and activity diagram model to use case diagram model

The rules of passage from model of the activity diagram to model of state diagram (Figure 4):

- R1: An object node transformed into a state.
- R2: A decision node transformed into a decision point.
- R3: A merge node transformed into a junction point.
- R4: A decision and merge node transformed into a junction point.
- R5: An initial node transformed into an initial state.
- R6: A final node transformed into a final state.
- R7: A control flow between two actions transformed into a transition.
- R8: A fork node transformed into a fork state.
- R9: A joint node transformed into a joint state.
- R10: A joint and fork node transformed into a joint and fork state.

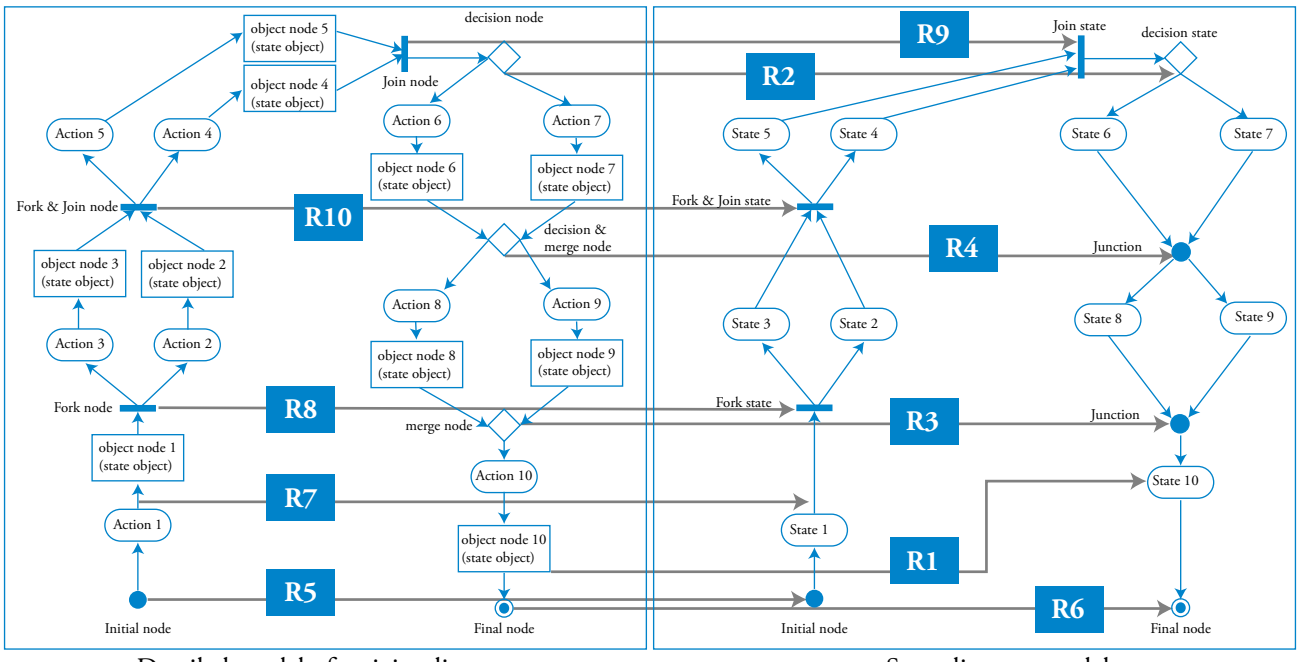

Figure 4. Schema of passage from activity diagram model to state diagram model

The state diagram model is transformed from model of activity diagram. First, object node is transformed into state and the control flow that lies between two tasks is transformed into a transition. For example, in hotel accommodation, the object node «Catalogue» with the state «displayed» becomes «Catalog displayed» in the state diagram model. However, the initial node is transformed into an initial state; the final node becomes a final state; the exclusive fork is transformed into a decision point; exclusive join become junction point; finally an exclusive fork & join node becomes a junction point.
The rules of passage from the model of activity diagram to the model of class diagram (Figure 5):

- Transform object nodes of model activity diagram as classes.
- Each state of an object becomes a class method.

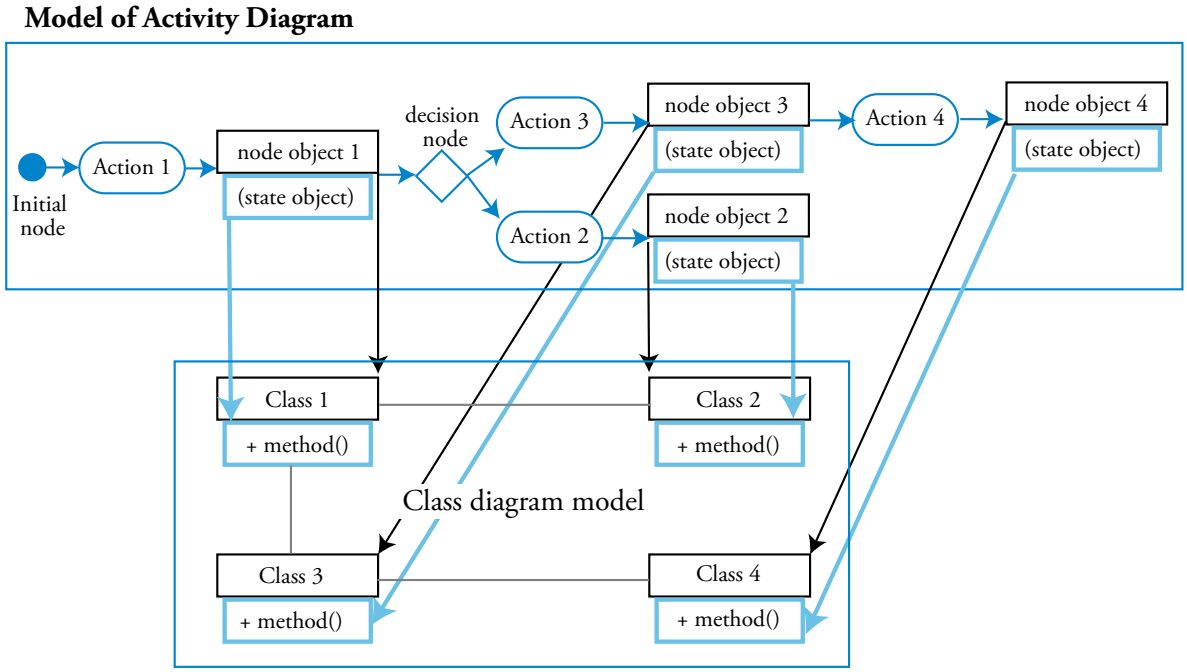

Figure 5. Schema of passage from activity diagram model to class diagram model

In this model, each object node is transformed into class. Then the states of an object node become functions in the class. For example, in hotel accommodation, the object node «reservation» with state «started» is transformed into class «reservation» that contains the method «start».
The rules of passage from the model of Data Flow Diagram and the model of class diagram to the model of package diagram (Figure 6):

- Each process is transformed into a package
- Classes resulting from the same process will be placed in the package that corresponds to the processes.

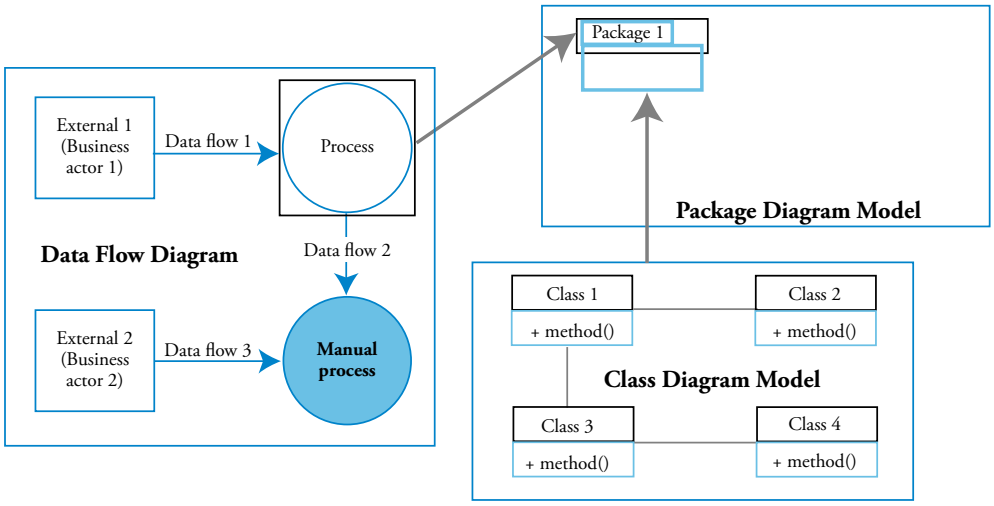

Figure 6. Schema of passage from DFD model and class diagram model to package diagram model

In this model, for example in hotel accommodation, the process such as «deal accommodation» becomes a package. In this section, we presented construction rules for structuring CIM in order to facilitate the transformation toward the PIM. For example, we defined an initial DFD model that represents the business process in a generic way as a set of processes (which may be detailed in 4 to 10 actions) each process will be detailed later in the model of activity diagram as several node objects and actions. Indeed, the object node will be transformed into class and the process will be transformed into package. Thereby the classes are distributed in packages, each package containing between 4 and 10 classes. Then, we presented transformation rules to move from business models, represented in the CIM level, to the analysis and design models, shown on the PIM level. In the next section we represent the use of construction rules and transformation rules in real case in order to substantiate our approach.

3. Case Study

In this section, we present a case study for sales through e-commerce to illustrate our approach of transforming the CIM level to the PIM level.
A customer can browse the catalog of products available and he can also see detailed information about each item. Then, he decides either to put a quantity of product in the cart or not. Each time the customer has the right to change the amount or eliminate completely the article from the cart. Once products that satisfy the needs of the customer are clearly selected, the latter starts the command. Then, he presents the payment information, and precise details of delivery.
An order agent begins treating the order, declaring the reservation of products specified by the customer. Then, the assembly worker collects reserved items, manually, from stock.
The assembly team leader checks quantity and quality of each product. Then the delivery agent carries the confirmed order, so that the customer gets his ordered products

3.1 Presentation of the CIM Level

According to MDA (OMG-MDA, 2014), the CIM level must be represented by business process models. Indeed, we find several standards enabling business process modeling such as BPMN, UML activity diagram, SoaML and DFD. In this paper, we based on UML activity diagram because it is an OMG standard for modeling business process, and on DFD because it is a simple standard since it contains a limited number of notations.

