N° d'ordre : 142/2017

UNIVERSITÉ SULTAN MOULAY SLIMANE
Faculté des Sciences et Techniques
Béni Mellal

Centre d'Etudes Doctorales : Sciences et Techniques

Formation doctorale : Mathématiques et Physique Appliquées

THÈSE

Présentée par

M. Badr HSSINA

Pour l'obtention du grade de

Docteur

Spécialité : Informatique

Construction d'un Système E-learning Adaptatif basé sur les Technologies du Web Sémantique

---

Soutenue le 16/12/2017 devant la commission d'examen :

Pr. Said MELLIANI	Professeur à la FST-USMS-Béni Mellal	Président
Pr. Mohammed BENATTOU	Professeur à la Faculté des Sciences-Kenitra	Rapporteur
Pr. Cherki DAOUI	Professeur à la FST-USMS-Béni Mellal	Rapporteur
Pr. Rochdi MESSOUSSI	Professeur à la Faculté des Sciences-Kenitra	Rapporteur
Pr. Mohamed FAKIR	Professeur à la FST-USMS-Béni Mellal	Examinateur
Pr. Abdelkrim MERBOUHA	Professeur à la FST-USMS-Béni Mellal	Directeur
Pr. Belaid BOUIKHALENE	Professeur à la FP-USMS-Béni Mellal	Co-directeur de thèse

Remerciements

"Si j'ai vu plus loin que d'autres, c'est parce que j'étais hissé sur des épaules de géants"

Isaac NEWTON

Le travail présenté dans ce document a été effectué au sein du Laboratoire de Traitement de l'Information et Aide à la Décision (TIAD) affilié au Centre d'Etudes Doctorales de la Faculté des Sciences et Techniques (FST) à l'Université Sultan Moulay Slimane de Béni-Mellal.

Je tiens tout d'abord à exprimer ma plus profonde estime et mes vifs remerciements à mon directeur de thèse le professeur **M. Abdelkrim MERBOUHA** pour la confiance qu'il m'a accordée en acceptant d'encadrer ce travail doctoral, pour ses multiples conseils et pour tout le temps qu'il a consacré à diriger cette recherche. J'aimerais également lui dire à quel point j'ai apprécié sa grande disponibilité et son respect sans faille des délais serrés de relecture des documents que je lui ai adressés. Enfin, j'ai été extrêmement sensible à ses qualités humaines d'écoute et de compréhension tout au long de ce travail doctoral.

Mes pensées vont tout particulièrement à mon co-directeur de thèse le professeur **M. Belaid BOUIKHALENE** qui m'a inculqué les principes de la recherche et qu'il m'a patiemment amené à formaliser les idées qui sont au cœur de ce travail. Au cours de nos nombreux entretiens, j'ai apprécié son écoute, sa rigueur et la profondeur de ses connaissances. Aussi, je le remercie sincèrement pour son encadrement considérable, ses idées précieuses et son soutien constant ainsi que son assistance morale qui m'a été d'une utilité inestimable.

Je remercie vivement tous les membres du jury de ma thèse qui ont pris de leur temps pour lire et juger mon travail ainsi que pour leur déplacement le jour de la soutenance.
Je remercie **M. Said MELLIANI** pour l'honneur qu'il m'a fait en acceptant de présider le jury de ma thèse.
Je remercie, également, infiniment les rapporteurs de ma thèse, **M. Cherki DAOUI** professeur à la Faculté des Sciences et Techniques de Béni Mellal, **M. Mohammed BENATTOU** professeur à la Faculté des Sciences de Kenitra et **M. Rochdi MESSOUSSI** professeur à la Faculté des Sciences de Kenitra, pour avoir consacré du temps à la lecture de cette thèse ainsi pour avoir soumis leur précieux jugement sur la qualité et le contenu de ce travail.
Je remercie **M. Mohamed FAKIR** d'avoir accepté d'être examinateur, malgré son emploi du temps très chargé.

Je tiens à exprimer mes sentiments les plus respectueux et ma profonde reconnaissance à mes chers parents, à ma chère femme, mes frères, mes sœurs et à tous les membres de ma famille qui m'ont apporté leur aide nécessaire au cours de mes études et qui m'ont supporté

dans les moments de stress et de difficultés.

Un grand hommage à tous mes collègues et amis du Lycée Qualifiant Atlas de Zaouit Echeikh, du Lycée Qualifiant Oued Eddahab d'Ouled Mbarek, de la Faculté des Sciences et Techniques de Béni-Mellal.

Enfin, je remercie toutes les personnes qui, de près ou de loin, ont apporté leur contribution à ce travail. Je leur exprime ici toute ma reconnaissance et ma sympathie.

Liste des publications

Publications dans des journaux :

1. Hssina, B., Bouikhalene, B., Merbouha, A. (2015). Towards an E-Learning Platform Multi-Agent Based On the E-Tutoring for Collaborative Work. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 9(4), 978-982
2. Hssina, B., Merbouha, A., Bouikhalene, B. (2014). Web services at the service of e-learning platforms. International Journal of Innovation and Applied Studies, 7(4), 1574.
3. Hssina, B., Bouikhalene, B., Merbouha, A. (2016, March). Evaluation of semantic similarity using vector space model based on textual corpus. In 2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV) (pp. 295-300), IEEE.
4. Hssina, B., Bouikhalene, B., Merbouha, A. (2017). An Ontology to Assess the Performances of Learners in an e-Learning Platform Based on Semantic Web Technology : Moodle Case Study. In Europe and MENA Cooperation Advances in Information and Communication Technologies (pp. 103-112). Springer International Publishing.
5. Hssina, B., Merbouha, A., Ezzikouri, H., Erritali, M., Bouikhalene, B. (2013). An Implementation Of Web Content Extraction Using Mining Techniques. Journal of Theoretical and Applied Information Technology, 58(3).
6. HSSINA, B., Merbouha, A., Ezzikouri, H., Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. Int. J. Adv. Comput. Sci. Appl, 4(2).
7. Hssina, B., Lamkhantar, S., Erritali, M., Merbouha, A., Madani, Y. (2017, June). Building of an Information Retrieval System Based on Genetic Algorithms. In International Conference on Mobile, Secure, and Programmable Networking (pp. 195-206). Springer

Conférences internationales :

1. 3ème Symposium International de traitement Automatique de la langue et culture Amazigh (SITACAM'13) à FST, Béni Mellal, Maroc, 2-4 Mai 2013. Participation avec une Communication orale intitulé « An Ontology-Based Intrusion detection for Vehicular Ad Hoc Networks»

2. First International Conference on Business Intelligence (CBI'14), à FST Béni Mellal, Morocco, 29-30 April 2014. Participation avec une Communication orale intitulé « Web services at the service of e-learning platforms»
3. The Second International Conference on Business Intelligence (CBI'15), à FST Béni Mellal. Participation avec une Communication orale intitulé « JADE multi-agent middleware applied to e-learning platforms»
4. First spring conference on applied science and computing. EST ESSAOUIRA du 30 au 31 Mai 2015. Participation avec une Communication orale intitulée « Semantic similarity between documents in collaborative work of learners on an e-learning platform»
5. 13th International Conference on Computer Graphics, Imaging and Visualization is held between 30,31 March - 1 April 2016, Beni Mellal City Morocco, CGIV'2016 . Participation avec une communication orale intitulée «Evaluation of semantic similarity using vector space model based on textual corpus».
6. Participation au concours international du Bigdathon-pédagogique organisé en France par un projet intitulé «Exploitation des données massives extraites des réseaux sociaux pour contribuer à l'apprentissage adaptatif». Ma participation a été au sein d'une équipe intitulée «atlas learning» et notre projet était parmi les 6 projets gagnants. Présenter et défendre l'idée de ce projet à la finale qui a été organisée du 19 au 23 septembre 2016 à l'atelier Canopé du POITIERS en France.
7. Europe Middle East and North Africa Conference on Technology and Security to Support Learning. Oral communication «An Ontology to Assess the Performances of Learners in an e-Learning Platform Based on Semantic Web Technology : Moodle Case Study». This conference is held between 3 - 5 October 2016 -Saidia - Morocco.
8. Conference on Information Technology (ACIT'2016). Oral communication «A hybrid approach of semantic similarity calculation for a Content-based recommendation of text documents on an e-learning platform». This conference is held between 6 - 8 December 2016 -Beni Mellal - Morocco.
9. Conference on Information Technology (ACIT'2016). Oral communication «Predicting learners' Performance in an E-Learning Platform based on Decision Tree Analysis». This conference is held between 6 - 8 December 2016 -Beni Mellal - Morocco.

Résumé

Les avancées rapides des technologies de l'information et de la communication ont des conséquences capitales sur l'évolution des méthodes d'apprentissage. À ce propos, le e-learning appelé aussi l'apprentissage électronique recouvre toutes les méthodes de formation à distance qui s'appuient sur les technologies de l'information et de la communication. L'enjeu pour ce type d'apprentissage est de fournir un accès efficace à la connaissance et un contenu adapté aux attentes des apprenants. La majorité des e-learning d'aujourd'hui manquent de méthodes pour assister le besoin des apprenants qui sont généralement hétérogènes en termes de capacités intellectuelles, rythme d'apprentissage, préférences, etc. Il faut alors fournir des mécanismes puissants pour organiser un tel apprentissage et adapter les décisions pédagogiques aux compétences et aux besoins particuliers de chaque apprenant.

Notre contribution dans ce domaine de recherche porte sur le développement d'une plateforme e-learning adaptatif qui permet de générer des parcours d'apprentissage adaptés au profil de l'apprenant et à l'objectif pédagogique fixé par l'enseignant. Nous avons étudié la problématique de l'adaptation comme un «problème d'optimisation», en utilisant les algorithmes génétiques qui sont fondés sur la théorie de l'évolution. Le but principal de notre approche est de chercher un parcours optimal à partir du profil de l'apprenant jusqu'à l'objectif pédagogique escompté en passant par des générations intermédiaires.

En outre, nous proposons un système de recommandation, considéré comme un sous-ensemble de systèmes e-learning adaptatifs. Ce système de recommandation sémantique permet de retourner des documents susceptibles d'intéresser l'apprenant. Une telle recommandation est basée sur une méthode hybride de calcul de similarité sémantique qui combine entre une ressource linguistique externe (WORDNET) et la représentation vectorielle des documents. Notre objectif final est d'orienter les apprenants et leurs suggérer des ressources à la base de leurs expériences d'apprentissage.

Par ailleurs, la gestion intelligente des connaissances circulant sur une plateforme e-learning est un défi majeur pour les concepteurs. C'est la raison pour laquelle nous avons conçu une solution que nous appellerons MASET (Système Multi Agents pour E-Tutorat des apprenants engagés dans le Travail collaboratif en ligne) qui vise essentiellement à aider les tuteurs à surveiller le travail collaboratif des apprenants à travers leurs diverses interactions. Ce système est fondé sur le middleware JADE (Java Agent Development Framework). Ainsi, nous utilisons le paradigme agent, dans notre système, pour bénéficier des points forts de ce paradigme notamment la modularité, l'autonomie et la flexibilité. Notons aussi que nous avons appliqué des algorithmes qui relèvent du domaine de datamining pour construire

un modèle prédictif basé sur les arbres de décision afin de prédire le niveau des apprenants dans leur parcours d'apprentissage au cours d'une formation en ligne.

Les mots clés : Arbre de décision, E-learning adaptatif, Système multi-agents, Similarité sémantique, web sémantique.

Abstract

Rapid advances in information and communication technologies have had a major impact on the evolution of learning methods. In this sense, e-learning also called online learning covers all distance learning methods that rely on information and communication technologies. The challenge for this type of learning is to provide effective access to knowledge and content that is responsive to learners' expectations. The majority of e-learning today lack methods to assist the need of learners who are generally heterogeneous in terms of intellectual abilities, pace of learning, preferences, etc. It is then necessary to provide powerful mechanisms for organizing such learning and adapting pedagogical decisions to the particular skills and needs of each learner.

Our e-learning research approach focuses on the development of an adaptive e-learning platform that enables the generation of learning paths adapted to the learner's profile and the pedagogical objective set by the teacher. We have studied the problem of adaptation as an «optimization problem», using genetic algorithms that are based on the theory of evolution. The proposed system seeks an optimal path from the learner's profile and the pedagogical objective through intermediate generations. In addition, we propose a recommendation system, considered as a subset of adaptive e-learning systems. This system of semantic recommendation makes it possible to return documents likely to interest the learner. Such a recommendation is based on a hybrid method of semantic similarity computation which combines an external linguistic resource (WORDNET) with the vector representation of the documents. Our ultimate goal is to guide learners and suggest resources for their learning experiences.

Moreover, the intelligent management of knowledge circulating on on an e-learning platform is a major challenge for designers. This is why we have developed a solution called MASET (Multi-Agent System for E-Tutoring of learners engaged in collaborative online work), which aims essentially at helping tutors to monitor the collaborative work of learners through their various interactions . This system is based on Java Agent Development Framework (JADE) middleware. Thus, we use the agent paradigm, in our system, to benefit from the strengths of this paradigm including modularity, autonomy and flexibility. Note also that we applied algorithms that fall within the domain of datamining to build a predictive model based on decision trees in order to predict the level of learners in their learning path during an online training.

Keywords : Adaptive E-learning, Decision tree, Multi-agent system, Semantic similarity, Semantic web.