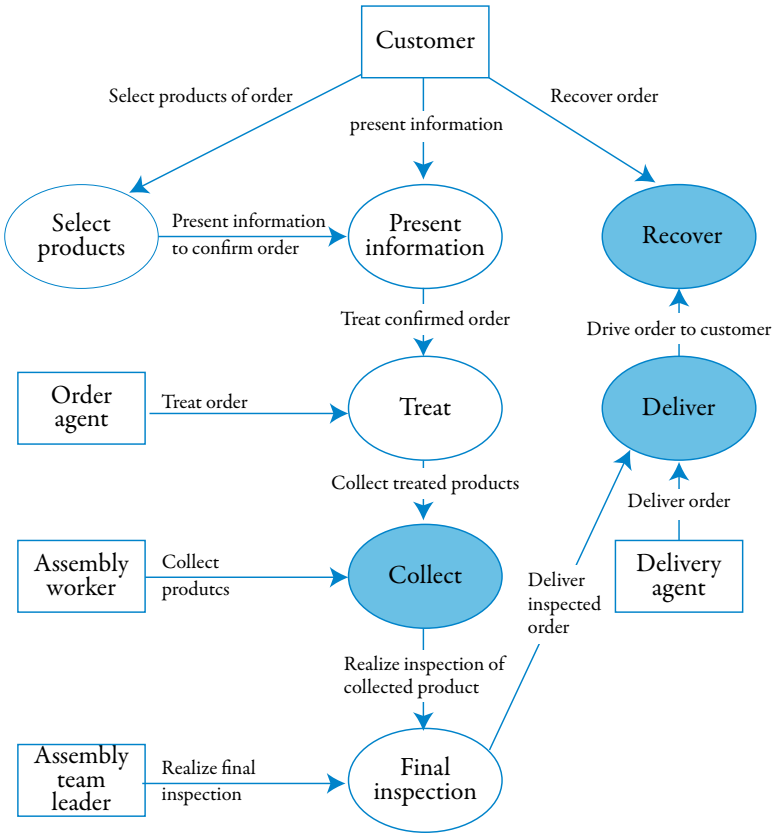

Figure 7. Data flow diagram model of «sales through e-commerce».

Figure 7 shows the business process model represented by Data Flow Diagram. We just specified processes and their sequence to present a business process in general. We tried to present the maximum of business actors to define a true business process, in which there is collaboration between several business actors. For example, instead of putting a single external «delivery service», we identified the externals: «assembly worker», «assembly team leader» and «delivery agent».

Figure 8 shows the second model on CIM level as a model of activity diagram. Through this model we individually detail each process of the previous model as several actions. However, in this model the process «select product for order» is analyzed. Also, we have identified all possible ways towards connections. Then we presented an object node with its state in the output of each action.

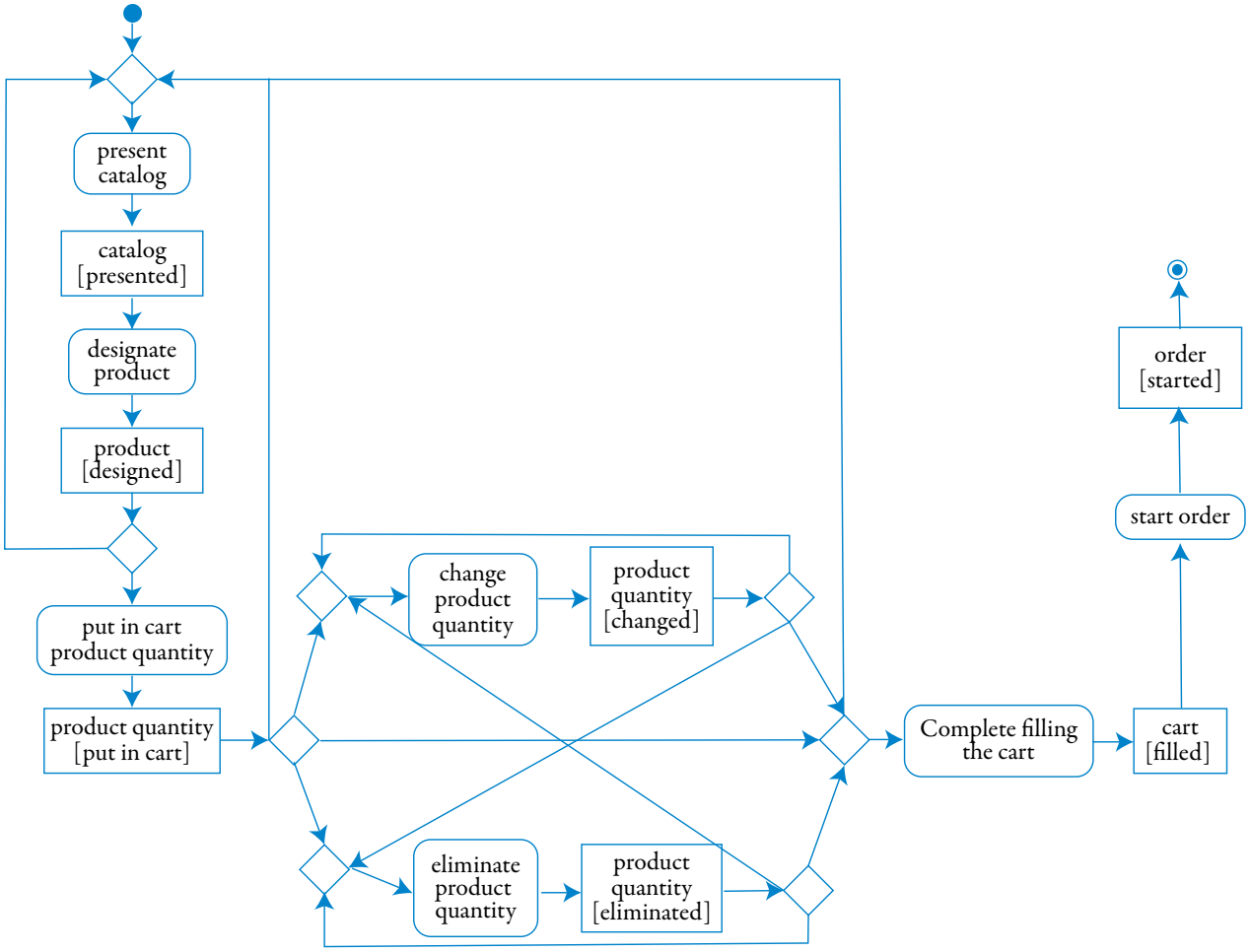

Figure 8. Detailed model of activity diagram of «select products for order"

3.2 Presentation of the PIM level

Figure 9 illustrates a model of use case diagram. In this model, the process «select product for order» -model of DFD- is transformed into a package. Then, the collaborator «customer» who performs the processes becomes actor. Then the actions that detail the processes in the model of the activity diagram are transformed to use cases. Decision nodes that lie between two actions become relationship «extend». However, control flows that lie between two actions become relationship «include."

Figure 10 shows state diagram model transformed from the model of activity diagram of CIM. In this model the states are obtained from nodes of objects. Then, the control flow, which connects two actions, is transformed into a transition. E.g. the object node «catalog» with state «presented» becomes «catalog presented» in state diagram model. Then, initial state is transformed as initial node; final node becomes a final state; node fusion is transformed to junction point; decision node becomes a decision point and fusion node is transformed into a junction point.

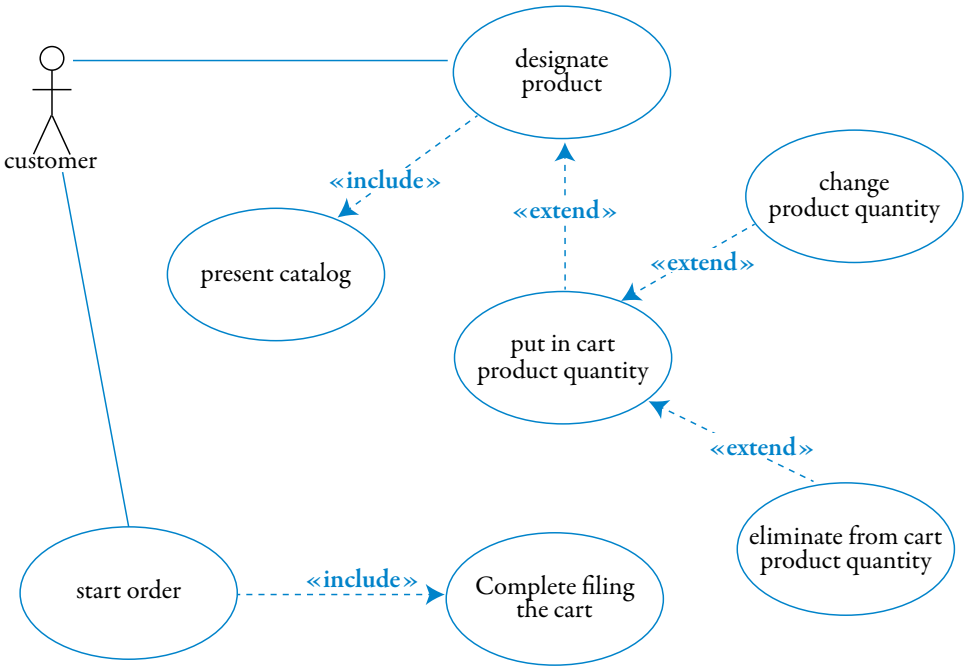

Figure 9. Use case diagram model of «select products of order».

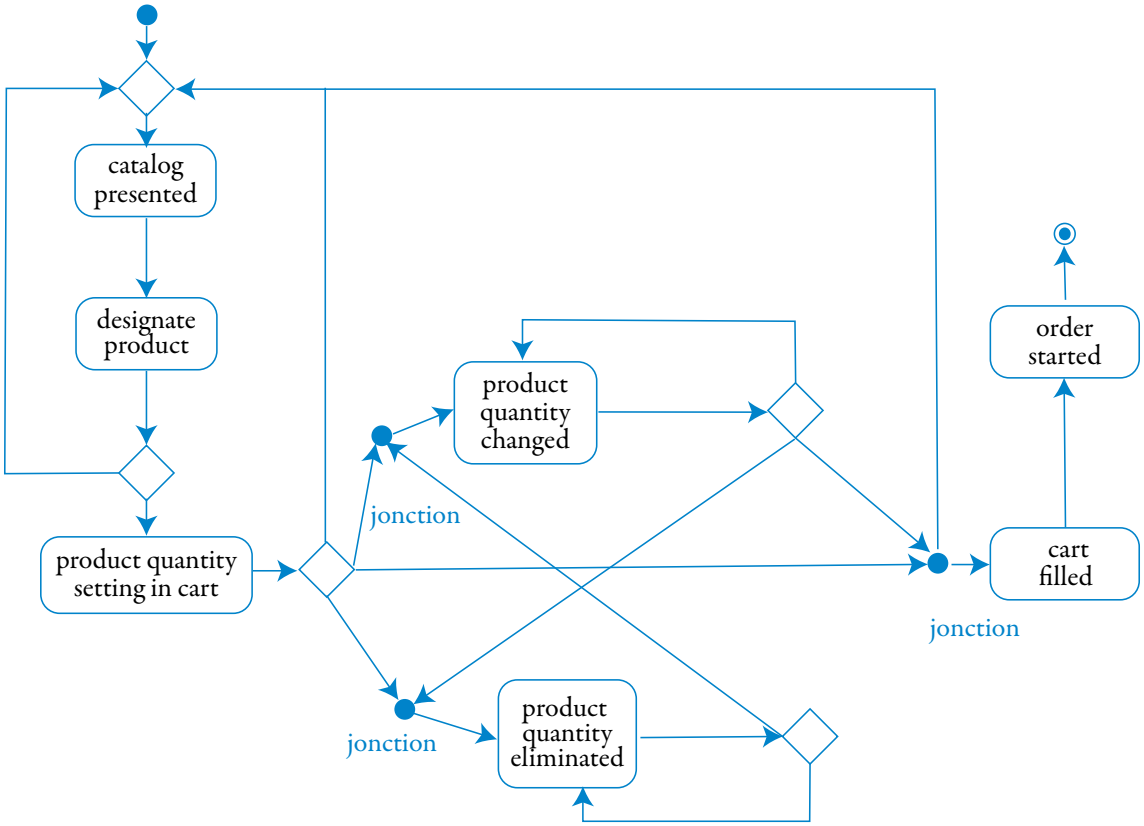

Figure 10. Model of state diagram of «select products of order"

Figure 11 shows the final objective of the PIM level that is the construction of a model of class diagram. This model is transformed from the model of the activity diagram. In this model the classes are transformed from object nodes. Then the states of an object are transformed into functions of the class. So the object node «order» with state «started» is transformed into class «order» that contains the «start» method. Figure 11 shows a model of the package diagram. So, the process «treat order» and «final inspection» become packages.

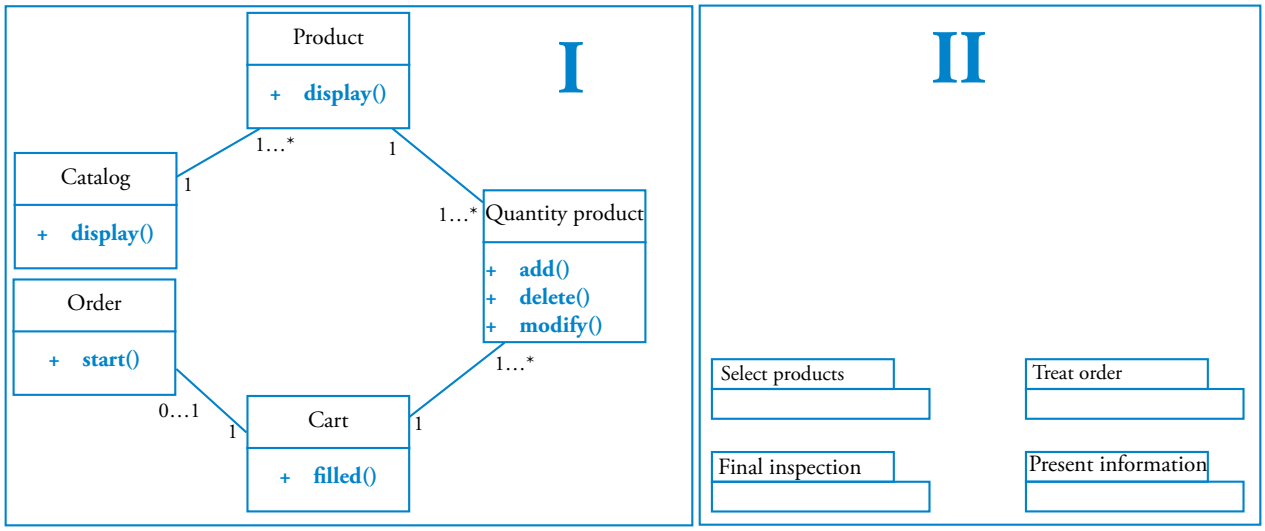

Figure 11. I: model of class diagram of «select products of order», II: model of package diagram of «sales through e-commerce»

This section provides a case study that was presented to validate construction rules of CIM level and transformation rules into the PIM level. In the next section we will present the difference between our proposal and the related works.

4. Related works

MDA does not provide indication about the standards that must be used to model different levels. The key principle of MDA is the use of models in different phases of an application development cycle. Specifically, MDA advocates the establishment of business process models on CIM level, analysis and design models on PIM level and code models on PSM level. Indeed relevant research use business process modeling standards such as BPMN, activity diagram, DFD, to model the CIM level. However most research based on UML diagrams to model PIM level because UML is recommended by MDA on this level.

4.1 Evaluation criteria

Our work is based on (Krouile *et al.*, 2013) to establish valuation criteria. According to (OMG-MDA, 2014), CIM is established by business process models and the PIM level is founded on one or multiple design models. Then, transformation rules allow moving from CIM models to PIM models. According to OMG, it is necessary to model all points of view in order to understand the future system.
The evaluation criteria are CIM business modeling, PIM completeness, transformation elements, and Assessment methodology. The CIM level must show the business process by one or more models of business process standards. The PIM level is complete if it holds one or more models for each modeling view: functional, static and dynamic views. However, the main model is class diagram model, because it contains information system structure and it is easily transformable into PSM.
Model transformations are based on source meta-model, target meta-model and transformation rules.

4.2 Analysis and discussion

On the CIM level, some researches focuses only on system requirement for modeling like in (Gutiérrez *et al.*, 2008). Other hybrid solutions are founded on business process and system requirement for modeling the CIM-level as in (Kherraf *et al.*, 2008). In these approaches, system requirements are early modeled on CIM-level to facilitate transformation towards PIM-level. In our method, we modeled business process on CIM-level by DFD and activity diagram.

On the PIM level, there is no approach that covers the three modeling views except (Kardoš *et al.*, 2010) (Rhazali *et al.*, 2015a) (Rhazali *et al.*, 2015d). Also, several approaches do not model the classes on the PIM level as (Zhang *et al.*, 2005) (Castro *et al.*, 2011) (Hahn *et al.*, 2008) (Mazón *et al.*, 2007) (Gutiérrez *et al.*, 2008), although without classes the source code is not easily obtained by transformation.
Our method covers the three modeling views. Therefore, in static view, we relies on class model, and package model for organizing classes.
All approaches define the transformation rules like in (Kherraf *et al.*, 2008) (Kardoš *et al.*, 2010).
Many standards are used in several approaches for modeling business process including :

- Data Flow Diagram (DFD) (Hoffer *et al.*, 2004) used in (Kardoš and Drozdová, 2010) and (Mokrys, 2012),
- Integration DEFinition (IDEF) (Mayer, 1995),
- XML Process Definition Language (XPDL) applied by (Mokrys, 2012),
- UML 2 Activity Diagram (AD) used in (Rodríguez *et al.* 2010).

Other notations or methodologies for modeling business process are cited by OMG (OMG-BPMN, 2011) including BPMN, Electronic Business using eXtensible Markup Language Business Process Schema Specifications (ebXML BPSS), UML EDOC Business Processes, Activity-Decision Flow (ADF) Diagram, RosettaNet.
We relied on above criteria evaluation, for comparing the CIM to PIM transformation approaches (Figure 12). the static view.

	Business modeling of CIM		**Completeness of PIM**						**Transformation**		
			Functional view		***Static view***		***Dynamic view***		***Source meta-model***	***Target meta-model***	***Transformation rules***
Studied papers	Attained	Representation	Attained	Representation	Attained	Representation	Attained	Representation			
(Kherraf *et al.*, 2008)	Yes	UML activity diagram			Yes	Component & class diagram					Yes
(Zhang *et al.*, 2005)											Yes
(Kardoš *et al.*, 2010)	Yes	DFD	Yes	UML Use Case & Sequence Diagram	Yes	Domain diagram	Yes	UML Activity diagram			Yes
(Rodríguez *et al.*, 2010)	Yes	BPMN & UML activity diagram	Ye	UML Use Case diagram	Yes	UML class diagram					Yes
(Castro *et al.*, 2011)	Yes	BPMN	Yes	UML Use Case diagram			Yes	UML Activity diagram	Yes	Yes	Yes
(Hahn *et al.*, 2010)	Yes	BPMN									Yes
(Mazón *et al.*, 2007)											Yes
(Gutiérrez *et al.*, 2008)							Yes	UML Activity diagram	Yes	Yes	Yes
(Rhazali *et al.*, 2015a)	Yes	BPMN	Yes	UML Use Case diagram	Yes	UML class diagram					Yes
Rhazali *et al.* 2015b)	Yes	BPMN & UML activity diagram	Yes	UML Use Case diagram	Yes	UML class diagram	Yes	UML state diagram			Yes
Rhazali *et al.* 2015c)	Yes	UML activity diagram			Yes	UML class diagram	Yes	UML state diagram			Yes
Rhazali *et al.* 2016a)	Yes	BPMN	Yes	UML Use Case diagram	Yes	UML class diagram	Yes				Yes
Our proposal	Yes	DFD & activity diagram	Yes	UML Use Case diagram	Yes	UML class & package diagram	Yes	UML state diagram	Yes	Yes	Yes

Figure 12. Comparison of studied papers through Evaluation criteria

We base on tree categories and each category contains some criteria :

- the category «business modeling CIM» contains the criteria «attained», «representation» and «simplicity»,

- the category «completeness of PIM contains the criteria «functional view», «static view» and «dynamic view», and each criterion contains two sub-criteria «attained» and «representation»,
- the category «transformation» contains the criteria: «source meta-model», «target meta-model» and «transformation rules».

The gray cell means that the criterion is not verified by the approach.
The principal key in MDA approach is the models transformation. Indeed, in MDA there are two elementary transformation kinds: CIM (computing independent model) to PIM (platform independent model) transformation and PIM to PSM transformation. However, most researches propose approaches transforming PIM to PSM (platform specific model), since there are multiple points in common between PIM level and PSM level. Nevertheless, transforming CIM level into PIM level is rarely addressed in research because these two levels are mainly different. All related works propose approaches to control model transformation from CIM to PIM according to MDA. However, our methodology is obtained from an analytical survey. Indeed, from the beginning, in CIM, we consider that we build business process models, which will be automatically transformed into PIM models. Our methodology is based on creating good CIM models, through well-defined construction rules, to facilitate transformation toward the PIM models. So, we establish a rich PIM level, considering the three classical modeling views : dynamic, functional and static. Use case diagram model interprets functional view, state machine diagram model represents the dynamic view, class and package diagram models show

5. Conclusions and future works

One of the main challenges in the software development process is the establishing of an approach that allows moving from models that define the business process to models which represent the analysis and design of software. Based on MDA, our approach provides a solution to the problem of transformation from business models represented on CIM level to design models represented on PIM level. This methodology results in a set of well-structured and useful classes in the process of software development. In this approach, we benefited from our experience in old transformation methods (Rhazali *et al.*, 2015a) (Rhazali *et al.*, 2015d) (Rhazali *et al.*, 2016a) (Rhazali *et al.*, 2015b) to provide, through semi-automatic transformation, a set of classes structured in packages from business models defined by DFD and activity diagram.
Our approach does not transform all notations existing in business process models into PIM level models such as the representation of the fusion node in the use case diagram, which is not considered by our work. Furthermore, several important elements on the PIM level are not automatically transformed from the CIM level like class properties and class relationships. Nethertheless, our approach provided a transformation methodology with some significant transformation rules for solving the transformation problem from CIM level to PIM level.
We aim to improve our approach in future works. In particular, the ongoing work is intended to improve the construction rules of CIM level and the transformation rules from CIM to the PIM in order to implement these transformations in a tool via the QVT. In addition, we plan to transform the models obtained on the PIM level to PSM models; indeed our ultimate goal is to provide the source code from the business models through automatic transformation.