Table des matières

Table des figures

Liste des tableaux

Abréviations

AAI	: Automatic Analysis of Interactions
ACL	: Agent Communications Language
ADL	: Advanced Distributed Learning
AHS	: Adaptive Hypermedia System
AICC	: Aviation Industry Computer-Based Training Committee
CART	: Classification And Regression Trees
CEN	: Comité Européen de Normalisation
CO	: Content Organization
DAWG	: Data Access Working Group
EAO	: Enseignement Assisté par Ordinateur
EIAH	: Environnements Informatiques pour l'Apprentissage Humain
IA	: Intelligence Artificielle
IDF	: Inverse Document Frequency
IMS-LD	: Instructional Management Systems Learning Design
IEEE	: Institute of Electrical and Electronics Engineers
IETF	: Internet Engineering Task Force
ITS	: Intelligent Tutoring Systems
JADE	: Java Agent DEveloppent framework
LCMS	: Learning Content Management System
LMS	: Learning Management System
LO	: Learning Object
LOM	: Learning Object Metadata
LTSC	: Learning Technology Standards Committee
MAS	: Multi-Agent System
MASET	: Multi Agents System for E-Tutoring Learners engaged in online collaborative work
Moodle	: Modular Object-Oriented Dynamic Learning Environment
ODE	: Ontology Design Environment
OWL	: Ontology Web Language
OWL-DL	: Ontology Web Language Description Logics
QTI	: Question Test Interoperability
RDF	: Resource Description Framework

RF	: Random Forests
RIF	: Rule Interchange Format
RS	: Recommendation System
SCO	: Shareable Content Object
SCORM	: Sharable Content Object Reference Model
SPARQL	: Protocol And Rdf Query Language
SWRL	: Semantic Web Rule Langage
TALN	: Traitements Automatiques de la Langue Naturelle
TF	: Term Frequency
TIC	: Technologies d'Information et de Communication
URI	: Uniform Ressource Identifier
URL	: Uniform Resource Locator
XML	: eXtensible Markup Langage

Introduction générale

1. Contexte et motivation

Ce travail présente la synthèse de quatre années d'investigation effectuée dans le cadre d'une thèse de doctorat en Informatique au sein du laboratoire TIADE (Traitement de l'Information et Aide à la Décision) de l'Université Sultan Moulay Slimane, Faculté des Sciences et Techniques, Béni Méllal. L'utilisation de e-learning qui se développe sans cesse dans de nombreuses universités et entreprises était parmi les premières motivations qui m'ont attiré vers ce sujet de recherche. De plus, le e-learning apparaît non seulement comme une nouvelle forme d'enseignement offrant de nombreux avantages mais également comme une solution impérative pour garantir un apprentissage de qualité.

En outre, le e-learning garantie une formation sans se déplacer. Ceci permet d'économiser et le temps et les dépenses ce qui favorise des conditions optimales de formation. Un formateur peut s'adresser à un grand nombre d'apprenants tout en assurant une relation individualisée avec chacun d'eux. Le temps d'apprentissage personnel est réduit. En se consacrant uniquement aux points qu'il souhaite approfondir, l'apprenant se forme plus rapidement. Le coût de son indisponibilité est donc fortement diminué.

Pour toutes ces raisons, notre recherche est orientée vers l'intégration des technologies du web sémantique dans e-learning en vue de l'adaptabilité d'objets pédagogiques aux profils des apprenants. L'objectif principal est la proposition d'une nouvelle architecture pour la conception d'une plateforme d'apprentissage adaptatif. Les travaux effectués ainsi que les résultats obtenus se résument comme suit :

- Générer un parcours d'apprentissage adapté au profil de chaque apprenant selon des objectifs pédagogiques fixés par l'enseignant en utilisant des algorithmes génétiques.
- Soutenir l'apprentissage personnalisé des apprenants hétérogènes, en proposant une contribution à un environnement éducatif qui permet de calculer la similarité sémantique entre des documents textes. Le but de notre contribution étant de recommander aux apprenants des documents pédagogiques similaires à leurs choix antérieurs.
- Fournir aux tuteurs une surveillance efficace des activités des apprenants lors d'un travail collaboratif en ligne grâce à un système multi-agents.
- Prédire la performance des apprenants dans leur parcours d'apprentissage, dans une formation, à travers un modèle prédictif basé sur les arbres de décision.

L'enjeu de cet effort de création est d'augmenter l'efficacité de l'apprentissage en ligne.

2. Problématique

Les systèmes e-learning existants proposent des parcours d'apprentissage fixés avec une adaptation d'affichage des pages selon les préférences des utilisateurs. D'autres systèmes utilisent des méthodes de création de cours, pour le structurer selon un format donné, pour pouvoir gérer les informations à afficher selon le profil de l'apprenant.

Aujourd'hui, les recherches relatives aux systèmes pédagogiques adaptatifs sont structurées généralement en trois groupes. Tout d'abord, on distingue celles qui sont liées aux *théories sur les méthodes d'apprentissage* qui suscitent un grand intérêt pour des pédagogies dites "actives" associées aux socioconstructivisme [1]. Cet intérêt, qui se manifeste à tous les niveaux scolaires, est en règle générale alimenté par une volonté de rendre les savoir-faire plus opérationnels et les actes d'apprentissage plus motivants.

Ensuite, les recherches relatives aux *systèmes hypermédias adaptatifs* [2] : Dans ce type de systèmes, les connaissances sur l'apprenant sont essentielles. On s'intéresse à son niveau de connaissance et d'expertise dans un domaine donné ainsi que ses préférences quant aux approches ou méthodes pédagogiques. Finalement, les recherches liées au *web sémantique adaptatif* [3] qui visent à enrichir les ressources disponibles sur le web avec des descriptions sémantiques de leur contenu. En plus, l'apprenant doit être guidé lorsqu'il consulte ses cours, ce qui va lui permettre de mieux comprendre ses démarches d'apprentissage et ainsi de pouvoir s'auto-évaluer. Le e-learning adaptatif va donc pouvoir aider l'apprenant à être plus autonome, à avoir une meilleure compréhension d'un cours et aussi de mieux gérer sa façon d'apprendre.

Comment pouvons-nous réponde à ce besoin en proposant à un apprenant un parcours d'apprentissage adapté à son profil et en lui suggérant des ressources pédagogiques qui peuvent l'intéresser ?
Le travail que nous présentons dans cette thèse contribue à répondre particulièrement à cette question, puisqu'il s'intègre dans une problématique générale des e-learning adaptatifs. De plus, ajouter une partie adaptative à une plate-forme e-learning nécessite une réflexion approfondie. Cependant, les moyens informatiques actuels permettent de définir un enseignement qui s'adapte aux résultats, aux comportements et aux goûts des apprenants. Ils les aident à prendre conscience des processus par lesquels ils apprennent.

3. Organisation de la thèse

Cette thèse traite les plateformes e-learning adaptatifs s'appuyant sur les technologies du web sémantique. Elle se compose de cinq chapitres, une introduction et une conclusion. **L'introduction** présente le contexte et les motivations de recherche du sujet de e-learning adaptatif ainsi que les objectifs et les contributions de cette investigation.

Le premier chapitre est réservé à l'état de l'art de la thématique de l'apprentissage en ligne. Nous commençons par l'historique de l'enseignement assisté par ordinateur puis

les systèmes tutoriels intelligents et les caractéristiques des environnements informatiques pour l'apprentissage humain. Nous décrivons dans la suite de cette section les avantages, les enjeux de e-learning, les normes et standards autour des objets pédagogiques. La dernière partie de ce chapitre explore le e-learning adaptatif en analysant minutieusement les composants, les techniques et les méthodes d'adaptation.

Dans **le chapitre 2**, nous allons tout d'abord commencer par définir le concept du web sémantique et spécifier les différentes couches de son architecture, puis nous allons présenter les apports de l'intégration des ontologies dans les plateforme e-learning.

Les développements réalisés dans **le chapitre 3** se composent de trois éléments principaux. Premièrement, nous exposons les systèmes de recommandation en particulier la recommandation basée sur le contenu. Deuxièmement, nous introduisons la notion de l'indexation sémantique et le calcul de similarité sémantique basé sur un corpus textuel. Troisièmement, nous terminons par les critères d'évaluation des systèmes de recommandation sémantique dans le cadre d'une plateforme e-learning.

Le chapitre 4 se focalise sur nos contributions réalisées dans le cadre de cette thèse. Premièrement, la génération d'un parcours d'apprentissage adapté au profil de l'apprenant. Deuxièmement, la proposition d'un système de recommandation sémantique des documents texte aux apprenants qui se base sur une nouvelle approche hybride de calcul de similarité sémantique.

Au cours du **chapitre 5** nous présentons notre système multi-agents pour e-tutorat des apprenants engagés dans le travail collaboratif en ligne. Puis, nous introduisons notre approche de la prédiction de la performance des apprenants en se basant sur les arbres de décision.

La thèse se termine par une **conclusion** générale dans laquelle nous présentons un bilan de nos travaux de recherches et nous traçons des perspectives qui nous permettraient d'améliorer ce qui a été proposé.

Chapitre 1

Etat de l'art sur les systèmes e-learning

"Dis-le moi et je l'oublierai ; Enseigne-le moi et je m'en souviendrai ; Implique-moi et j'apprendrai"

- Benjamin Franklin, *Écrivain américain du XVIIIe siècle*

Introduction

Le développement rapide des nouvelles technologies de l'information et de la communication a rendu la formation en ligne une forme d'apprentissage privilégiée et une pratique très courante. En effet, le e-learning est un processus d'apprentissage à distance qui s'appuie sur des ressources multimédias, permettant à une ou à plusieurs personnes de se former à partir de leur ordinateur.

Dans ce chapitre nous allons présenter en premier lieu, un bref historique des systèmes e-learning et les standards utilisés pour représenter les objets pédagogiques. Puis, nous allons citer les différents acteurs et les principales plateformes d'apprentissage en ligne existantes. Ensuite, nous allons introduire la notion du tutorat en ligne qui fait partie des problématiques étudiées dans cette thèse.

Nous terminons ce chapitre, en introduisons une nouvelle forme d'apprentissage en ligne qui connaît un grand succès ces dernières années, c'est le e-learning adaptatif dont l'objectif principal est la personnalisation des ressources d'apprentissages pour chaque apprenant. Plus précisément, grâce à cette forme d'apprentissage, un apprenant pourra se former en fonction de son rythme d'apprentissage, de ses besoins, de ces compétences et de ses disponibilités.

1. Historique de l'Enseignement Assisté par Ordinateur (EAO)

L'idée des systèmes d'apprentissage automatisés remonte au début des années soixante, avec l'avènement de l'intelligence artificielle, qui a donné lieu aux systèmes d'EAO (En-

seignement Assisté par Ordinateur) [4]. Ces derniers provoquent une forte évolution de l'utilisation de l'ordinateur dans l'enseignement. Cette évolution, illustrée notamment par l'apparition successive de sigles différents (EIAO [1] puis EIAH [2]) prend appui sur des avancées successives d'ordre théorique (progrès de l'intelligence artificielle) ou encore technologique (mise au point de nouveaux dispositifs d'interaction homme-machine, explosion des possibilités d'accès à l'information et diversification des moyens de communications). À ce propos, Tchounikine définit l'EIAH comme : «Un environnement qui intègre des agents humains (apprenant ou enseignant) et artificiels et leur offre des conditions d'interactions, localement ou à travers les réseaux informatiques, ainsi que des conditions d'accès à des ressources formatives locales ou distribuées» [5].
Ce que nous pouvons retenir de cette définition c'est que les EIAH sont des systèmes informatiques destinés à favoriser l'apprentissage chez les apprenants à l'aide d'un ensemble de ressources, tout en étant guidés à des degrés divers par des acteurs (ou agents).

Par ailleurs, ce n'est que dans les années quatre-vingts qu'a apparu les STI (Systèmes Tutoriels Intelligents) [6], qui sont des systèmes d'enseignement qui possèdent un contenu sous forme de base de connaissance (qui spécifie ce qui doit être enseigné), des stratégies d'enseignement (qui spécifient la manière d'enseigner ce contenu) ainsi qu'une connaissance sur le niveau de l'apprenant, afin de lui fournir dynamiquement un contenu pédagogique.

Nous présenterons dans la suite une description de l'architecture des STI et ses différentes composantes.

2. Les Systèmes Tutoriels Intelligents (STI)

Les systèmes tutoriels intelligents sont des environnements d'apprentissage informatisés issus de l'EAO qui visent à personnaliser la formation en ligne. Ainsi, ils ont été développés pour répondre aux limites de l'EAO en ayant recours à l'intelligence artificielle pour mettre en place des systèmes plus souples et interactifs qui s'adaptent aux besoins spécifiques de l'apprenant en évaluant et en diagnostiquant ses problèmes afin de lui fournir l'aide nécessaire. En effet, tout comme un tuteur humain, les systèmes de ce type ont le potentiel d'amener l'apprenant à réaliser une activité d'apprentissage de la meilleure façon possible. Les STI placent l'apprenant au centre du processus d'apprentissage. C'est dans ce but que parfois ils exposent immédiatement le contenu du domaine à l'apprenant, et dans d'autres cas ils présentent directement les exercices qui permettront d'assimiler de nouvelles connaissances [7] [8].

1. Enseignement Intelligemment Assisté par l'Ordinateur
2. Environnement Informatique pour l'Apprentissage Humain

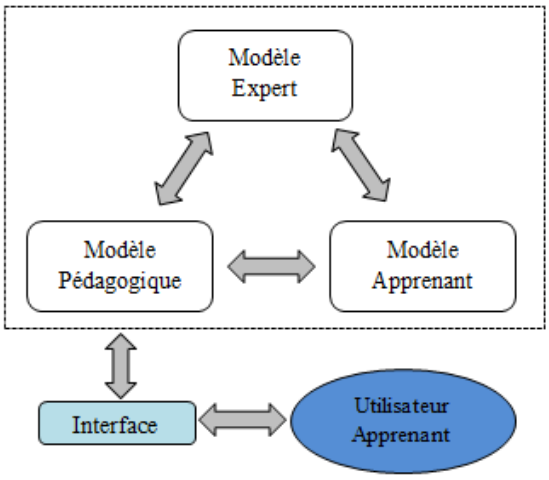

FIGURE 1.1 – Architecture classique d'un système tutoriel intelligent

De manière générale, l'architecture conceptuelle d'un STI se compose de quatre modèles principaux [9] : Le modèle expert, le modèle pédagogique, le modèle apprenant et le modèle interface (figure 1.1) :

Le modèle expert : Aussi appelé modèle du domaine. Il représente l'expertise de l'enseignant dans le domaine, c'est à dire toute la connaissance nécessaire au processus d'enseignement, nous parlons de connaissance reliée à l'expertise du domaine. En général, un modèle expert doit aussi posséder un savoir-faire, c'est-à-dire une expertise sur la manière de résoudre les problèmes du domaine.

- Le modèle pédagogique : Ce modèle met en œuvre des stratégies pédagogiques pour enseigner la connaissance d'un domaine donné. Toute stratégie pédagogique doit être basée sur des principes pédagogiques et psychologiques formels. Les besoins de l'apprenant doivent être identifiés, et pris en compte dans la stratégie d'enseignement. Globalement, ce modèle sert à trois fonctions principales [10] : il contrôle la présentation du contenu enseigné à l'apprenant, il doit être capable de répondre aux questions de l'apprenant et pouvoir déterminer quand les apprenants ont besoin d'aide et quel type d'aide leur fournir.

- Le modèle apprenant : Ce modèle permet d'identifier, pour un apprenant, son niveau de compréhension du domaine de connaissance. Selon McCalla et Greer [11], l'implantation du modèle de l'apprenant est essentielle à l'adaptation du système d'apprentissage aux besoins des apprenants. VanLehn [12] décrit les différences et les similarités entre le modèle expert et le modèle apprenant en termes de conceptions erronées ou de conceptions manquantes. Les conceptions manquantes peuvent être décrites comme des conceptions possédées par l'expert mais pas par l'apprenant, tandis que les conceptions erronées sont des connaissances (fausses) possédées par l'apprenant, mais pas par l'expert.

- Le modèle interface : Il représente la couche de communication (les interactions) entre l'apprenant et le système. De façon générale, l'interface d'un STI doit être conviviale et modélisée de manière à ne pas poser des problèmes de compréhension supplémentaires à

l'apprenant. L'interaction humain-machine dans les STI est particulièrement complexe car les usagers de ces systèmes travaillent avec des concepts qu'ils ne maîtrisent pas bien. Dans ce contexte, une interface incorrectement modélisée peut remettre en cause tout le processus d'enseignement. Selon Miller [13], il est important de tenir compte de deux aspects fondamentaux pour créer une interface de STI. En premier lieu, l'interface doit permettre un enseignement clair et direct. En second lieu, elle doit faciliter l'interaction de l'apprenant avec le contenu enseigné.

En résumé, l'architecture classique d'un STI comporte un modèle expert qui décrit quoi enseigner, un modèle pédagogique qui décrit la stratégie de l'enseignement, un modèle interface qui assure l'interrelation entre l'apprenant et le système et finalement un modèle de l'apprenant qui peut adapter l'apprentissage en tenant compte de celui-ci. Un tel comportement d'adaptation à l'apprenant par les STI est possible puisqu'ils possèdent des composantes «intelligentes», soit une base de connaissances et un moteur d'inférence qui exploite les connaissances de la base. Ces techniques venant de l'intelligence artificielle font donc en sorte que l'environnement est capable d'imiter le tuteur en «raisonnant» à l'aide des connaissances incluses dans sa base de données.

Nous abordons dans ce qui suit, le concept de e-learning d'une manière plus détaillée pour revenir plus tard à la notion du tutorat en ligne.

3. E-Learning

La formation en ligne, l'apprentissage en ligne, l'e-formation ou encore le e-learning, désignent l'ensemble des solutions permettant l'apprentissage par des moyens électroniques. Le « e » dans e-learning est une référence explicite aux technologies de l'information et de la communication.
Mais quel est exactement e-learning ?

Il est difficile de trouver une définition communément acceptée. Dublin en 2003 affirme que l'un des mythes au sujet de e-learning est que «tout le monde sait ce que vous voulez dire quand vous parlez de e-learning : cependant, le terme e-learning signifie différentes choses pour différentes personnes» [14].

Est-il de cours en ligne pour les étudiants à distance ? Est-il l'utilisation d'un environnement d'apprentissage virtuel pour soutenir la prestation de l'éducation sur un campus ? Est-il un outil en ligne pour enrichir, améliorer et étendre la collaboration ? Est-il totalement l'apprentissage ou partie de l'apprentissage mixte en ligne ?
Ci-dessous un bref aperçu des différentes définitions de e-learning qui répond à ces questions :

3.1. Définition

D'après la commission européenne (2001), le e-learning est «l'utilisation des nouvelles technologies multimédias de l'internet pour améliorer la qualité de l'apprentissage en facilitant d'une part l'accès à des ressources et à des services, d'autre part les échanges et la collaboration à distance». En d'autres définitions, Gallagher (2001) décrit e-learning comme «l'utilisation des technologies numériques afin de soutenir et de livrer une partie ou la totalité de l'enseignement et d'apprentissage d'une unité d'étude particulier» [15].

En effet, le e-learning est une modalité pédagogique et technologique qui a d'abord concerné la formation continue, l'enseignement supérieur puis la formation en entreprise, c'est-à-dire au service d'un apprenant mûr ayant une certaine autonomie dans l'organisation de son processus d'apprentissage. Cependant, aux pays développés le e-learning est offert de la maternelle à la formation continue, incluant les didacticiels, hypermédias, tuteur intelligent [16], etc. Dans ces pays, des mesures de réduction des coûts d'accès à l'internet pour les centres d'éducation et de formation ont contribué au succès de e-learning.

Finalement, le terme e-learning évoque donc une alliance nouvelle entre des pratiques pédagogiques et des technologies de communication. Il semble cependant, comme pour les évolutions récentes des organisations, que le e-learning possède maintenant des caractéristiques émergentes (organisation et gestion de la connaissance, approches collaboratives) qui le rendent différent des outils pédagogiques antérieurs.

3.2. Les avantages de e-learning

Comme on pouvait s'y attendre, e-learning présente des avantages par rapport à l'apprentissage traditionnel en classe. Les exemples les plus significatifs sont l'accessibilité, la flexibilité et la performance [17]. En effet, Christian Depover présente dans son livre quelques avantages généraux de e-learning [18] :

- ◇ E-learning propose une large gamme d'outils pour permettre aux enseignants et aux apprenants d'être innovants, créatifs et ingénieux dans toutes les activités d'apprentissage. Les enseignants et les apprenants peuvent facilement personnaliser les ressources d'apprentissage numériques à leur style d'apprentissage.
- ◇ E-learning offre une formation dans un temps assez court par apport à la formation classique en accélérant le délai d'apprentissage. Cette formation donne la possibilité de déterminer le rythme et le temps d'apprentissage ainsi que les cours à étudier.
- ◇ E-learning crée des communautés de pratique en ligne. Ainsi, l'internet peut motiver les apprenants, les enseignants, les communautés de spécialistes et des experts à partager des idées, des informations et de bonnes pratiques.
- ◇ E-learning peut assurer une participation efficace et un accès plus équitable à un enseignement supérieur, en fonction des besoins des apprenants.
- ◇ E-learning offre un environnement d'apprentissage personnalisé grâce à des consignes et des services d'orientation. Il peut aider les apprenants à trouver le cours en fonction de leurs besoins et préférences.

- ◇ E-learning fournit des mondes virtuels d'apprentissage où les apprenants peuvent apprendre à travers des simulations, des jeux, du contrôle à distance d'outils et de dispositifs du monde réel.
- ◇ E-learning n'est jamais en grève, et plus vous l'utilisez, plus son coût relativement devient bas. Dans une formation à distance nous consommons moins de moyens par rapport à la formation traditionnelle. De plus, e-learning élimine certains coûts (transport, papier, location de salle, etc).

Néanmoins, même si le e-learning détient plusieurs avantages, il n'en demeure pas moins que ce système possède aussi ses propres limites.

3.3. Les limites de e-learning

Les limites de e-learning peuvent être vues comme des problématiques à étudier dans les laboratoires de recherche pour l'élaboration d'une offre de formation de qualité :

L'autodiscipline : Les apprenants doivent faire preuve de rigueur et de discipline, particulièrement s'ils sont isolés dans une formation à distance. Les caractéristiques des personnes diffèrent d'une personne à une autre, l'apprenant qui est moins discipliné ou moins organisé peut être incapable de suivre une formation. Il n'est pas toujours facile de s'octroyer du temps pour étudier ou s'auto-former sur son lieu professionnel ou encore chez soi.

L'isolement : En effet, la formation en ligne implique une notion de distance et c'est là qu'un premier signe de faiblesse peut apparaître. Il n'est pas toujours évident de se former par l'intermédiaire d'un logiciel à la place d'un véritable formateur. L'ordinateur est certes un formidable outil qui permet de véhiculer un nombre incalculable d'informations et d'apprentissages mais il demeure un objet froid, impersonnel avec lequel il est difficile de converser et d'échanger [19].

Maîtrise des outils : Le e-learning nécessite une maîtrise suffisante des outils informatiques et d'internet afin de suivre une formation. Pour la majorité des étudiants, il est difficile d'être familiers avec ce nouveau style d'apprentissage à distance sans encadrement. C'est pourquoi, dans chaque cours nous devons ajouter des guides pour aider l'apprenant à suivre son cours.

3.4. Les enjeux de e-learning

L'amélioration des systèmes e-learning est un enjeu majeur des chercheurs dans le domaine de l'enseignement à distance. Les concepteurs de ces systèmes sont focalisés dans leurs recherches, particulièrement, sur l'apprenant, l'enseignant et le contenu pédagogique [17] [20]. Nous pouvons résumer leurs enjeux en trois catégories :

- Mettre l'apprenant au centre des intérêts du système : L'apprenant doit avoir un certain degré d'autonomie pour organiser des tâches, à les mener en bonnes conditions, à les évaluer, à gérer des ressources, à travailler avec des pairs, à maîtriser les techniques et

les outils d'apprentissage, etc. Pour faire face au problème de l'isolement de l'apprenant, un tel système éducatif doit avoir des méthodes d'accompagnement humain de l'apprenant.

- Attribuer des rôles nouveaux aux enseignants : Les systèmes e-learning exigent la modification du rôle de l'enseignant. Alors, le nouvel enseignant devra donc devenir un modérateur capable d'aider les apprenants dans leurs nouveaux environnements de formation. Pour cette raison, les enseignants doivent acquérir de nouvelles compétences, qui ne sont pas uniquement techniques et technologiques, mais encore pédagogiques aussi bien qu'éthiques. Les enseignants ne sont plus appelés désormais à travailler isolément mais au contraire à partager leurs expériences et à prendre conscience des potentialités qu'offre la diversité [21].

- Rendre le système e-learning plus efficace et adaptable aux processus d'apprentissage : Une plateforme e-learning doit fournir un accès à la connaissance plus efficace, plus solide et plus adapté aux attentes des apprenants. L'un des défis posé par les nouveaux styles d'apprentissage est l'adaptabilité. Le e-learning d'aujourd'hui nécessite des méthodes pour assister au besoin des apprenants. Les apprenants qui suivent un processus de formation sont, généralement, hétérogènes au point de vue intelligences, capacités, background, personnalités, préférences, etc. Donc, il faut fournir des mécanismes puissants pour organiser une telle formation. Aussi, l'apprentissage doit être un service en ligne adaptable, initié par les profils des apprenants.

Pour atteindre ces enjeux, les développeurs des systèmes e-learning proposent un ensemble de normes et standards, dans un souci d'équité d'accès, de partage et de mutualisation des acquis éducatifs.

4. Normes et standards autour de e-learning

L'intérêt de e-learning ne se limite pas à proposer un contenu pédagogique aux apprenants mais aussi de faciliter la mise en place de ce contenu aux différents acteurs. En effet, l'idée de réexploiter les ressources pédagogiques en ligne, a donnée lieu à des normes et des standards, afin de faciliter leurs usages, leurs réutilisations et leurs interopérabilités. C'est un autre enjeu de e-learning car seuls quelques organismes comme ISO [3], UIT [4] ou encore le CEN [5], sont accrédités à développer les normes. Avant d'introduire les principaux standards et normes, une clarification de vocabulaires est nécessaire pour enlever la confusion entre ces deux concepts :

- Une norme est un ensemble de règles de conformité, dicté par un organisme de normalisation au niveau national ou international.
- Un standard est un ensemble de recommandations émanants d'un groupe restreint d'utilisateurs réunis autour d'un consortiums et non par des organismes internationaux.

3. International Organization for Standardization
4. Union Internationale des Télécommunications
5. Comité Européen de Normalisation

La différence est cependant faible et les anglo-saxons utilisent le terme standard pour désigner une norme.

En somme, les normes visent à faciliter le transfert ou «l'interopérabilité des contenus de formation entre le système de gestion des apprentissages, c'est-à-dire de rendre possibles l'utilisation et la réutilisation des contenus entre plateformes, en faisant disparaître les contraintes de conversion». Pour ce faire, il est nécessaire de définir un modèle d'objet d'apprentissage dans lequel on prévoit la syntaxe et la sémantique des métadonnées pour la description complète d'un objet pédagogique [22].