6. References

Blanc, X. (2005). *MDA in action*. Ed. Eyrolles.
De Castro, V., Marcos, E., & Vara, J.M. (2011). *Applying CIM-to-PIM model transformations for the service-oriented development of information systems*. Journal of Information and Software Technology, 53 (1), 87-105.
Gordijn, J., & Akkermans, J.M. (2003). *Value based requirements engineering: exploring innovative e-commerce idea*. Requirements Engineering Journal, 8 (2), 114–134
Grammel, B., & S. Kastenholz, A. (2010). *Generic traceability framework for facet-based traceability data extraction in model-driven software development.* in Proceedings, 6th ECMFA Traceability Workshop held in conjunction ECMFA. (pp. 7-14). Paris, France.
Gutiérrez, J.J., Nebut, C., Escalona, M.J., Mejías, M., Ramos, I.M. (2008). *Visualization of use cases through automatically generated activity diagrams*. In Proceedings of the 11th International Conference MoDELS'08, Toulouse, France.
Hoffer, J.A, George, J.F, Valacich, J.S. (2004). *Modern system analysis and design*. Prentice Hall ISBN 0-13-145461-7, 2004.
Kardoš, M., Drozdová, M. (2010). *Analytical method of CIM to PIM transformation in Model Driven Architecture (MDA)*, Journal of information and organizational sciences, vol. 34, pp. 89-99.

Kherraf, S., Lefebvre, É., Suryn, W., (2011). *Méthodologie de transformation du CIM en PIM dans l'approche MDA*. Thèse de doctorat électronique, Montréal, École de technologie supérieure.

Kherraf, S., Lefebvre, É., Suryn, W. (2008). *Transformation from CIM to PIM using patterns and Archetypes*. In ASWEC'08, 19th Australian Software Engineering Conference, Perth, Australia.

Kriouile, A., Gadi, T., Balouki, Y. (2013). *CIM to PIM Transformation: A criteria Based Evaluation*. International Journal Computer Technology & Applications, 4(4), 616-625.

Mayer, R., Menzel, C., Painter, M., Perakath, B., de Witte P. and Blinn T. (1995). *Information Integration For Concurrent Engineering (IICE) - IDEF3 Process Description Capture Method Report*. Technical Report September 1995. available at http://www.idef.com/pdf/idef3_fn.pdf

Mazón, J., Pardillo, J., Trujillo, J. (2007). *A model-driven goal-oriented requirement engineering approach for data warehouses*. In Proceedings of the Conference on Advances in Conceptual Modeling: Foundations and Applications, ER Workshops, Auckland, New Zealand, pp. 255–264.

Mokrys, M. (2012). *Possible transformation from Process Model to IS Design Model*. In First International Virtual Conference Slovakia, pp. 71–74.

OMG-BPMN. (2011). *Business Process Model and Notation (BPMN)-Version 2.0*. Boston, USA: OMG.

OMG-MDA. (2014). *Object Management Group Model Driven Architecture (MDA) MDA Guide rev. 2.0*. Boston, USA: OMG.

OMG-QVT. (2015). *Meta Object Facility (MOF) 2.0 Query/View/Transformation Specification, V1.2*. Boston, USA: OMG.

OMG-SoaML (2012). *Service Oriented Architecture Modeling Language (SoaML) – Specification for the UML Profile and Metamodel for Services (UPMS)*. OMG document: ad/2012-05-10. Available at <**http://www.omg.org/spec/SoaML/1.0.1/PDF**>.

OMG-UML. (2011). *OMG Unified Modeling LanguageTM (OMG-UML)*, Infrastructure, **http://www.omg.org /spec /UML/2.4.1/Infrastructure**. August 2011.

Osis, J., Asnina, E., & Grave, A.. (2007). *Formal Computation Independent Model of the Problem Domain within the MDA*. ISIM.

Rhazali, Y., Hadi, Y., & Mouloudi, A. (2015a). *A Methodology of Model Transformation in MDA: from CIM to PIM*. International Review on Computers and Software, 10 (12), 1186-1201. DOI: http://dx.doi.org/10.15866/irecos.v10i12.8088.

Rhazali, Y., Hadi, Y., & Mouloudi, A. (2015b). *A Methodology for Transforming CIM to PIM through UML: From Business View to Information System View*. In Proceedings Third World Conference on Complex Systems. Marrakech, Morocco. DOI: 10.1109/ICoCS.2015.7483318.

Rhazali, Y., Hadi, Y., & Mouloudi, A. (2015c). *Disciplined Approach for Transformation CIM to PIM in MDA.* In Proceedings, 3rd International Conference on Model-Driven Engineering and Software Development. (pp. 312 – 320). Angers, France.

Rhazali, Y., Hadi, Y., & Mouloudi, A. (2015d). *Transformation Approach CIM to PIM: From Business Processes Models to State Machine and Package Models*. In Proceedings, the 1st International Conference on Open Source Software Computing. (pp. 1 – 6). Amman, Jordan. DOI: 10.1109/OSSCOM.2015.7372686.

Rhazali, Y., Hadi, Y., & Mouloudi, A. (2016a). *A Based-Rule Method to Transform CIM to PIM into MDA.* International Journal of Cloud Applications and Computing, 6 (2). DOI: 10.4018/IJCAC.2016040102.

Rhazali, Y., Hadi, Y., & Mouloudi, A. (2016b). *CIM to PIM Transformation in MDA: from Service-Oriented Business Models to Web-Based Design Models*. International Journal of Software Engineering and Its Applications, 10 (4), 125-142. DOI: 10.14257/ijseia.2016.10.4.13.

Rhazali, Y., Hadi, Y., & Mouloudi, A. (2016c). *A New Methodology CIM to PIM Transformation Resulting from an Analytical Survey*. In Proceedings of the 4th International Conference on Model-Driven Engineering and Software Development. (pp. 266-273). Rome, Italy. DOI: 10.5220/0005690102660273

Rhazali, Y., Hadi, Y., & Mouloudi, A. (2016d). *Model Transformation with ATL into MDA from CIM to PIM Structured through MVC*. Procedia Computer Science - Journal - Elsevier. doi:10.1016/j.procs.2016.04.229.

Rodríguez, A., García-Rodríguez de Guzmán, I., Fernández Medina, E., Piattini, M., 2010. *Semi-formal transformation of secure business processes into analysis class and use case models: an MDA approach.* Information and Software Technology 52 (9) (2010) 945–971.

Roques, P., 2004, *UML in Practice: The Art of Modeling Software Systems Demonstrated through Worked Examples and Solutions*. Wiley.

Schmidt, D.C. (2006). *Guest Editor's Introduction: Model-Driven Engineering*. IEEE Computer, 39 (2), 25 - 31.

Zhang,W., Mei, H., Zhao, H., & Yang, J. (2005). *Transformation from CIM to PIM: A Feature-Oriented Component-Based approach*. In Proceedings MoDELS. (pp. 248-263). Montego Bay, Jamaica.

Recherche & Développement

Réalisation d'un classeur pédagogique numérique

Méthodologie et contexte

The realization of an educational workbook software

Methodology and context

Ali Sadiqui
ISTA Meknes, OFPPT, Meknes, Maroc
sadiqui2000@yahoo.fr

Résumé

Utilisées pour concevoir des applications en impliquant pas à pas le client, les méthodes agiles sont de plus en plus adoptées par les équipes de développement. C'est dans ce cadre théorique qu'a été conçu et piloté le projet objet de cette contribution.
Ledit projet est motivé par le fait que depuis quelques années les formateurs appartenant à l'Office de la Formation Professionnelle et de la Promotion du Travail (Maroc) recourent quotidiennement à un classeur pédagogique sur support papier pour organiser les activités pédagogiques relatives à la formation, à l'évaluation et à la présence des étudiants. Étant sur support papier, l'utilisation dudit classeur se révèle fastidieuse en ce sens qu'elle exige un temps considérable et, de ce fait, affecte négativement la progression de l'apprentissage. Il a fallu donc penser à alléger son usage.
Nous avons alors, en collaboration avec une équipe de praticiens et avec le client effectif, conçu et expérimenté une version numérique de cet outil pédagogique. Le produit final a impacté positivement la gestion de la classe.

Abstract

Used to design applications involving step by step the customer, agile methods are increasingly adopted by development teams. It is within this theoretical framework that has been designed and piloted the project purpose of this contribution.
The said project is motivated by the fact that in recent years the trainers belonging to the Office of Vocational Training and Work Promotion (Morocco) daily have recourse to a paper educational workbook to organize educational activities related to training, evaluation and student's presence. Being on paper, it requires considerable time and, thus, affects negatively the learning progress. It was, therefore, necessary to think about simplifying its use.
So we, in collaboration with a team of practitioners and the actual customer, have designed and tested a digital version of this educational tool. The final product has impacted positively classroom management.

Mots-clés

classeur pédagogique, logiciels pédagogiques, logiciels libres, Outils didactiques, outils pédagogiques, méthodes Agiles, Extreme-Programming.

Keywords

paper educational workbook, educational software, free software, didactic tool, educational tools, Agile Extreme Programming.

1. Introduction

Placer le client au cœur du processus de développement des logiciels est le fondement des méthodes de développement dites «agiles» (Houy *et al*, 2013) ; (Cantone et Marchesi, 2014) ; (Vickoff, 2009) ; (Abbas *et al*, 2008). Plusieurs méthodes se regroupent sous cette bannière. Elles sont fondées sur un développement itératif et incrémental dans lequel la recherche de solutions aux problèmes rencontrés s'appuie sur une collaboration de pair à pair. Ces méthodes définissent des groupes de pratiques qui visent à favoriser le travail avec les spécificités de tout un chacun intégré dans la création du logiciel. Le développement d'un système informatique devient alors une activité motivante où chaque membre se voit confier une part de responsabilité.
La présente étude a pour objectif la réalisation d'un logiciel informatique qui permet d'élaborer un classeur pédagogique selon les exigences de la démarche agile et de l'organisme demandeur.
Dans la seconde section de cet article, nous présenterons les méthodes agiles et plus particulièrement la méthode Extreme Programming. Puis, dans la troisième section, nous présenterons le contexte de notre projet, incluant une brève présentation de l'organisme concerné, son système de gestion et le classeur pédagogique utilisé. Présenter la solution adoptée pour créer ledit logiciel et les résultats obtenus fera l'objet de notre quatrième et dernière section.

2. Les méthodologies Agiles

Les méthodes traditionnelles se fondent sur un enchaînement séquentiel des différentes étapes de développement, depuis les spécifications jusqu'à la validation du système, selon un planning préétabli. Elles s'efforcent à définir toutes les exigences et les spécifications dès le début supposant que celles-ci restent immuables. Cette vision rassurante est cependant bien loin de la réalité des projets. En effet, des changements peuvent intervenir suite à une modification des besoins du client, ou bien suite à des erreurs découvertes lors de la phase de conception ou lors de l'implémentation du système. Il devient évident qu'un logiciel est a priori un exercice difficile, sauf dans le cas d'applications extrêmement simples ou de rares contextes connus et maîtrisés.
Les méthodologies classiques sont fondées sur le principe que le coût relié à une modification du logiciel augmente d'une manière exponentielle avec le temps. Et par conséquent, il est judicieux de définir tous les aspects du produit et de concentrer la plupart des décisions avant de procéder à sa réalisation.

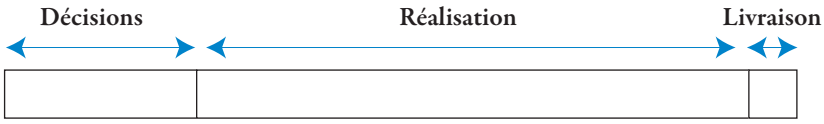

Figure 1. Processus de développement dans les approches classiques

Si les méthodologies classiques sont peu ouvertes au changement, les méthodes agiles, par contre, se proposent de réserver un accueil favorable au changement partant du fait que ce dernier est une composante incontournable dans le processus de développement. Elles partent du fait qu'il est plus rentable et plus pertinent de prendre les décisions progressivement et le plus tard possible au lieu de chercher à les spécifier et les concevoir complètement dès le départ. Pour ce faire, la prise de décision sera faite tout au long du projet grâce à des cycles de réunions appelés itérations.

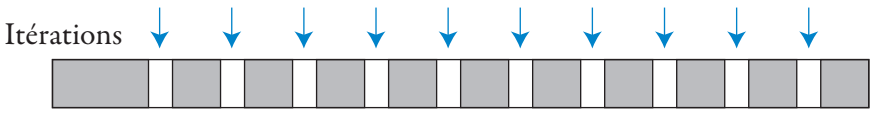

Figure 2. Processus de développement dans les approches agiles

En conséquence, le grand gagnant de cette approche est d'abord le client qui se trouve intégré dans tout le processus de développement. A chaque itération, il établit lui-même les fonctionnalités à implémenter, collabore avec l'équipe pour définir tous les détails et reçoit une nouvelle version du logiciel qui intègre les mises à jour en question.
L'une des méthodes Agiles les plus connues est la méthode Extreme Programming (XP) (Lassenius *et al*, 2015) ; (Rao *et al*, 2014) ; (Beck et Andres, 2005) ; (Cros, 2004) ;. Elle a été principalement développée par Kent Beck et Ward Cunningham. Cette méthode est définie par un ensemble de pratiques qui visent à coordonner le travail d'une équipe de développement. Elle se caractérise par les quatre éléments suivants : des cycles de développement courts, de tests intenses et une validation en continu, en plus d'une collaboration totale du client tout au long du processus de développement.

Notre projet, baptisé «PLANIFORM» (la planification de la formation), est une application concrète de ladite méthodologie. En effet, il a fait intervenir plusieurs participants dans les domaines de l'éducation et de l'informatique et a tracé pour objectif l'élaboration d'un projet informatique qui doit répondre à certaines exigences, d'ordre administratif, technique et pédagogique, de l'organisme demandeur, et selon la démarche Extreme Programming dont l'intérêt n'est plus à démontrer.

3. Le contexte du projet

Nous allons, en quelques lignes, faire une brève présentation du contexte qui nous a permis de concevoir et d'expérimenter notre projet.

3.1. Présentation de l'OFPPT

Créé en 1974, l'Office de la Formation Professionnelle et de la Promotion du Travail (OFPPT) est un établissement public ayant pour vocation d'accompagner les jeunes et les entreprises en matière d'apprentissage et de développement de compétences.
Le parcours de formation qu'il procure permet à son public cible de construire les compétences nécessaires permettant une insertion facile et rapide dans le monde du travail ou de la création d'entreprises.
Il s'adresse également aux demandeurs d'emploi pour les aider à actualiser ou à élargir leurs compétences professionnelles via des formations qualifiantes et aux entreprises, via des formations continues, pour développer les compétences de leurs employés.

3.2. Le système de formation à l'OFPPT

Le régime de formation à l'OFPPT est modulaire et les formations sont proposées selon la hiérarchie «domaine – filière – spécialité». Par exemple la spécialité «Développent Web et Multimédia» est une option de la filière «Techniques de Développement Informatique» appartenant au domaine «NTIC».
Il est à noter que certaines filières sont fusionnées avec leurs spécialités pour des raisons relatives à la simplification de leur gestion.
Chaque module est un ensemble d'informations et de connaissances qui permettent d'acquérir une compétence professionnelle précise reliée à une filière ou une spécialité. Les modules font partie d'un ensemble plus global : la filière ou la spécialité.
Un module est le découpage d'une compétence donnée en un certain nombre d'objectifs qui forment un tout autonome. Les objectifs sont divisés en objectifs de premier niveau et en objectifs de second niveau et commencent toujours par un verbe d'action (Présenter, Définir, etc.). Un module peut être acquis via plusieurs modalités : cours théoriques, travaux pratiques, travaux dirigés, projets ou stages, le tout organisé de manière cohérente, selon une masse horaire prédéfinie et une logique de progression. Ces modules visent l'acquisition d'une ou de plusieurs compétences ciblées.
C'est à travers ces divers modules que les stagiaires acquièrent de nouvelles connaissances qui leur permettent de finaliser l'apprentissage d'un métier donné.
Chaque module s'achève par un contrôle des compétences, une sorte d'examen appelée examen de fin de module (EFM) qui peut être théorique ou pratique.

3.3. Le système de gestion

Dans un établissement relevant de l'OFPPT, l'encadrant est un formateur qui met en œuvre des actions de formation et d'évaluation du public ciblé. Il dispense un ou plusieurs modules, qui relèvent d'une ou de plusieurs filières. Chaque formateur a un emploi du temps hebdomadaire variable durant l'année. Lors de chaque séance, de 2 heures 30 minutes en général, il doit aborder les objectifs préétablis du module assuré.
Le formateur est tenu de préparer une fiche pédagogique pour chaque séance, où il indique les objectifs abordés, les documents et les méthodes pédagogiques utilisés. Il doit aussi mentionner, dans le cahier de textes (cahier journal) les mêmes objectifs ainsi que l'état d'absence des stagiaires. Chaque module est évalué sous forme de contrôles continus et d'un EFM. La formule pour calculer la note du module est donnée comme suit :
Note = ((moyenne des notes des Contrôles Continus)+(2*note EFM))/3
On désire gérer le suivi de la progression du programme ainsi que le suivi des évaluations des modules et des absences des stagiaires.
Toutes ces actions pédagogiques se font actuellement de façon manuelle à travers un outil appelé «le classeur pédagogique» complètement basé sur le support papier.

3.4. Le classeur pédagogique, éléments de composition

Parmi les éléments clés de ce classeur pédagogique, on trouve la fiche **«Planification et suivi de la réalisation des modules de formation»** (Figure 3). Elle est composée de trois zones essentielles :

Zone 1 : indique des informations générales sur le module : filière et groupe concerné, masse horaire, objectifs du module, etc.

Zone 2 : intitulée **«Prévision par séance»**, elle indique le numéro et la date ainsi que les objectifs prévus pour cette séance. Chaque objectif commence par un verbe d'action (présenter, définir, etc.). Les objectifs mentionnés peuvent être des objectifs de premier niveau, ou détaillés en objectifs secondaires. Cette partie devrait faire l'objet d'une conception pédagogique selon un modèle préétabli. Le formateur doit y préciser minutieusement les démarches pédagogiques à mettre en œuvre pour atteindre les objectifs ciblés.

Zone 3 : intitulée **«Réalisation par séance»**, elle indique les objectifs réalisés pour chaque séance. Dans le meilleur des cas, tous les objectifs prévus seront réalisés. Dans le cas contraire (un rythme lent de l'avancement de la séance, un problème technique ou un empêchement imprévu lors de déroulement de la séance, etc.) les objectifs réalisés seront mentionnés et le reste sera programmé pour la prochaine séance.

On établit aussi dans cette zone le cumul des heures réalisées pour ce module, cela a un intérêt particulier étant donné qu'il permet d'indiquer la progression de la réalisation du module.

Dans la zone réservée, on indique la liste des stagiaires absents.

Planification et suivi de la réalisation des modules de formation

Filière :						Module :			
Masse horaire du module :						Nombre de séances :			
1ère année		2ème année		3ème année		Groupe :	Nombre de stagiaires :		
Objectif du module : **ZONE 1**									

Prévision par séance				Réalisation par séance				
N°	Date de la séance	Objectif opérationnel de la séance	Durée	Date	Contenu réalisé	Durée en heures réalisée	Durée en heures cumul	Stagiaires absents
					A prévoir pour la prochaine séance :			
		ZONE 2			**ZONE 3** A prévoir pour la prochaine séance :			
					A prévoir pour la prochaine séance :			
					A prévoir pour la prochaine séance :			

Figure 3. la fiche «Planification et suivi de la réalisation des modules de formation»

3.5. Cahier des charges du projet

Il fallait penser à la conception et la mise en œuvre d'un logiciel permettant de mieux gérer toutes ces informations pendant toute l'année scolaire et pour tous les publics ciblés. Notre référence est la norme ISO 9126, qui définit la qualité d'un logiciel, que nous avons essayé d'appliquer en prenant en considération le contexte présenté auparavant. Cela a permis de fixer certaines de ses fonctionnalités :

La capacité fonctionnelle

C'est la capacité d'un logiciel à répondre aux exigences et aux besoins explicites ou implicites des usagers. En effet, le projet doit respecter tous les éléments existant dans le classeur pédagogique exigé par la Direction Générale, sans aucune modification, en respectant aussi le contenu et la forme de tous les états de sorties.

La facilité d'utilisation

C'est la capacité d'un logiciel à être manipulé sans demander beaucoup d'efforts qui pourraient entrainer son rejet. Pour cela, le projet doit être facile à apprendre et à exploiter. En plus, une utilisation incorrecte d'un débutant ne doit pas entraîner son dysfonctionnement.

La fiabilité

C'est la capacité d'un logiciel à rendre des résultats corrects : tous les calculs effectués (le cumul des heures supplémentaires, les moyennes des notes des modules des stagiaires, etc.) doivent respecter les particularités et les exigences de l'OFPPT.

La maintenabilité

C'est la capacité d'un logiciel à être facilement modifiable. Le projet doit être extensible et nécessiter peu d'efforts pour celui qui veut y ajouter de nouvelles fonctions.

La portabilité

C'est la capacité d'un logiciel à fonctionner dans un environnement matériel ou logiciel différent de son environnement initial.