Cette standardisation des «objets pédagogique» permet de favoriser les échanges et la mutualisation des données entre machines.

4.1. Objets Pédagogique

Actuellement, le terme d'objet pédagogique (Learning Object) est devenu central dans les environnements e-learning et fait l'objet de nombreux travaux au sein des organismes internationales de normalisation [22]. Plusieurs organismes ont proposé de structurer un document pédagogique autour d'un ensemble d'objets pédagogiques. Généralement la description du contenu de ces objets pédagogiques est basée sur des métadonnées. Plusieurs modèles ont été mis au point au cours des dernières années. Ces modèles ont généralement des buts distincts, mais s'accordent sur l'idée d'aboutir à des composants pédagogiques réutilisables [23].

4.1.1. Définition

Le groupe de travail LTSC [6] définit un objet pédagogique comme «toute entité numérique ou non, qui peut être utilisée pour l'enseignement ou l'apprentissage», cette définition est donc trop générale, non utilisable en pratique et montre la difficulté à définir clairement ce qu'est un objet pédagogique. D'autres définitions plus restrictives sont proposées dans la littérature qui décrivent un objet pédagogique comme «des petites unités d'apprentissage qui sont suffisamment petits pour être intégrés à une activité pédagogique, une leçon, un module ou un cours».

Donc, cette diversité des définitions peut s'expliquer par les tensions présentes autour du concept d'objet pédagogique [24]. À ce sens, la figure 1.2 illustre les objectifs d'un objet pédagogique.

6. Learning Technology Standards Committee

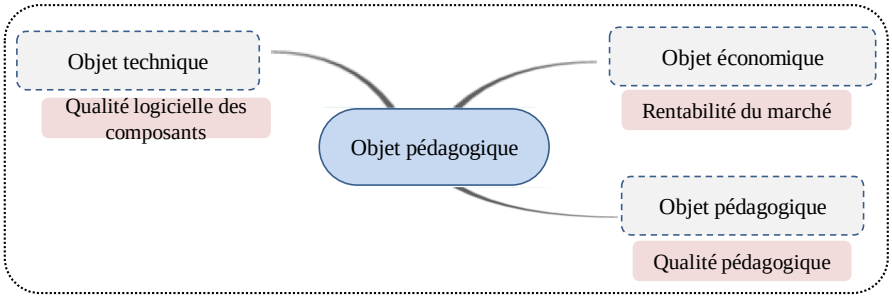

FIGURE 1.2 – L'objet pédagogique, un concept au centre de plusieurs objectifs

Au niveau économique, il s'agit de diminuer les coûts de production sans pour autant diminuer la qualité. Cela implique d'avoir des composants réutilisables et partageables. Au niveau pédagogique, le but est la mise en place des unités d'apprentissage qui définissent les modalités précises d'acquisition, de validation ou de communication d'une ou plusieurs connaissances. Enfin au niveau technique, l'intérêt de l'approche par objet, largement mise en avant dans le développement informatique, n'est plus à démontrer ; il rend possible la réutilisation de composants dans de multiples contextes.

Afin d'approfondir le concept d'objet pédagogique, nous présentons dans la partie suivante ses principales caractéristiques, et une description de son cycle de vie.

4.1.2. Caractéristiques

Chaque objet pédagogique est menu d'un ensemble de caractéristiques qui décrient sa qualité technique, pédagogique et descriptive [25]. Nous recensons dans cette partie les caractéristiques principales d'un objet pédagogique :

- La granularité : Les objets pédagogiques les plus petits, ou les granules, représentent des éléments bruts tels qu'une simple phrase, un paragraphe explicatif, une figure, une animation, etc. Selon (Hodgins, 2002), nous pouvons définir cinq niveaux de granularité différents [26] :

- ◇ L'élément brut de données, au niveau de granularité le plus bas, correspond à des éléments de contenu situés purement au niveau des données ;
- ◇ L'objet d'information se focalise sur une information simple. Il peut servir à expliquer un concept, illustrer un principe ou décrire une procédure ;
- ◇ L'objet d'application est un ensemble d'objets d'information qui se focalisent sur un objectif pédagogique unique ;
- ◇ L'assemblage s'étend à des objectifs pédagogiques plus larges, il correspond aux leçons ou aux chapitres ;
- ◇ La collection, au niveau de granularité le plus élevé, correspond à des cours ou même à des cursus.

- La réutilisabilité : L'objectif de la réutilisabilité est d'avoir des objets pédagogiques élémentaires pouvant être utilisés dans des contextes et dans des buts multiples [27]. Ces

composants doivent par conséquent être autonomes. Ils peuvent être produits séparément, mais doivent pouvoir être modifiés pour correspondre aux besoins des utilisateurs. Par exemple, un auteur qui conçoit un objet pédagogique qui explique le fonctionnement d'un moteur doit idéalement éviter de faire référence à d'autres objets pédagogiques, car il pourrait être utilisé séparément dans un autre contenu et dans un autre contexte.

- L'agrégation : Un objet pédagogique peut certes être réutilisé, mais il peut aussi être créé par agrégation d'autres objets pédagogiques de granularités plus fines. Il répond ainsi au besoin d'appropriation de la part de l'enseignant, et de mise en contexte pour répondre aux besoins spécifiques du public cible d'apprenants [27].

- L'accessibilité : Il est indispensable de pouvoir retrouver facilement un objet pédagogique ; il doit être étiqueté avec des métadonnées pour être stocké et référencé dans une base de données. Ce processus est appelé « indexation ». L'accès à un objet pédagogique est efficace lorsque le coût engendré par sa recherche en vue de sa réutilisation est inférieur au coût nécessaire pour créer un objet pédagogique équivalent. L'objet pédagogique doit être diffusé le plus largement possible, ce qui rend nécessaire la possibilité d'échanger et de communiquer entre les systèmes de stockage. La qualité et la quantité des métadonnées jouent également un rôle important [27].

- L'interopérabilité : Les contenus pédagogiques sont conçus et développés par des organisations et des formateurs différents, constituant généralement des sources de données sémantiquement hétérogènes. De ce fait, l'interopérabilité entre ces contenus est complexe puisque chaque système peut posséder son propre modèle de description et d'encodage des objets pédagogiques et son propre langage d'interrogation. La recherche et la combinaison des résultats deviennent alors un travail long et fastidieux. C'est pourquoi, l'adoption de normes de description permettant de définir de façon modulaire des solutions d'apprentissage souples et adaptables nécessitera le respect d'un ensemble précis de règles permettant l'interopérabilité des différents composants [25].

4.2. Cycle de vie d'un objet pédagogique

Le cycle de vie d'un objet pédagogique est une succession d'évènements liés à son évolution au cours du temps. Il s'agit surtout de prendre en compte de l'évolution de son contenu et de sa description par les métadonnées qui jouent un rôle essentiel pour faciliter sa recherche, son expertise, son apprentissage et son utilisation par les acteurs de e-learning [24]. La figure 1.3 illustre les principales étapes qui constituent le cycle de vie d'un objet pédagogique :

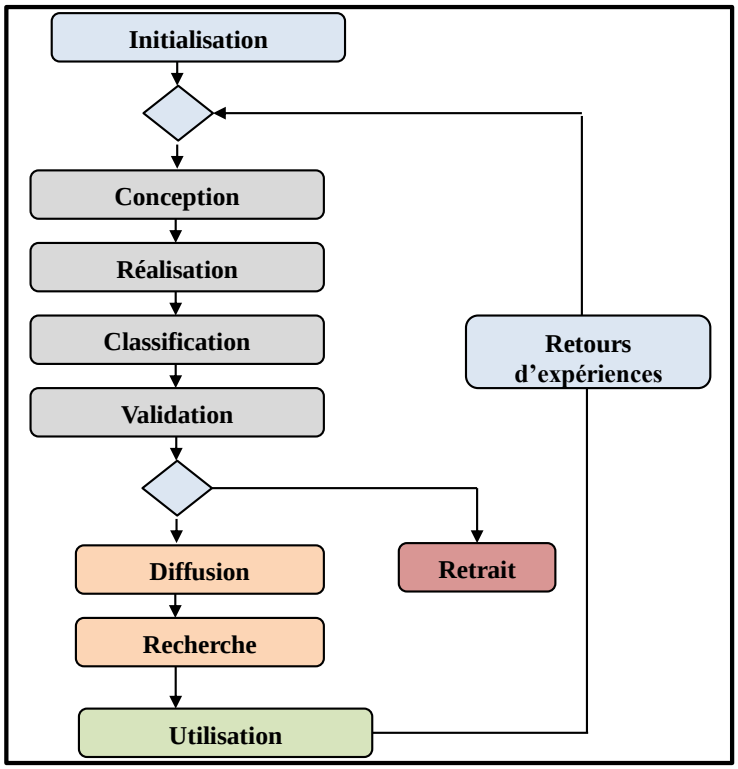

FIGURE 1.3 – Cycle de vie d'un object pédagogique

- La phase de l'initialisation : Au cours de l'étape d'initialisation, l'objet pédagogique ne contient qu'une description des intentions du comité de la formation pour lancer la création d'un nouvel objet pédagogique pour répondre à une nouvelle demande de formation. Il s'agit de définir : Quels sont les objectifs de l'objet pédagogique ? Quelle est la discipline concernée ? Quels sont les prérequis nécessaires ? À quel type de public s'adresse-t-il ?
- La phase de la conception : L'objet pédagogique entre ensuite dans l'étape de conception. Il s'agit d'une étude permettant de répondre aux objectifs fixés en phase d'initialisation. Les experts du domaine, les ingénieurs pédagogiques, les scénaristes et les enseignants visent à définir les caractéristiques spécifiques de l'objet pédagogique : Que va-t-il contenir ? Comment vont s'enchaîner ses différentes parties ? Quel type d'interactivité faut-il mettre en place ?
- La phase de la réalisation : Puis, l'objet pédagogique passe dans l'étape de réalisation. Il s'agit de le concrétiser en le rendant exploitable : Comment appliquer la mise en forme ? Lorsque l'objet pédagogique est réalisé sous forme de produit multimédia, les créateurs graphiques, les spécialistes de l'image, du son, de la vidéo, et du multimédia, les ergonomes interviennent dans la réalisation : comment créer les illustrations et les animations graphiques nécessaires ? Quel type de support est le mieux adapté ? Comment créer les bandes son et animations vidéo quand elles sont nécessaires ? Comment rendre l'utilisation de l'objet pédagogique la plus attirante possible ?
- La phase de la classification : L'objet pédagogique rejoint plus tard l'étape de classification. Il ne subit pas de modification mais cette phase importante permet à un documentaliste, à un archiviste multimédia, de le classer par rapport à des systèmes de classification. Cela permet de situer l'objet pédagogique par rapport aux autres objets et offre un moyen

pour le retrouver facilement dans un domaine particulier. Cela facilite également la diffusion de l'objet pédagogique à l'extérieur de l'organisation responsable de sa production. La classification n'est en fait qu'une partie de l'indexation.
- La phase de la validation : L'objet pédagogique passe ensuite dans l'étape de validation. Il s'agit d'une validation globale : il faut soumettre à l'avis de différents experts pour juger la qualité de l'objet pédagogique par apport à son contenu, sa forme et sa description.
- La phase de retrait : L'objet pédagogique peut terminer son parcours par son retrait. Il ne peut alors plus être utilisé. l'objet pédagogique ne peut alors plus être recherché par les responsables de formation, même si le document numérique continue à être hébergé.
- La phase de la diffusion : Une fois l'objet pédagogique validé, il doit passer à l'étape de diffusion. Il s'agit de permettre sa distribution. Il faut comprendre que la diffusion se fait pour des objets pédagogiques prêts à l'emploi.
- La phase de la recherche : Pendant l'étape de recherche, le responsable pédagogique a besoin de trouver l'objet pédagogique qui correspond le mieux à la formation qu'il souhaite mettre en place.
- La phase de l'utilisation : Lors de l'étape d'utilisation, l'objet pédagogique est intégré à un dispositif permettant son exploitation. En effet, les apprenants, les enseignants et les tuteurs utilisent généralement les plateformes pédagogiques pour tirer profit de l'objet pédagogique.
- La phase retours d'expériences : Il s'agit, lors de l'étape de retours d'expériences, d'analyser les avis des utilisateurs qui auront pu être collectés lors de la phase d'utilisation pour étudier les évolutions ou adaptations possibles et/ou nécessaires de l'objet pédagogique.

Pour conclure, les objets pédagogiques doivent être autonomes, c'est-à-dire indépendants du support de diffusion et des plateformes d'apprentissage. Outre cela, un objet pédagogique doit être interopérable, d'où l'apparition de plusieurs travaux pour la mise en place de standards et de normes pour la description d'un objet pégagogique.

4.3. Normes pour la description d'un object pédagogique

Dans cette partie de l'état de l'art, nous présentons les principales normes et modèles de métadonnées décrivant des objets pédagogiques dans le domaine de e-learning. En effet, les métadonnées permettent une recherche sémantique efficace des ressources pédagogiques sur le web et leur traitement automatique. En outre, différentes normes ont été définies pour aider à l'élaboration et à la représentation des objets pédagogiques.

L'application de ces normes garanti non seulement l'interopérabilité mais également la qualité des systèmes e-learning. Nous allons citer dans la suite les normes les plus connues en e-learning en l'occurrence LOM [28], AICC [29], SCORM [30] et IMS-LD [31] :

4.3.1. LOM

Le LOM (Learning Object Metadata) est un standard international qui permet de décrire un objet pédagogique en utilisant des métadonnées. Ces objets peuvent être des livres, des scénario pédagogiques, des sites web, des logiciels, etc [32]. Le LOM définit une ressource

pédagogique comme étant une entité, numérique ou physique, qui peut être utilisée, réutilisée ou référencée dans des applications de e-learning. Il fournit un schéma de données conceptuel qui définit la structure d'une instance de métadonnées pour une ressource pédagogique. Les éléments permettant la description des ressources pédagogiques sont groupés dans les catégories suivantes (figure 1.4) :

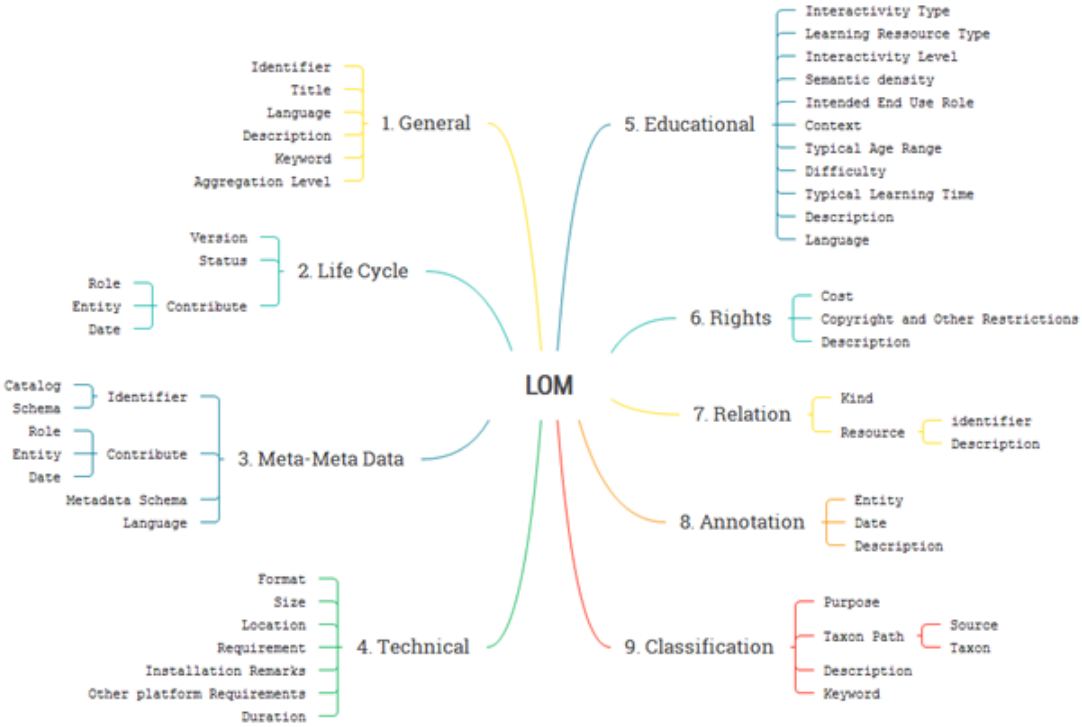

FIGURE 1.4 – Une représentation schématique de la hiérarchie des éléments LOM

1. **Général :** Catégorie qui regroupe toutes les informations générales pour la description de la ressource pédagogique ;
2. **Cycle de vie :** Catégorie qui décrit l'historique et l'état courant de la ressource pédagogique ;
3. **Meta-métadonnée :** Catégorie qui décrit des informations à propos des métadonnées elles-mêmes ;
4. **Technique :** Catégorie qui décrit les exigences et les caractéristiques techniques de la ressource ;
5. **Pédagogie :** Catégorie qui décrit les caractéristiques pédagogiques de la ressource ;
6. **Droit :** Catégorie qui décrit les droits de propriété intellectuelle et les conditions d'utilisation de la ressource ;
7. **Relation :** Catégorie qui décrit les relations entre les ressources pédagogiques ;
8. **Annotation :** Catégorie qui offre des commentaires sur l'utilisation pédagogique de la ressource, ainsi que sur le(s) créateur(s) du commentaire ;

9. **Classification :** Catégorie qui décrit si la ressource pédagogique fait partie d'un système de classification particulier.

Notons que les métadonnées du standard LOM ont été complétées par des métadonnées particulières. En effet, plusieurs extensions du standard LOM ont vu le jour à titre d'exemple LOM-FR [33], SupLOMFR [34], etc.

4.3.2. AICC

AICC est l'acronyme de (Aviation Industry CBT [7] Committee), cette norme est le fruit d'une association de compagnies d'aviation qui utilise depuis longtemps les technologies de l'information pour former ses pilotes [35]. Elle permet de garantir certaines spécificités : gestion du chargement d'un contenu dans un système e-learning, standardisation de la communication entre le contenu et le système, adaptation de la pédagogie du contenu en fonction de l'apprenant.

Cette norme a progressivement été étendu à l'ensemble des problématiques liées à la formation électronique. La compatibilité avec cette norme permet notamment l'interopérabilité entre plates-formes et contenus hétérogènes offrant ainsi des possibilités d'évolution et d'enrichissement élevé. AICC normalise encore la description du déroulement d'une formation selon des paramètres de réussite ou d'échec d'un apprenant, de son profil, de ses compétences de départ, etc [29].

4.3.3. SCORM

SCORM (Shareable Content Object Reference Model) est un ensemble de standards et de spécifications utilisé pour normaliser les communications et les formats d'échange de données, et définit précisément les paquets pour le transfert de fichiers [36]. Le standard SCORM a été créé afin de permettre le partage et la réutilisation des objets d'apprentissage. Puisque les objets crées avec SCORM contiennent un protocole universel de communication avec plusieurs plateformes, ils n'ont pas besoin d'être recréés lorsque l'on choisit de changer de plateforme. SCORM est composé de trois grands volets [37] (figure 1.5) :

- Le modèle d'agrégation du contenu.
- L'environnement d'exécution.
- Le modèle de séquencement et de navigation.

7. CBT : Computer-Based Training

FIGURE 1.5 – Le contenu du package SCORM

- Le modèle d'agrégation du contenu : Il assure la promotion de méthodes cohérentes en matière de stockage, d'identification, de conditionnement d'échange et de repérage du contenu. Il permet aux responsables de la conception et de la mise en œuvre de la formation de regrouper les ressources appropriées dans le but d'offrir un parcours individualisé de formation. Ce volet comporte trois niveaux de métadonnée [24]. Premièrement, l'Asset qui représente la plus petite unité de ressources pédagogiques (page web, image, etc.) et qui ne communique pas avec une plateforme de formation. Deuxièmement, le SCO (Sharable Content Object) composé d'un ou de plusieurs Assets, le SCO utilise l'environnement d'exécution du SCORM pour communiquer avec une plateforme de formation. Troisièmement, le CO (Content Organization) permet de représenter la structure des contenus. Le CO réunit les ressources pédagogiques pour constituer une activité pédagogique.

- L'environnement d'exécution : Cette composante décrit les exigences du système de gestion de l'apprentissage nécessaire à la gestion de l'environnement d'exécution. Elle fonctionne à partir d'un API (Application Programming Interface) et permet aux grains (SCO) de communiquer avec les plateformes et autres applications d'une manière standardisée.

- Le modèle de séquencement et de navigation : Ce modèle permet la coordination des activités pédagogiques en configurant le mode de parcours. Il décrit le cheminement des activités pédagogiques et les conditions de sélection et d'affichage des ressources pédagogiques. Ce modèle range les activités entre eux dans une organisation hiérarchique décrivant une arborescence d'activités. Les règles de séquencement sont utilisées pour déterminer l'ordre des activités présentées à l'apprenant [37].

4.3.4. IMS-LD

La spécification IMS-LD (Instructional Management Systems Learning Design) [31] fait appel à des concepts pédagogiques permettant de modéliser les unités d'apprentissage. Elle prend en compte une grande variété de modèles pédagogiques et utilise la métaphore théâtrale, ce qui implique l'existence de rôles, de ressources et de scénarios d'apprentissage. Une pièce est divisée en un ou plusieurs actes et elle est conduite par plusieurs acteurs qui peuvent assumer différents rôles à différents moments. Chaque rôle doit réaliser un certain

nombre d'activités pour achever le processus d'apprentissage. De plus, tous les rôles doivent être synchronisés à la fin de chaque acte avant de traiter l'acte suivant.

Dans IMS-LD, les activités caractérisées par des objectifs et des prérequis possèdent une structure spécifique, utilisent des ressources et produisent des résultats. Ces résultats peuvent être réinjectés dans d'autres activités. Le modèle IMS-LD permet de spécifier le déroulement d'une unité d'apprentissage, il utilise la norme LOM pour la description des métadonnées relatives aux ressources et reconnaît les objets pédagogiques comme une partie des environnements d'apprentissages. Il place également l'activité au centre du processus. La figure 1.6 montre les relations entre les différents concepts qui constituent IMS-LD [38] :

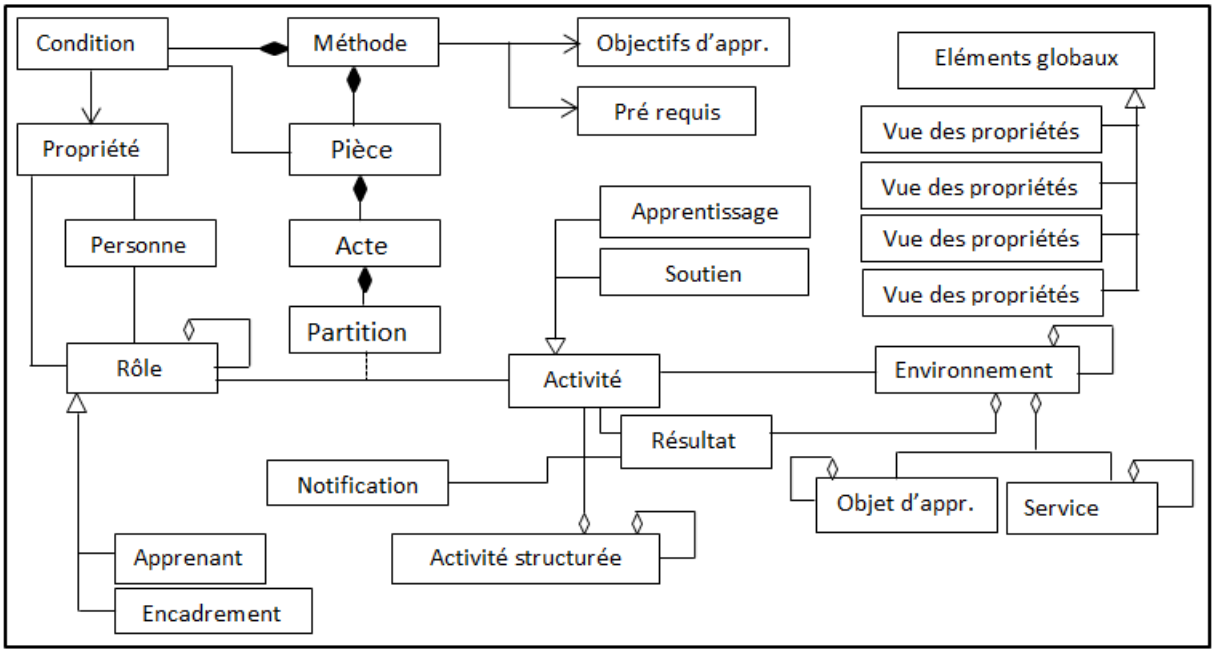

FIGURE 1.6 – Architecture de la spécification IMS-LD

Le développement des solutions de formation assistée par les technologies numériques comporte des étapes de conception de deux types tout au long du processus d'ingénierie pédagogique [39] : La conception didactique qui se concentre sur les connaissances et la conception pédagogique qui s'intéresse à la mise en œuvre des stratégies autour des situations d'apprentissages [38]. La modélisation des environnements d'apprentissage consiste à utiliser le paradigme des objets pédagogiques et à s'appuyer sur une scénarisation avec les notions d'unité d'apprentissage, de ressources et d'activités pédagogiques.

4.3.5. IMS-QTI

La spécification QTI (Question Test Interoperability) est issu des travaux de l'organisation IMS [40] dont le but est de promouvoir le développement de l'apprentissage collaboratif et coopératif à distance. Elle permet de représenter la structure de données d'une question

(assessmentItem) et d'un test (assessment) ainsi que de leurs résultats correspondants [41].

La figure 1.7 illustre la représentation faite par Gamazo [42], de la spécification IMS QTI. Cette dernière traite des questions (assessmentItems) et des essais (assessmentTests). Plus précisément, il propose une architecture logicielle consistant en un dépôt (itemBank) géré par la (itemBankManager) qui stocke les (assesmentItems) pouvant être inclus et réutilisés dans différents (assesmentTests) dans un (learningSystem) donné. Il y a aussi un (authoringTool) permet aux auteurs de gérer (assessmentItems) et un (testConstructionTool) pour construire (assessmentTests). Les tuteurs configurent les matières dans la (learningSystem) pour les apprenants, qui peuvent répondre à (assessmentTest) par un (assessmentDeliverySystem).