4. La conduite du projet

4.1. Présentation de l'équipe

Notre équipe fut formée de cinq personnes, toutes exerçant au sein de l'OFPPT. Chacun des membres a pris un rôle en conformité avec la méthode XP. Le projet a été piloté par un coordonnateur (Manager et Coach), qui avait pour rôle de planifier les réunions et de coordonner les efforts. Faisaient aussi partie de l'équipe un vérificateur (Tracker), deux développeurs et un testeur.

4.2. Expérimentation du logiciel

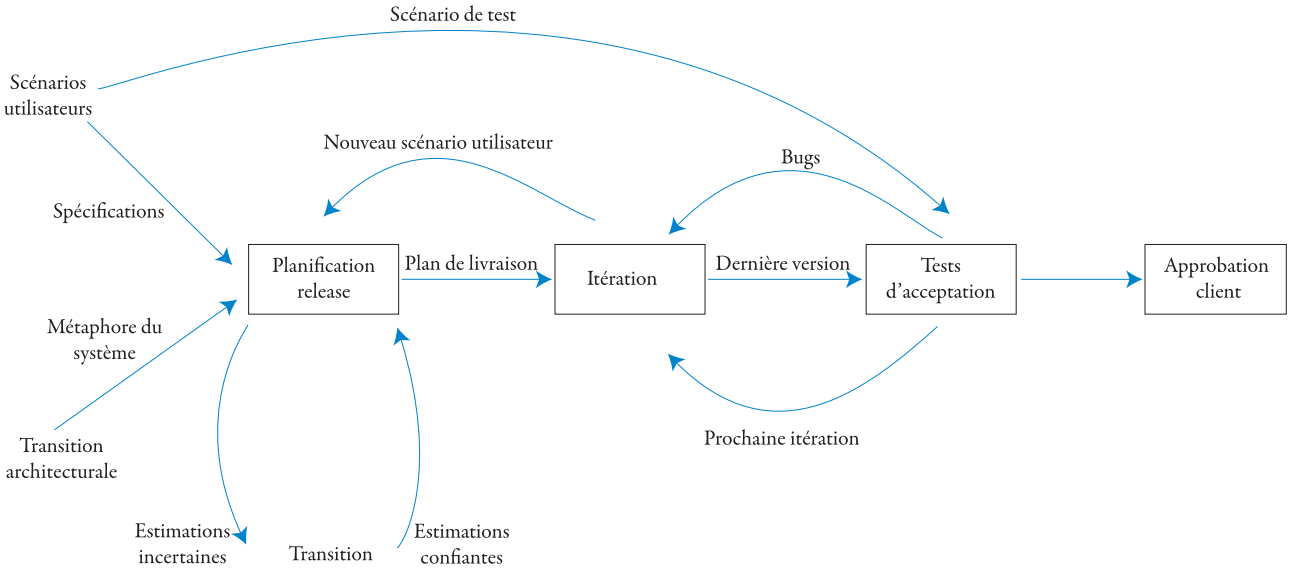

Figure 4. Cycle de développement de la méthode «Extreme Programming».

Le projet a fait aussi intervenir une trentaine de formateurs, tous exerçant à l'OFPPT. Leur domaine d'intervention au sein de notre organisme inclut toutes filières et tous niveaux confondus.

Le projet avait démarré par une phase d'exploration d'un mois, qui avait pour objectifs de définir le contenu fonctionnel de l'application, établir un premier plan de développement pour le projet, et produire la toute première version du logiciel. Le fait que notre équipe ainsi que les formateurs étaient tous disponibles à l'établissement durant toute la semaine nous a beaucoup facilité le pilotage du projet.

La planification du projet a été réalisée conjointement avec les formateurs au cours de séances dédiées, organisées régulièrement tout au long du projet.

Les itérations ont été fixées à deux semaines et les réunions ont été établies dans les établissements de formation. L'équipe a adopté des horaires qui lui permettent de conserver tout au long du projet l'énergie nécessaire pour produire un travail de qualité et mettre en œuvre efficacement les régulations qui s'imposaient.

4.3. Résultats obtenus

4.3.1. Le feed-back des formateurs

Notre projet a été développé et amélioré en référence à la fois aux objectifs sus cités et à l'analyse des pratiques des formateurs. Durant toute une année, notre équipe leur livrait des versions du logiciel (frequent releases) pour qu'il soit adapté et optimisé à leurs besoins et pour leur offrir une réactivité et un confort d'utilisation maximal sans oublier la correction des erreurs et des anomalies produites.
Il est à noter que les documents du classeur pédagogique, sur support papier, restent toujours des éléments exigés par l'administration pour tout contrôle et suivi d'avancement du programme. Cependant, ils sont imprimables sous forme d'états de sorties pour être classés sur ledit classeur.
Notre projet ne visait donc pas à remplacer l'ancien système, mais se fixait plutôt comme objectif de doter les formateurs d'une solution souple pour la gestion et la production desdits documents.
En outre, le produit final a permis à tous les formateurs impliqués dans cette expérience de gérer leurs emplois du temps, de faire le suivi des heures supplémentaires et de planifier les dates des EFM, entre autres.

4.3.2. Présentation de PLANIFORM

Les différentes réflexions menées lors de la conception de ce projet nous ont orientés vers le choix de le doter de 6 modules : Vues, Groupes, Séquences, Listes des stagiaires, États, et Évaluations.
Les modules ont été développés simultanément, car ils définissent le contenu fonctionnel minimal d'une application utilisable par les formateurs. Cependant, ils ont subi des améliorations ainsi que des corrections durant la période du projet.
Le développement de ces modules s'est basé sur des composants libres et dont le code source est disponible. Ce dernier a été modifié et adapté selon les besoins de notre projet.

Le module «Vues»

Il permet de gérer les emplois de temps, les jours d'absence, les jours fériés et les congés de maladie, etc. il permet aussi de gérer les contenus des séances. En effet, nous avons la possibilité de mettre en forme le texte saisi, de lui ajouter des schémas, des documents, des liens, etc., de réexploiter ce contenu d'un groupe à l'autre, et de marquer les absents.

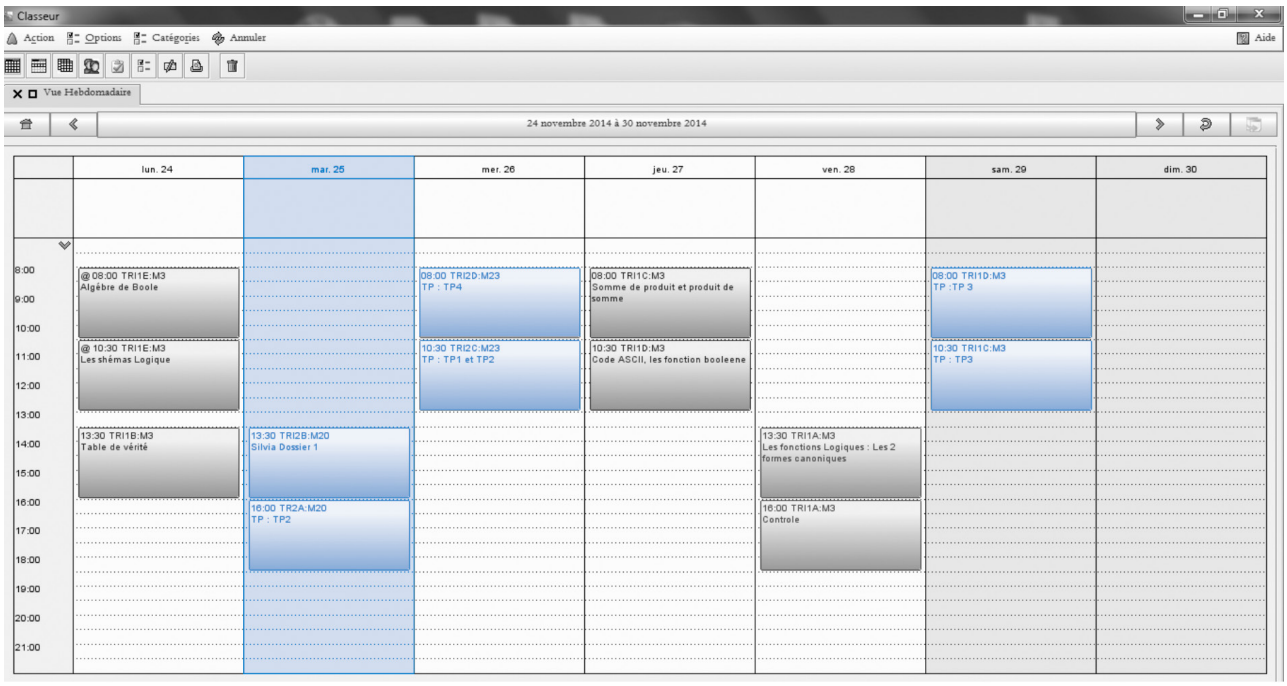

Figure 5 : Capture écran de la vue «hebdomadaire»

Le module «Séquences»

À partir des objectifs établis par la Direction de la Recherche et de l'Ingénierie de la Formation (DRIF), les formateurs doivent élaborer des séquences pédagogiques qu'ils vont suivre tout au long de l'année, les segmenter et les organiser en activités pédagogiques. Ce module leur permet de faciliter la création et l'utilisation des séquences pédagogiques. En effet, nous avons la possibilité de mettre le lien avec le guide pédagogique élaboré par la DRIF. Ce module permet aussi d'échanger les préparations des séquences avec d'autres collègues pour les améliorer.

Le module «Groupes»

Il permet de gérer les groupes et les modules, planifier les dates des EFM et faire le suivi de la progression du programme annuel.

Le module «Listes des stagiaires»

Il permet de gérer les stagiaires, importer la liste des stagiaires pour chaque groupe et avoir un accès rapide aux absences des stagiaires et à leurs évaluations.

Le module «États»

Il permet d'imprimer les préparations pédagogiques pour chaque module, les documents du classeur pédagogique, le bilan d'absence annuelle ou par module des stagiaires et l'emploi du temps d'une période définie.

Le module «Évaluations»

Il permet de gérer les évaluations, saisir les notes des évaluations et les imprimer.

4.4. Les composants utilisés

Deux composants libres essentiels ont été utilisés pour le développement de notre projet à savoir «Borg» et «Jasperreports». Ce choix est justifié par le fait que ces derniers ont répondu à plusieurs recommandations notamment le type de licence, la disponibilité du code source et d'autres facteurs comme la plateforme requise, le langage de programmation ayant servi à sa création, le type de base de données mis en œuvre, etc.
En conséquence, le logiciel qui constitue la pierre angulaire de notre projet et qui a été développé en langage Java est un logiciel libre. Nous avons veillé à ce que le logiciel et les outils sus cités suivent les mêmes termes des licences General Public License (GNU).