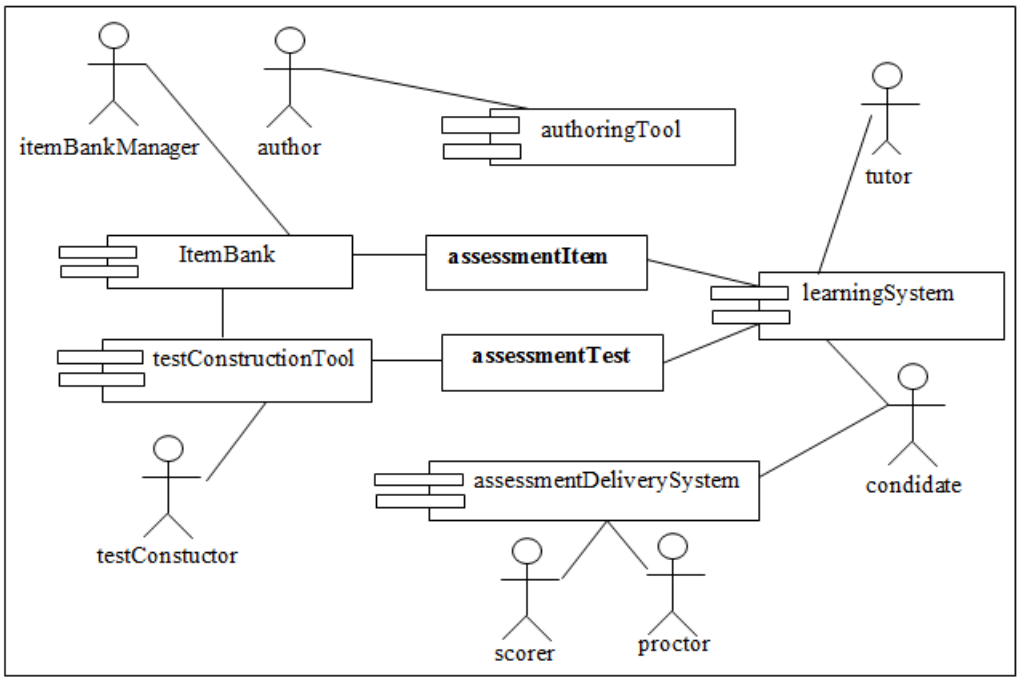

FIGURE 1.7 – Architecture étendu IMS QTI

En résumé, QTI est un standard de représentation des évaluations développés par l'organisme IMS. Il s'agit d'un modèle de données qui définit la structure des questions, les réponses et les résultats de l'évaluation.

Finalement, les concepteurs des systèmes e-learning savent qu'ils ont intérêt à utiliser des environnements qui répondent aux normes et qui sont conformes aux standards de généralisation et d'interopérabilité des ressources pédagogiques.

C'est la raison pour laquelle, nous introduisons dans la section suivante les plateformes d'apprentissage (LMS et LCMS) qui intègrent un ensemble de fonctionnalités conçues pour mettre en œuvre, suivre et gérer les apprentissages, les contenus, les progressions des apprenants et leurs interactions.

5. LMS et LCMS

L'aspect important d'un projet de e-learning est le déploiement. Une fois que votre contenu d'apprentissage est produit, vous devez le mettre à la disposition des apprenants. La question à se poser est la suivante : comment afficher votre ressource en ligne ? Pour répondre à cette question, vous aurez à évaluer vos besoins de l'administration des cours, le partage des ressources, la communication et la collaboration. Cela vous aidera à choisir le système adéquat pour vous : LMS [43] ou LCMS [44]. Mais quelle est la différence entre ces deux acronymes ?

5.1. LMS

Les LMS (Learning Management System) désignent les plates-formes de gestion de la formation par le media internet [43]. Ces systèmes permettent la diffusion des contenus pédagogiques, la gestion de la formation, la gestion des étudiants (inscription, authentification, modification du profil), la gestion des cours et les résultats des étudiants.

Le LMS possède d'autres fonctionnalités telles que les outils de classes virtuelles, la gestion des connaissances, et même des solutions de création de cours. Ils permettent aussi une gestion centrale de la formation, la gestion des utilisateurs (responsables de formation, et étudiants), ainsi que les progrès des étudiants entre les différents modules de formation. Les LMS permettent la gestion de la formation en mode connecté et en mode déconnecté, avec synchronisation des résultats.

5.2. LCMS

Les LCMS (Learning Content Management System) désignent les systèmes de gestion de contenu d'apprentissage qui se concentrent principalement sur la création de contenu [44]. En d'autres termes, les développeurs et les administrateurs créent du matériel, tels que des articles, des tests, des jeux, des vidéos et de petites unités de contenu numérique (grains pédagogiques), qui sont ensuite rapidement assemblés, réutilisés puis adaptés à différents cours selon les besoins des apprenants. Les LCMS réduisent les efforts de développement et permettent de réutiliser facilement du contenu numérique.

Les outils LCMS et LMS sont des outils très proches, quelques fois, complémentaires. Cependant, la distinction entre ces deux outils n'est pas évidente dans la mesure où les plates-formes LMS intègrent souvent en standard des fonctionnalités de LCMS, et vice versa. En effet, historiquement les LMS (premières solutions apparues sur le marché de e-learning), proposaient des applications de création de contenus. La gestion des compétences est une des fonctionnalités partagées par les deux types de solutions. Pourtant, cette gestion est plus performante dans les LMS puisqu'elles sont centralisées et permettent d'établir le profil et les compétences d'un apprenant par rapport à toutes les formations qu'il a pu faire. La gestion des compétences dans les LCMS s'intéresse davantage aux compétences acquises lors d'une formation.

Pour mieux comprendre la distinction et la complémentarité de ces deux familles de solutions, nous proposons dans la suite une comparaison entre les deux plateformes d'apprentissage.

5.3. Comparaison entre les deux systèmes LMS et LCMS

Le tableau 1.1 illustre les caractéristiques communes entre les deux plateformes d'apprentissage LMS et LCMS :

	LMS	LCMS
Inscription des étudiants	Oui	Non
Affectation de tuteurs	Oui	Non
Rapports de formation	Oui	Non
Bibliothèque de formation	Non	Oui
Gestion de compétences	Oui	Non
Analyse des écarts	Oui	Non
Création de cours	Oui	Oui
Accès au cours	Oui	Non
Accompagnement des étudiants	Oui	Non
Classes virtuelles	Oui	Non

TABLE 1.1 – Comparaison entre les deux systèmes LMS et LCMS

La principale différence entre un LMS et un LCMS est que le LCMS est plus axé sur le développement du contenu pédagogique, sur sa gestion et sur sa création. En outre, les plateformes d'apprentissage sont souvent utilisées de manière interchangeable, et malgré les différences qui existent entre ces plateformes, elles possèdent de nombreuses caractéristiques communes.

6. Exemples de plateformes de formation à distance

Dans cette section, nous allons décrire en détail quelques plates-formes dédiés au e-learning :

6.1. Moodle

Le LMS Moodle a été conçu comme une plateforme de e-learning extrêmement modulaire [45]. Il est compatible avec les normes SCORM, AICC, QTI et IMS. Le terme Moodle est l'acronyme de "Modular Object-Oriented Dynamic Learning Environment" qui désigne en français "Environnement orienté objet d'apprentissage dynamique modulaire". La communauté d'utilisateurs de Moodle est l'une des plus importantes dans le domaine de e-learning open-source. De nombreuses sociétés proposent de l'assistance et du conseil autour de Moodle. La conception de la plateforme Moodle a été influencée par les travaux de recherches

doctorales de Martin Dougiamas (auparavant administrateur de la plateforme WebCT) [46]. Martin Dougamias a étudié les apports du constructivisme social dans la pédagogie en ligne. Moodle est considérée comme une plateforme d'apprentissage en ligne servant à créer des communautés d'apprenants autour de contenus et d'activités pédagogiques. Elle est dotée d'un système de gestion de contenu (SGC) performant [47].

6.2. Claroline

Claroline [48] est le LMS le plus utilisé dans le monde de l'éducation, et moins dans celui de l'entreprise. Il est compatible avec les normes SCORM et QTI. Claroline est une plateforme libre et gratuite. Elle est développée en 2002 par l'université de Louvain en Belgique pour la formation à distance et le travail collaboratif entre les apprenants. Elle permet aux formateurs de créer des espaces de cours en ligne et de gérer des activités de formation sur internet. La plateforme Claroline bénéficie de l'appui d'une communauté mondiale d'utilisateurs et de développeurs, elle permet de créer sans coût de licence des espaces de travail et des cours en ligne.

6.3. Ganesha

Ganesha [49] met l'accent sur les parcours individualisés et le suivi des apprenants. Cela le différencie de la plupart des autres LMS qui sont plutôt orientés « contenu ». Ganesha est développé par l'entreprise ANEMA [49] qui propose également des services pour le développement de contenu et de la mise en place de la plateforme. Ganesha est compatible avec les normes SCORM et AICC. Les parcours pédagogiques sont constitués de séquences par le créateur de cours. Le contenu est caractérisé par une granularisation extrême qui permet une individualisation de l'apprentissage. On fait appel intensivement au processus d'ingénierie pédagogique dans la création de parcours. L'apprenant est intégré dans des groupes ou classes à qui on affecte des parcours pédagogiques.

6.4. Dokeos

Dokeos [50] est un projet dérivé du projet Claroline mais évoluant vers un style plus commercial basé sur les nouvelles technologies. Récemment un style plus professionnel et plus moderne a été adopté par la plateforme, elle utilise un site web fonctionnel et très complet, des services payants sont proposés pour des hébergements, des formations, etc.
Dokeos est aussi un réseau de sociétés de services qui fournissent du conseil et d'autres services : développement, formation, notamment auprès de grandes entreprises et des administrations publiques.

6.5. Récapitulatif sur les plateformes présentées

Le tableau 1.2 résume les caractéristiques des différentes plateformes décrites précédemment.

		Plateformes			
		Moodle	Claroline	Ganesha	Dokes
Adaptabilité	Personnalisation par les formateurs du menu de navigation	Non	Oui	Non	Non
	Personnalisation par l'apprenant de l'interface du cours	Non	Non	Non	Oui
	Personnalisation du contenu en fonction du profil de l'apprenant	Non	Non	Non	Non
Collaboration	Forum	Oui	Oui	Oui	Oui
	Wiki	Oui	Oui	Non	Oui
	Groupes d'apprenants	Oui	Oui	Oui	Oui
	Communautée d'apprentissage	Oui	Non	Oui	Oui
Evaluation	Description, réponse courte, mise en correspondance, texte à trous, texte lacunaire où des mots manquants	Oui	Oui	Oui	Oui
	QCM Questionnaire à choix multiples	Oui	Oui	Oui	Oui

TABLE 1.2 – Comparaison entre les différentes plateformes e-learning existantes

Nous remarquons qu'aucune de ces plateformes ne propose une adaptation des contenus suivant le profil de l'apprenant, seuls les paramètres d'affichage peuvent être modifiés.

7. Le tutorat en ligne

La réussite d'une formation e-learning ne réside pas uniquement dans la qualité des outils de formation mais aussi dans l'accompagnement humain mis en place pour optimiser la formation. Cet accompagnement humain se traduit par le tutorat en ligne [51]. Le tutorat est une relation formative entre un enseignant, le tuteur, et un apprenant (ou un petit groupe des apprenants). Il se distingue de l'enseignement classique par l'impliquation des professeurs et des apprenants pour une formation individualisée et flexible. Le tuteur n'a pas forcément toutes les connaissances que doit maîtriser l'apprenant au fil de sa formation, car son rôle n'est pas d'apporter des réponses aux problèmes posés mais de guider l'apprentissage.

7.1. E-tutorat : élément essentiel dans le processus d'apprentissage en ligne

Le tutorat s'applique dans des contextes différents dans lesquels le tuteur peut avoir différents statuts. Un tuteur en entreprise aura généralement un statut juridique soumis à une règlementation stricte alors que le rôle de tuteur en école pourra être pris par n'importe quel enseignant ou même étudiant. Dans le contexte de l'enseignement, nous pouvons définir

le tutorat comme une forme d'aide individualisée qui vise à apporter en dehors du contexte de la classe, une aide personnalisée. Nous parlons alors de coaching pédagogique. En outre, cette définition large peut prendre de nombreux aspects, notamment selon :

- À qui on s'adresse : à un étudiant seul ou à un groupe d'étudiants.
- Le tuteur est un professeur, ou un autre étudiant, nous parlons de "tutorat par les pairs" [52].

Selon Bourdet "Le tutorat en ligne est l'accompagnement à distance d'un apprenant ou d'un groupe d'apprenants par les moyens de communication et de formation que permettent aujourd'hui l'informatique, le multimédia et internet" [51].
En résumé le tutorat est une fonction pédagogique à part entière, dont la finalité est d'évaluer, encourager et renforcer les capacités des apprenants à travailler de manière autonome, individuellement ou en groupe.

7.2. Le rôle du tuteur sur une plateforme e-learning

Les établissements confient généralement aux personnes tutrices la responsabilité de l'encadrement des étudiants à distance. Selon Paquette [53], le tuteur se trouve dans une position d'intermédiaire pour favoriser le dialogue constructif entre les apprenants et les soutenir dans leurs transactions. Il met en perspective les différentes modalités d'intervention du tuteur.

Paquette établit une distinction entre tutorat individuel et tutorat collectif. Le tutorat individuel se réalise de personne en personne, entre le tuteur et un apprenant. Il a pour objectif de guider l'étudiant dans son processus de connaissance, de le conseiller, de lui permettre de prendre conscience de ses capacités et l'aider à apporter les changements qu'il souhaite dans ses pratiques. Le tutorat collectif s'adresse au groupe d'apprenants ; le tuteur représente l'établissement devant les étudiants et tente notamment de mettre en place un sentiment d'appartenance à une communauté.

7.3. Le travail collaboratif en ligne

Le travail collaboratif [8] désigne un travail qui n'est plus fondé sur l'organisation hiérarchisée traditionnelle mais,plus précisément, un nouveau mode de travail où collaborent de nombreuses personnes grâce aux technologies de l'information et de la communication.
Le travail collaboratif en ligne est un moyen actif par lequel l'apprenant travaille à la construction de ses connaissances, c'est une stratégie pédagogique qui favorise l'échange, l'interaction entre les apprenants et le partage d'un but commun [54].

Nous définissons, l'apprentissage collaboratif comme toute activité d'apprentissage réalisée par un groupe d'apprenants ayant un but commun, étant chacun source d'information, de motivation, d'interaction, d'entraide et bénéficiant chacun des apports des autres, de l'alliance du groupe et de l'aide d'un formateur facilitant les apprentissages individuels et

8. "Ce qu'un enfant peut faire aujourd'hui en collaborant avec autrui, peut le faire seul demain" Vygotski

collectifs [55].
Mais il faut distinguer entre travail collaboratif et travail coopératif ?

7.3.1. Distinction entre travail collaboratif et travail coopératif

Dans le cadre d'un travail coopératif, il y a une répartition claire du travail à réaliser. Il est assigné à chaque élève une tâche claire. Par la suite, les travaux individuels de chaque élève sont rassemblés et forment le travail final. Dans ce cas, l'apprenant est responsable de sa propre production mais il doit néanmoins apprendre à interagir avec les autres apprenants afin que le travail final puisse être cohérent. L'apprentissage coopératif est une méthode d'enseignement selon laquelle des élèves travaillent ensemble en petites équipes afin d'atteindre un but commun [56].

Par contre, dans le cadre d'un travail collaboratif, il n'y a aucune répartition du travail entre les apprenants. Ces derniers travaillent tous ensemble à chaque étape de l'élaboration du travail. Il est donc impossible, une fois le travail réalisé, d'identifier le travail fourni par chacun. D'où il ne faut pas confondre «travail collaboratif» et «travail coopératif» qui sont différents puisque le travail coopératif est une coopération entre plusieurs personnes qui interagissent dans un but commun mais se partagent les tâches, alors que le travail collaboratif se fait en collaboration du début à la fin sans diviser les tâches [57].

7.3.2. Acteurs et tâches du travail collaboratif en ligne

Le travail collaboratif en ligne fait intervenir comme acteurs : le tuteur, l'apprenant, le modérateur et le secrétaire. D'après [55] les trois tâches d'un travail collaboratif sont : la coordination, la communication et le partage.

- **La coordination** : vise l'agencement efficace des activités, des personnes et des ressources pour atteindre un but. Pour mieux canaliser et coordonner les énergies et les activités du groupe, trois variables sont à contrôler par l'enseignant ou par les apprenants : la tâche, la composition du groupe et l'animation.
- **La communication** : amène l'apprenant à exprimer des idées dans le but de les partager avec le groupe, à faire des liens entre les idées exprimées pour permettre l'émergence des nouvelles idées et à structurer les idées pour leur donner du sens et construire des connaissances. La coordination constitue, avec la communication, l'une des deux pierres angulaires de la collaboration.
- **Le partage** : l'apprentissage collaboratif invite les étudiants à partager, à instaurer des interdépendances positives et à s'investir dans des productions communes. Le partage pédagogique peut prendre plusieurs formes : partage des idées, des activités (apprentissage, enseignement, évaluation des apprentissages), des matériels pédagogiques ou des connaissances pédagogiques.

7.3.3. Les indicateurs d'analyse automatique des interactions

Les concepteurs des plateformes e-learning adoptent l'Analyse Automatique des Interactions (AAI) [58] qui ont comme objectif principal l'analyse des interactions en vue d'assister les différents acteurs, en particulier, le tuteur dans le suivi des différentes activités des apprenants et des groupes d'apprenants (figure 1.8).

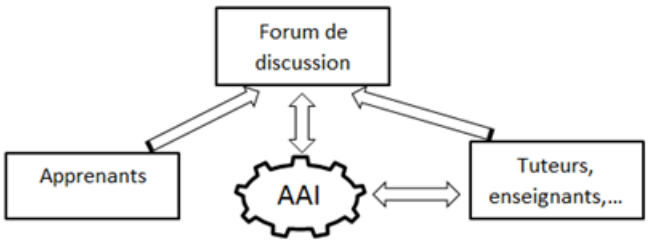

FIGURE 1.8 – Schéma général des utilisateurs des outils d'analyse des interactions

La fonction d'AAI consiste à capter, filtrer et traiter les données de l'environnement informatique afin de produire des indicateurs d'analyse de l'action et de l'interaction [58]. Plusieurs systèmes d'AAI ont été développés, citons par exemple l'outil I-Bee [59], qui s'adresse aux élèves qui discutent de façon asynchrone via un forum. L'outil I-Bee produit comme résultat d'analyse trois indicateurs :

- La popularité de chaque sujet de discussion,
- Le degré d'activité de chaque participant,
- Le sujet principal de discussion de chaque participant.

Un autre outil d'analyse automatique des interactions est DIAS [60] qui offre un grand nombre d'indicateurs, afin d'assister les étudiants participant à un forum, ainsi que les tuteurs, les administrateurs du forum. De plus, l'outil Analytic Tool est un ensemble de mécanismes reliés à l'environnement de collaboration asynchrone Knowledge Forum [61], qui offrent aux enseignants un ensemble d'indicateurs (relatifs à l'évolution du vocabulaire, ou une visualisation du champ sémantique des thèmes élaborés par les élèves,...). La diversité et la pertinence des indicateurs retournés par ces systèmes d'AAI fait émerger la question de leur réutilisation dans des contextes et environnements différents.

La nature d'un indicateur correspond aux aspects de l'interaction qu'il tend à faire émerger. Il est relié (directement ou indirectement) à une ou plusieurs des dimensions suivantes :

- **Dimension cognitive** indiquant quelques caractéristiques sur les opérations cognitives de l'individu ou du groupe, relatives à l'acquisition des connaissances lors des activités d'apprentissage.
- **Dimension sociale** liée aux activités de communication, de coopération ou de collaboration d'un groupe ou d'une communauté des apprenants.

- Dimension affective liée à la situation affective des apprenants.

Dans ce qui suit nous expliquons en détail les différents types d'indicateurs :

8.3.3.1 Indicateurs de nature cognitive :

Ces indicateurs concernent les interactions des participants liées au contenu de l'activité [62]. Ainsi, nous avons défini deux catégories d'indicateurs de nature cognitive :

◇ *Indicateurs relatifs au déroulement de l'activité d'apprentissage :*
Dans cette catégorie nous avons l'indicateur «Profondeur de Discussion» [63] qui affiche sous forme de graphe les messages relatifs à une discussion dans le forum en fonction du temps. Le graphe est sous forme d'arbre ; où les nœuds présentent les messages étiquetés par l'identifiant de l'apprenant expéditeur, et les flèches relient chaque message au message source. Cet indicateur permet d'avoir une traçabilité de la progression dans un forum en fonction du temps.

◇ *Indicateurs relatifs au contenu de l'activité d'apprentissage :*
Dans cette catégorie nous pouvons citer trois indicateurs : «Productivité des apprenants dans le mail», «Productivité des apprenants dans le forum» et «Productivité des groupes». Ces indicateurs sont calculés à partir des données recueillies dans les espaces forum et mail en se basant sur le nombre de documents déposés par chaque apprenant.

8.3.3.2 Indicateurs de nature sociale :

Ces indicateurs se réfèrent aux modes ou à la qualité de communication et de collaboration au sein d'un groupe. Parmi les indicateurs ayant une valeur d'interprétation relativement élevée, nous pouvons noter ceux qui favorisent *la prise de conscience de l'espace du travail*, ceux qui rendent compte de *la qualité de la collaboration* au cours de la discussion et ceux qui fournissent *un état des relations établies entre les participants.*

◇ *Indicateurs relatifs à la prise de conscience de l'espace du travail :*
Concerne les actions et les contributions des autres membres, fondées sur des indicateurs simples comme «le nombre des nouveaux messages postés», «le nombre des messages non lus» par chaque individu. Un indicateur similaire mais plus élaboré est celui de «complexité» qui représente la complexité des interactions et rend explicite le degré de difficulté à poursuivre toutes les conversations dans un forum [64].

◇ *Indicateurs relatifs à la qualité de la collaboration :*
La majorité des indicateurs existants permettent de caractériser la participation. Ainsi, le «degré de présence» dans un forum mesure la distribution et la fréquence des contributions des participants depuis l'ouverture d'une discussion sur un forum [65]. Le «niveau d'interaction dans un forum» mesure la distribution et la fréquence des contributions des participants selon qu'ils initient un nouveau fil de discussion ou répondent à un message précédent [66]. D'autres indicateurs soutiennent la coordination entre les membres, ainsi «coordination» mesure le degré de communication qui apparaît entre les membres d'un

groupe.

◇ *Indicateurs de l'état des relations entre apprenants :*
Une sous-catégorie significative des indicateurs sociaux représente les relations sociales des participants à un forum. Ainsi les diagrammes d'analyse de réseaux sociaux représentent entre autres des informations relatives aux relations établies au sein d'un groupe tel le «degré de centralité des acteurs» [67]. Ce type de diagramme permet également de repérer les membres isolés, ainsi que ceux qui dominent les interactions [68]. Nous pouvons citer aussi la «cohésion du groupe» qui représente l'habilité d'un groupe à tenir ses membres, c'est-à-dire le nombre minimum de participants qui déconnectent le groupe, s'ils partent.

8.3.3.3 Indicateurs de nature affective :

La participation effective dans un processus d'apprentissage nécessite une maturité émotionnelle, de la prise de conscience, de l'empathie, du contrôle, et une prise en compte des émotions des autres personnes [69]. Les qualités et les habiletés de nature affective interviennent de manière significative dans la construction des relations dans un groupe.

Cette dimension affective est apparue très récemment dans le champ de l'analyse des interactions. Avec les systèmes existants, nous identifions par exemple, l'indicateur «motivation individuelle» qui représente la motivation en fonction du temps [70]. Pour le calculer, les individus sont invités à indiquer et exprimer leur propre niveau de motivation, chaque fois qu'ils font une contribution lors d'un travail collaboratif.

8.3.3.3 Synthèse sur l'ensemble des indicateurs :

Dans tous les cas, il est à noter que la dimension affective s'utilise plutôt pour des fonctions d'observation et d'autorégulation et n'influence pas, jusqu'à présent, les indicateurs qui peuvent intervenir sur les fonctionnalités. De nouvelles tendances de recherche se consacrent à la définition d'indicateurs significatifs pouvant être calculés de manière automatique.

Concernant les indicateurs sociaux, certains chercheurs considèrent que ces indicateurs produisent plutôt des vues abstraites fonctionnant comme des substituts aux communications orales [70]. Pourtant, ces indicateurs sociaux contribuent à faire émerger des aspects des processus cognitifs. En ce sens, il semble assez prometteur de continuer la recherche d'identification d'indicateurs sociaux appropriés susceptibles d'offrir un soutien aux utilisateurs des espaces d'interactivité en ligne. Enfin, les indicateurs de nature affective constituent une nouvelle dimension qui devrait être explorée de façon plus systématique dans les années à venir.

Dans la section suivante nous illustrons, le plus simplement possible, les systèmes d'apprentissages adaptatifs dont l'objectif principal est d'adapter les décisions pédagogiques aux compétences et besoins particuliers de chaque apprenant. Ces systèmes pédagogiques adaptatifs promettent de prendre en considération le profil de l'apprenant (ses connaissances, ses préférences, ses aptitudes...) dans la construction d'un parcours pédagogique unique et adapté.

8. E-learning adaptatif

De nos jours, de nombreuses solutions ont été mises en pratique dans le cadre de l'apprentissage à distance. La majorité de ces solutions privilégie une approche centrée sur la mise à disposition de ressources pédagogiques de qualité. Ainsi, de nouvelles approches émergent, la qualité du service pédagogique rendu dépend de la capacité de ces nouvelles approches à fournir aux apprenants, d'une part, des contenus pédagogiques adaptés à leur profil et d'autre part, des processus qui les guident véritablement dans leur processus d'apprentissage. Les systèmes pédagogiques adaptatifs ont pour objectif de répondre à ce besoin.

Par ailleurs, e-learning adaptatif [71] est une nouvelle tendance dans le domaine de l'éducation et de la formation, permettant d'adapter le cours à l'apprenant. Son objectif est d'amener chacun à progresser, quelle que soit sa forme d'intelligence ou son niveau de départ, d'éviter les frustrations ou les abandons en cours de route, de créer chez les apprenants davantage de plaisir et d'envie d'apprendre.
C'est une théorie d'enseignement d'origine américaine, datant des années 1970, qui permet d'individualiser les parcours d'apprentissage.

D'après Bourdeau (2010), un système pédagogique adaptatif est basé sur trois modèles différents [72] :
◇ Un modèle des ressources pédagogiques qui contient la description des ressources pédagogiques proposées aux apprenants ;
◇ Un modèle de l'apprenant qui décrit les différents types de connaissances relatives aux apprenants ;
◇ Un modèle d'apprentissage qui définit les différentes méthodes et activités d'apprentissage mises en œuvre pour atteindre des objectifs.

Pour introduire le plus simplement possible les systèmes d'apprentissages adaptatifs, concentrons-nous sur le schéma de la figure 1.9.