4.5. Présentation de l'outil «Borg»

L'outil Borg est un outil libre qui combine un calendrier et un système de suivi de projets personnels. Le calendrier prend en charge toutes sortes de rendez-vous, de gestion des événements et des listes de tâches simples à réaliser. Il fait partie des logiciels appelés les gestionnaires d'informations personnelles ou de projets (PIM), tels que Microsoft Outlook, Mozilla Calendar, Palm Desktop, Yahoo Calendar, etc. Ces logiciels proposent des fonctionnalités agenda, bloc-notes, post-it, rendez-vous, carnet d'adresses. Les versions professionnelles peuvent aussi être associées à des fonctionnalités de courrier et de téléphonie (VoIP, SMS), couplés à des logiciels de gestion de la relation client et de cartographie. Elles peuvent également servir les multi-utilisateurs de façon à partager des messageries, des calendriers et des emplois du temps de réunions.
Notre projet hérite de certaines des fonctionnalités de cet outil. Cependant, son code source était largement modifié et revu étant donné que les objectifs escomptés diffèrent en grande partie de ceux ayant fait l'objet de sa création.
Plusieurs fonctionnalités lui ont été intégrées et d'autres ont été abandonnées. Les attributs de classes ont été également revus et, par conséquent, la conception de la base de données a été modifiée. Et cela pour l'adapter à nos besoins et pour mener à bien notre projet. Nous estimons que seulement 30% du code source original a été conservé.
C'est sur la base de ce composant que nous avons développé les modules : Vues, Groupes, Séquences, Listes des stagiaires et Évaluations.

4.6. Présentation de JasperReports

JasperReports est un outil libre de création d'état de sortie, offert sous forme d'une bibliothèque qui peut être intégrée dans tous types d'applications Java. Cet outil se base sur des fichiers XML (dont l'extension est en général .jrxml) pour la présentation des états. Il peut être couplé à d'autres outils (iReport ou JasperStudio par exemple) pour faciliter sa mise en œuvre dans une application Java, classique ou orientée web.

Cet outil nous a servi pour produire les documents du classeur pédagogique. Ces derniers peuvent être lus, directement dans l'application, imprimés ou exportés dans une variété de formats de documents, y compris HTML, PDF, Excel, OpenOffice et Word. Sa facilité d'utilisation et d'intégration nous a permis de réduire considérablement le temps et l'effort de développement.
Ce composant nous a aidés à développer le module États.

4.7 Autres détails à préciser

- Le projet est disponible sur les liens :
  - `http://www.mediafire.com/download/46gjogchlu67jpo/PlaniformSetup.exe`
  - `https://mon-partage.fr/f/nSiVKEeT/`
- La documentation de PLANIFORM est directement intégrée au logiciel. Un fichier pdf permet d'expliquer les différentes étapes pour son utilisation. De plus, la vidéo disponible sur le lien : `https://www.youtube.com/watch?v=bUTJJXnwn6c` permet de combler tout besoin d'information. Deux fichiers Excel sont aussi disponibles pour faciliter la procédure d'importation.
- La base de données choisie est le «H2» vu sa facilité d'être intégrée dans notre projet. Elle est déjà intégrée dans l'outil «Borg» et suit les mêmes termes de licence.
- Une copie des termes de licence est intégrée dans la rubrique «Licence» sous le menu Aide.
- Avant la version 2, le projet a pris le nom de «Classeur».

5. Conclusion

PLANIFORM est un outil simple à utiliser, pratique et intuitif, destiné à l'élaboration du classeur pédagogique selon les exigences de notre établissement. Par ailleurs, le projet prend en considération les paramètres théoriques auxquels nous nous sommes référés. Il permet de réduire considérablement le temps et l'effort pour la réalisation des documents afférents et, par conséquent, permet au formateur de se focaliser sur l'approfondissement des contenus pédagogiques et sur l'amélioration des compétences des étudiants.
Ce projet peut facilement être intégré dans d'autres contextes similaires (enseignement général, supérieur, etc.) étant donné que dans sa conception, il a été tenu compte de plusieurs facteurs qui peuvent se retrouver partiellement ou totalement dans ces contextes.
Le projet va, dans la prochaine étape, bénéficier d'autres fonctionnalités visant à offrir plus de satisfaction aux différents intervenants. Nous projetons, en fait, de continuer à nous investir dans des recherches participatives qui permettraient d'élargir et d'optimiser son utilisation dans divers contextes universitaires.

6. Références

Abbas, N., Gravell, A., Wills, G. (2008). *Historical roots of Agile methods: where did «Agile Thinking» come from?* In Abrahamsson, P., Baskerville, R., Conboy, K., Fitzgerald, B., Morgan, L., Wang X. Agile processes in software engineering and extreme programming. 9th International Conference XP2008, Limerick, Ireland, June 2008, Proceedings in Software Engineering. Limerick: 10 - 14 Juin 2008, Berlin: Springer. 94-103.

Beck, K., Andres, C. (2005). *Extreme programming explained: embrace change* (2e édition). Upper Saddle River NJ : Pearson Education.

Cantone, G., Marchesi, M. (2014). *Agile processes in software engineering and extreme programming*. Proceedings of 15th international conference, XP 2014. Berlin : Springer.

Cros, T., (2004). *Maîtriser les projets avec l'extreme programming : pilotage par les tests-client.* Toulouse : Éditions Cépadues.

Houy T., Fernandez V., Khalil C., (2013). *Les méthodes agiles de développement informatique*. Paris: Presses des Mines.

Lassenius, C., Dingsøyr, T., Paasivaara, M. (2015). *Agile processes, in software engineering, and extreme programming*. Proceedings of 16th international conference, XP 2015, Helsinki: 25-29 Mai 2015, Berlin : Springer.

Rao, G. S., Krishna, C. V., Rao, K. R. (2014). *Extreme Programming for service-based application development architecture*. Proceedings of the 2014 Conference on IT in Business, Industry and Government (CSIBIG). Indore: 8-9 Mars 2014. Mishra, D. K., Sheikh, R., Excellent Publishing Services.

Vickoff, J.P. (2009). *Méthode agile, Les meilleures pratiques, Compréhension et mise en œuvre*. Vanves (Hauts-de-Seine) : Editions QI.

Fiche équipe

Laboratoire Image et Reconnaissance de Formes – Systèmes Intelligents et Communicants (IRF – SIC)

Image and Pattern Recognition – Intelligent and Communicative Systems

Driss Mammass
Ecole Supérieure de Technologie Agadir, Université Ibn Zohr, Agadir – Maroc
mammass@uiz.ac.ma

Hassan Douzi
Faculté des Sciences, Université Ibn Zohr, Agadir – Maroc
h.douzi@uiz.ac.ma

Résumé

Les recherches du laboratoire IRF - SIC sont menées dans trois équipes de recherche (IRF, R2IS et MS2I) et concernent principalement :

- Le traitement de l'image, la reconnaissance de formes et leurs applications : classification, watermarking et applications, analyse de documents, reconnaissance de l'écriture, du caractère et de la signature manuscrite, imagerie satellitaire et applications, analyse vidéo et suivi du mouvement …
- La recherche d'information : document multimédia électronique et typographie numérique, entrepôt de données, fouille des données complexes et de graphes, réseaux sociaux, archivage et indexation de documents en arabe.
- L'ingénierie des systèmes : ingénierie des systèmes d'information, système décisionnel, big data et cloud computing, systèmes temps réel et embarqués, architectures orientées services et systèmes répartis, services (modélisation, conception, QoS, etc.), sécurité des réseaux (fixe et mobile) et des systèmes.
- La modélisation, simulation, interaction et intelligence : environnements informatiques d'apprentissage humain, l'ingénierie de la connaissance et du Web sémantique, la modélisation à base d'agents et la simulation des systèmes sociaux écologiques complexes, le traitement automatique du langage naturel & apprentissage automatique, les logistique et transport et les automates cellulaires pour simuler la dynamique spatio-temporelle.

Abstract

The IRF-SIC laboratory comprises three research teams (IRF, R2IS, and MS2I). Its research themes are the following :

- *Image Processing and Pattern Recognition and their Applications : classification, watermarking and applications, document analysis, handwritten, character and handwritten signature recognition, satellite imagery and applications, videos analysis and motion tracking, etc.*
- *Information Retrieval : electronic and multimedia document, and digital typography; data warehouse, data mining and graphs; social networks, archiving and indexing of Arabic documents*
- *Systems Engineering : information system engineering; business intelligence, big data and cloud computing; Real-time and embedded systems, service-oriented architecture and distributed systems, services (modeling, design, QoS, etc.), systems and network security (fixed and mobile).*
- *Modeling, Simulation, Interaction and Intelligence: Intelligent tutoring systems, knowledge engineering and semantic web, agent-based modeling and simulation of complex ecological social systems, Natural Language Processing and Machine Learning, logistic and transport, and cellular automata for spatio-temporal dynamics simulation.*

Mots-clés

Reconnaissance de formes, traitement d'images, recherche d'information, ingénierie des systèmes, modélisation, simulation, interaction et intelligence.

Keywords

Pattern Recognition, Image Processing, Information Retrieval, Systems Engineering, Modeling, Simulation, Interaction and Intelligence.

1. Présentation du laboratoire

Les recherches de l'équipe IRF remontent à plusieurs années. Elles sont focalisées sur les thèmes liés à l'analyse et au traitement de l'image et à la reconnaissance des formes avec différentes applications : **Extraction de l'information de documents, reconnaissance de caractères, vérification de signatures manuscrites, compression d'images, tatouage d'images, Analyse de vidéos et suivi de mouvement.** Plusieurs thèses ont été soutenues concernant ces thématiques et d'autres sont en cours.
L'équipe MS2I (ex-équipe SIC dans l'ancienne accréditation 2005 -2014), forte de sa grande productivité scientifique (4 thèses soutenues, 11 thèses en encadrement et un grand nombre de publications et communications) a élargi ses champs de recherche aux thèmes de la modélisation, la simulation, l'interaction et l'intelligence et plus particulièrement aux environnements informatiques d'apprentissage humain, l'ingénierie de la connaissance et du Web sémantique, la modélisation à base d'agents et la simulation des systèmes sociaux écologiques complexes, le traitement automatique du langage naturel & apprentissage automatique, la logistique et le transport et les automates cellulaires pour simuler la dynamique spatio-temporelle.
L'équipe R2IS est une équipe de recherche nouvellement mise en place par le laboratoire IRF-SIC afin de s'ouvrir sur de nouveaux axes de recherches.
R2IS a comme principal axe de recherche la recherche d'information et l'ingénierie des systèmes. Son principal objectif est de promouvoir et de mener des recherches pluridisciplinaires dans le domaine de la recherche d'information et de l'ingénierie des systèmes

Cet axe constitue le domaine d'expertise de l'équipe et traite essentiellement des thèmes suivants :

- Document multimédia électronique et typographie numérique.
- Entrepôt de données.
- Ingénierie des systèmes d'information.
- Fouille des données complexes et de graphes.
- Archivage et Indexation de documents en arabe.
- Système décisionnel, big data et Cloud computing.
- Réseaux sociaux.
- Systèmes temps réel et embarqués.
- Architectures orientées services et systèmes répartis.
- Services (modélisation, conception, QoS, etc.).
- Sécurité des réseaux (fixe et mobile) et des systèmes.

2. Le laboratoire en chiffres

Le laboratoire rassemble 19 chercheurs permanents. Il compte à son actif :

- 17 doctorats soutenus entre 2010 et 2016.
- 32 doctorats en cours.
- 6 Thèses en cotutelle internationale, dont 5 soutenues.
- 10 conférences internationales organisées.
- 81 publications dans des revues internationales entre 2010 et 2016.
- 90 communications dans des conférences nationales et internationales entre 2010 et 2016.
- 13 ouvrages ou périodiques de recherche entre 2010 et 2016.

Neuf projets de recherche ont été financés entre 2010 et 2016.
Des formations doctorales ont été portées par le laboratoire : une UFR DESA (2005- 2007) et une UFR doctorat (2006-2010). Il participe activement dans des masters : Mathématiques et Applications, Mathématiques Appliquées et Sciences de l'Ingénieur. Il participe également à la formation doctorale» Mathématiques, informatique et applications» dans le cadre du Centre d'études doctoral de l'université Ibn Zohr.

3. Manifestations organisées

Une dizaine de manifestations ont été organisées :

- First International Conference on Image and Signal Processing (ICISP'2001) à Agadir ainsi que les éditions suivantes ICISP'2003 (Agadir), ICISP'2008 (Cherbourg), ICISP'2010 (Québec), ICISP'2012 (Agadir)et ICISP'2014 (Cherbourg) et ICISP'2016 (Québec).
- 9ème Conférence Maghrébine sur les Technologies de l'Information, MCSEAI'06, 07-09 décembre 2006 à Agadir.
- 1er colloque régional sur «L'enseignement des mathématiques vu par la pédagogie d'intégration», 8 mai 2011 à Agadir.
- 7ème édition de la Conférence Internationale sur les Environnements Informatiques pour l'Apprentissage Humain EIAH' 2015, 3-5 juin 2015 à Agadir.

4. Ouvrages de recherche et revues

Mammass, D., El Yassa, M. & Nouboud, F. (2011). *Structure pré topologique et Applications*. Editions Universitaires Européennes.
Mammass, D., Nouboud, F. & El Moataz, A. (2012).Special Issue of the International Journal of Future Generation Communication and Networking Vol. 5, No. 4, Sciences and Engineering Research Society.
El Moataz, A., Lezoray, O., Nouboud, F. & Mammass, D. (2008). Lecture Notes in Computer Sciences, Vol. 5099, Springer.
El Moataz, A., Lezoray, O., Nouboud, F., Mammass, D. & Meunier, J. (2010).Lecture Notes in Computer Sciences, Vol. 6134, Springer.
Mammass, D., Nouboud, F. & El Moataz, A. (2010). International Journal on Graphics, Vision and Image Processing (GVIP), Vol. 10, Issue VI D.
El Moataz, A., Mammass, D., Lezoray, O., Nouboud, F., Aboutajdine, D.(2012). Lecture Notes in Computer Sciences, Vol. 7340, Springer.
Essaaidi, M. & Nemiche, M., 2012 IEEE International Conference on Complex Systems - ISBN 978-1-4673-4764-1, IEEE.
Nemiche, M. & Essaaidi, M. (2013). Systems Research and Behavioral Science, Vol. 30, Issue 6, John Wiley & Sons, Ltd
Nemiche, M. & Essaaidi, M. (2013).International Journal of Systems, Control and Communications Vol. 5, Issue 3/4 (Double Special Issue), Inderscience Enterprises Ltd.
Nemiche, M. & Essaaidi, M. (2013). International Journal of Applied Evolutionary Computation Volume 4, Issue 3 (Special Issue), IGI Global. 2014
Essaaidi, M. & Nemiche, M., IEEE Second World Conference on Complex Systems, ISBN 978-1-4799-4647-1, IEEE
Bricage, P. & Nemiche, M. , 2013Acta Europeana Systemica, Vol. 2, Issue 1 (Special Issue).
El Moataz, A., Lezoray, O., Nouboud, F., Mammass, D. (2014). Lecture Notes in Computer Sciences, Vol.8509, Springer.

5. Projets de recherche financés

- Indexation Intelligente de Documents en Langue Arabe (AIDA) Programme Euromed 3+3, INRIA de Nancy, LITIS de Rouen, Labged (Annaba – Algérie), ENIS de Tunisie et Université de Barcelone, 2009-2011
- Télédétection et Systèmes d'Information Géo Référenciés : Application à la Cartographie Thématique de la Région d'Agadir «convention CNRST – CNRS SP I09/10» - CESBIO – CNES – Toulouse – France, 2009-2011.

- Indexation par le contenu et archivage de fonds documentaires arabes, Action intégrée MA/233/10 - INRIA de Nancy, LITIS de Rouen et Labged – Univ. D'Annaba- 2010-2013.
- SEIW : santé et extraction d'images basées sur le watermarking, Action intégrée AI : MA/21/ 279 - Polytech'Orléan et Uinv Hassan 2. 2012-2015.
- Projet PALMERA - Projet Européen (Région Souss Massa Draa et Université Laguna -Iles Canaries), 2012-2013.
- Archivage et Indexation de documents en Arabe, Axe de recherche du pole National STIC. 2011.
- Projet PROTARS II : P41/25 :» Extraction de l'information de documents et vérification de signatures manuscrites» CNRST – 2006-2008.
- Aprendizaje y Gestion Del Conocimiento En La Formacion De Formadores' Action complémentaire B/7269/06 de la coopération Maroco-espagnole - Centro de la Innovación para la Sociedad de la Información (CICEI), Universidad de Las Palmas de Gran Canaria, 2006-2007.
- Indexation et Archivage du Patrimoine Historique du Sud Marocain, Laboratoire de Recherche sur les Sociétés du Souss, des Oasis et du Sahara (LARSSOS) – Faculté des Lettes et Sciences Humaines d'Agadir.

J'ai lu

"Langage C", par Najib Tounsi

"Langage C", by Najib Tounsi

Zineb Kacemi
Kacemi.c@gmail.com

Résumé

"Langage C" est un livre qui s'adresse aux personnes désireuses d'apprendre C, un des langages de programmation les plus utilisés.

Abstract

"Langage C" is a book intended for people who wish to learn C, one of the most used programming languages.

Mots-clés

Langage C, Langage de programmation.

Keywords

C Language, Programming Language.

"Langage C", ouvrage paru en 2016 aux Editions Universitaires Européennes en français, est un cours complet sur le langage C, rédigé par Najib Tounsi, professeur d'informatique à l'Université Mohammed V de Rabat. Bien qu'ancien, ce langage est toujours d'actualité et de ce fait, l'auteur a jugé bon de faire part de sa longue expérience dans l'enseignement.
L'auteur commence par l'introduction de la structure générale et des éléments de base du C, avant d'enchaîner sur la présentation des structures de contrôle, des structures de données et enfin, les fonctions qui permettent de composer un programme.
Le livre est riche en exemples illustratifs, accompagnés à différents endroits de schémas qui aident à la visualisation des concepts. Il propose également des exercices variés de compréhension et d'entrainement. J'ai apprécié en particulier les remarques de l'auteur quand il attire l'attention du lecteur sur les erreurs récurrentes chez les débutants, ou quand il expose les bonnes pratiques à adopter dans le codage. Par ailleurs, Najib Tounsi anticipe sur les questions que le lecteur se pose habituellement, ce qui ne manquera pas d'éclairer ce dernier sur des aspects nécessaires à la compréhension.
Même si cet ouvrage est présenté comme étant destiné aux débutants en programmation, il semble raisonnable de l'adresser aux étudiants ayant déjà quelques bases de la programmation. Il s'adresse également à ceux qui connaissent déjà le langage C mais qui souhaitent le revisiter.
L'ouvrage est donc à recommander, même si l'on peut regretter qu'avec toutes ses qualités, un nombre non négligeable de fautes de langue et de mise en forme du texte aient échappé au contrôle.

Appel aux articles

e-TI, Revue électronique des Technologies de l'Information

Appel aux articles pour la 10e édition

Call for articles for the 10th édition

en accès libre : `www.revue-eti.net`

La Revue électronique desTechnologies de l'Information sollicite aussi bien les universitaires que les industriels pour présenter leurs résultats de recherche, leurs réflexions et les tendances dans les thématiques liées aux Technologies de l'Information (TI). Son but est de favoriser les échanges des connaissances en TI, entre pays du Nord et du Sud. La revue e-TI est en libre accès. Elle est référencée par plusieurs moteurs de recherches tels que DOAJ, EBSCO, IMIST, revue.org, ResearchGate, et Scholar Google.
Vos articles soumis seront évalués en continu pour la 10ème édition, et selon le résultat, ils sont insérés dans le site web de la revue, au fur et à mesure de leur acceptation. L'évaluation est réalisée par les pairs par double évaluation anonyme (voir la déclaration éthique). Le délai d'évaluation est au plus de deux mois. A la clôture de la 10ème édition, l'ensemble des articles parus sera rassemblé en un volume, en version papier ainsi qu'en version électronique, téléchargeable à partir du site de la revue. :

- **Etat de l'art** propose un état de l'art d'un thème dans le domaine des TI ;
- **Recherche** rassemble des articles portant sur la théorie, la conception, la spécification ou l'implémentation d'outils liés aux TI. Les articles traitant d'axes émergents sont particulièrement les bienvenus ;
- **R&D** met en évidence des expériences de recherches et de développement ainsi que leur validation ;
- **Usage et formation** donne un aperçu des recherches concernant la perception des TIs du point de vue de l'usager et la formation au TIs ;
- **Fiche équipe** présente une équipe de recherche afin d'encourager la coopération ;
- **J'ai lu** introduit et critique un ouvrage récent.

La liste suivante suggère (mais ne limite pas) les thèmes de e-TI :

- Systèmes d'information
- Approche à base de composants
- Méthodologie de conception, Ingénierie dirigée par les modèles
- Bases de données, *Big Data*
- Sémantique, ontologie
- Intelligence artificielle, *data mining*, base de connaissances, SIAD
- Technologies Internet et Web, *cloud computing*, *web mining*
- Systèmes répartis, interopérabilité et intégration, SOA
- Mobilité, pervasivité, MDM
- Qualité, sécurité et aspects non fonctionnels
- Utilisabilité, accessibilité, personnalisation, environnement collaboratif
- Applications *e-gov, e-business*, SIG.

Soumission

Les auteurs sont invités à soumettre des articles en **français** ou en **anglais** dans les rubriques citées : 22 pages maximum pour les articles **Etat de l'art**, 20 pages maximum pour les articles de **Recherche** et ceux de la rubrique **Usage et formation**, 10 pages maximum pour les **Expériences R&D**, 2 pages maximum pour les **Fiches des équipes** et la rubrique **J'ai lu**. Ces articles doivent respecter les instructions aux auteurs. Le *template* des articles ainsi que les recommandations sont téléchargeables sur le site de la revue.
La soumission est à effectuer via le site EasyChair eTI10 <`https://easychair.org/conferences/?conf=eti10`> ou par voie électronique à l'adresse `eti@revue-eti.net`.

Date limite de la soumission

31 mars 2017.