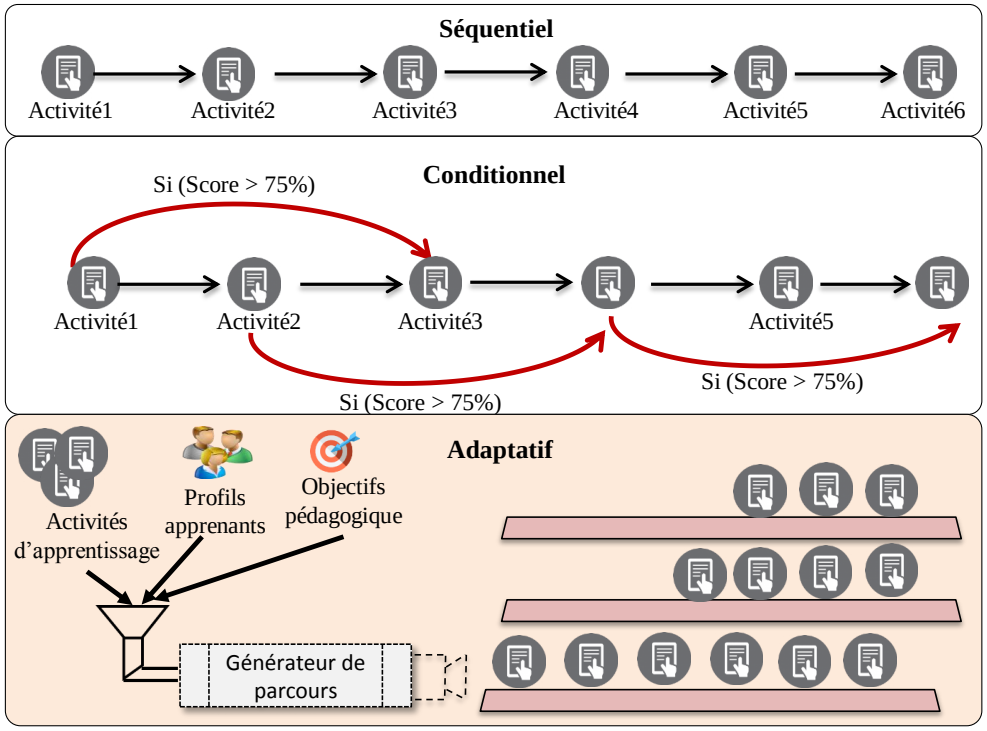

FIGURE 1.9 – Comparaison entre les types de parcours pédagogiques

Ce schéma illustre trois type de parcours pédagogiques :
le parcours séquentiel : C'est le e-learning traditionnel où nous devons suivre des parcours séquentiels. L'apprenant suit un apprentissage linéaire (une activité, puis l'autre, etc).
le parcours conditionnel : Le parcours est conditionné par les résultats obtenus dans les activités précédentes.
Le parcours adaptatif : Le parcours de l'individu est généré selon ; les ressources pédagogiques, le profil de l'apprenant (ses connaissances de bases, ses préférences, ses aptitudes) et l'objectif d'apprentissage.

Les méthodes et les techniques d'adaptation assurent la construction de parcours individualisés en utilisant les trois modèles. Les approches des e-learning adaptatifs existantes n'utilisent pas obligatoirement ces trois modèles. Par exemple, les approches issues du domaine des systèmes pédagogiques « intelligents » décrivent les processus d'apprentissage et peu le modèle de l'apprenant, alors que les approches issues du domaine des systèmes hypermédias adaptatifs [73] ne prennent pas en compte le processus d'apprentissage et se concentrent sur le modèle de l'apprenant. Par ailleurs, les modèles des ressources pédagogiques peuvent être très différents dans leur contenu et dans leur forme d'une approche à une autre. Les technologies actuelles dans le domaine du e-learning (les plateformes LMS, les normes telles que SCORM, LOM ou IMS) semblent en tout cas désormais matures pour porter des systèmes pédagogiques adaptatifs.

8.1. Système hypermedia adaptatifs

Depuis quelques années, les documents du web ne sont plus seulement textuels, plusieurs documents multimédia (image, son, vidéo, animation . . .) sont utilisés. L'apport et les bénéfices du multimédia s'avèrent très évident, notamment dans les systèmes de e-learning. L'hypermédia, obtenu par la fusion des techniques de l'hypertexte et du multimédia, présente plusieurs avantages dans le cadre de l'apprentissage électronique.

En effet, d'une part, la composante multimédia améliore l'aspect visuel de l'apprentissage, et ainsi renforce l'intérêt de l'apprenant par rapport au système d'apprentissage [74]. D'autre part, le composant hypertexte améliore la qualité de l'apprentissage grâce à sa structure non linéaire, en aidant l'apprenant à construire sa connaissance [75]. Cependant, un hypermédia ne peut pas offrir des contenus personnalisés. Les apprenants ont alors accès au même ensemble de contenus sans prendre en compte leurs différences : niveau de connaissances, intérêts, motivations, objectifs, etc. Différentes recherches dans le domaine des hypermédias adaptatifs essayent de comprendre les relations entre le profil de l'apprenant, le contexte dans lequel se déroule l'apprentissage et les contenus, à des fin d'adaptation selon les besoins de l'apprenant.

8.2. Généralités sur l'adaptation

8.2.1. L'adaptation

Dans la littérature, plusieurs définitions sont utilisées pour référencer la notion d'adaptation. L'adaptation c'est l'action d'adapter ou de s'adapter qui veut dire ajuster, joindre, rattacher. Cette définition est générale et ne permet pas de donner une précision sur la notion de l'adaptation dans le domaine de l'apprentissage. (Paccou ,2002) [76] définit l'adaptation comme "une tentative de modifier le comportement interactif d'un système en considérant à la fois les besoins individuels des apprenants humains et les conditions propres à l'environnement de l'application". En effet, deux éléments sont importants pour la réalisation de l'adaptation : les besoins des apprenants et les éléments propres à l'environnement de l'application.

Pour (Vieville, 2005) [77], l'adaptation correspond à un processus par lequel un sujet, lorsqu'il enregistre une variation de l'environnement, modifie les paramètres d'un objet, à partir d'un modèle de référence, dans le but d'accomplir une tâche spécifique. Dans cette définition, pour aboutir à une adaptation efficace et complète, trois domaines sont évoqués : un sujet qui va réaliser l'adaptation, un objet qui va être adapté, un modèle de référence constituant le référentiel sur lequel le sujet va se baser pour adapter l'objet. Cette définition prend en compte également les variations de l'environnement qui spécifient les conditions de déclenchement du processus de l'adaptation.

8.2.2. La personnalisation

L'auteur (Bollet, 2002) [78], définit la personnalisation comme "toute interaction avec le client dans laquelle le contenu, l'offre ou le message a été taillé sur mesure pour un apprenant ou groupe d'apprenants spécifiques". Cette définition affirme que la personnalisation porte sur deux critères qui sont l'interaction et les aspects de personnalisation à savoir les messages envoyés à l'apprenant, les offres ou les contenus qui lui sont présentés. Pour (Chen, 2005) [79], le principal élément de la personnalisation est la prise en compte d'éléments propres à l'environnement de l'application. Alors que l'adaptation, prend en charge, en plus des éléments de l'environnement, les modifications techniques nécessaires à l'emploi des éléments. Ce qui en résulte c'est que la personnalisation n'est pas un concept distinct de celui de l'adaptation, mais plutôt une sous-catégorie de ce dernier (figure 1.10).

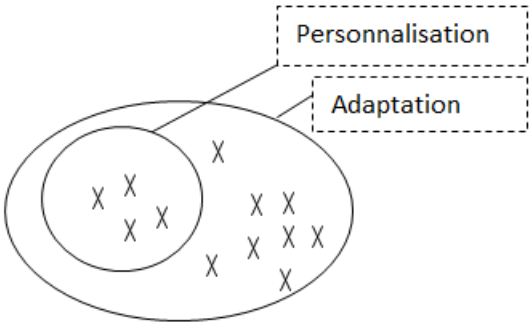

FIGURE 1.10 – Positionnement de la personnalisation par rapport à l'adaptation

8.2.3. Le processus d'adaptation

D'après (Villanova, 2002) [80], Il existe deux modes d'adaptation dans le e-learning adaptatif (l'adaptabilité et l'adaptativité) en fonction de qui prend l'initiative ; le système ou l'apprenant :

- Le système d'apprentissage : Nous parlons alors d'*adaptativité*, c'est la capacité d'un système de modifier automatiquement sa présentation en fonction des caractéristiques des apprenants.

- L'apprenant : Il s'agit de l'*adaptabilité*, c'est la possibilité pour un utilisateur de faire des changements sur certains paramètres d'un système d'apprentissage, pour qu'il adapte son comportement et le parcours d'apprentissage en conséquence.

L'adaptation est déterminée en liaison avec le degré d'implication de l'apprenant et du système (figure 1.11).

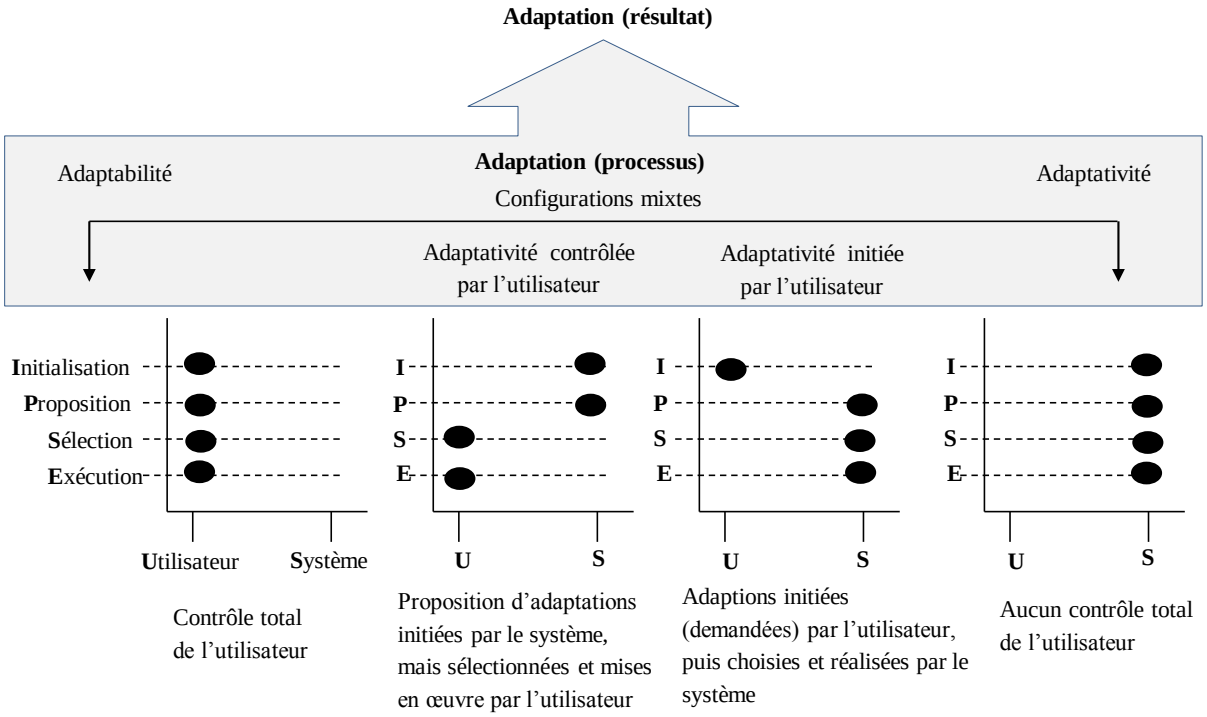

FIGURE 1.11 – Différents types de processus d'adaptation : de l'adaptabilité à l'adaptativité

L'adaptabilité est donc réalisée lors de l'initialisation du système. Les données utilisées par le système sont statiques et donc inchangeables au cours de l'utilisation. Inversement, l'adaptativité est réalisée au fur et à mesure de l'avancement de l'utilisation du système. Les données pour la réalisation de l'adaptation sont dynamiques et sont mises à jour par les systèmes au cours de l'interaction avec l'apprenant. La figure 1.11 illustre quatre types de processus d'adaptation :
Type 1 : Contrôle total de l'utilisateur.
Type 2 : Propositions d'adaptations initiées par le système, mais sélectionnées et mises en œuvre par l'utilisateur.
Type 3 : Adaptations initiées (demandées) par l'utilisateur puis choisies et réalisées par le système.
Type 4 : Aucun contrôle de l'utilisateur (Adaptation automatique réalisée par le système).

8.3. Principales composantes d'un système pédagogique adaptatif

Le e-learning adaptatif est un moyen d'apprentissage et d'enseignement qui utilise un système de tutorat intelligent capable d'adapter, dynamiquement, et de manière systémique, l'apprentissage en ligne. Nous citons dans cette section les principales composantes d'un système pédagogique adaptatif : (figure 1.12).

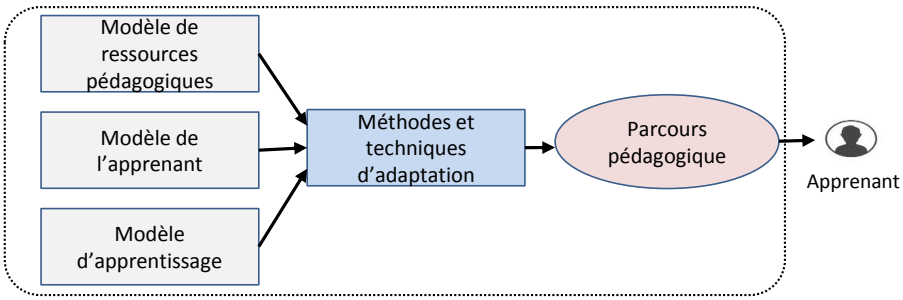

FIGURE 1.12 – Les méthodes et techniques d'adaptation sont appliquées aux trois modèles (ressources pédagogiques, apprenant et apprentissage).

Modèle des ressources pédagogiques : exprime la connaissance sur le sujet enseigné. Cette connaissance peut être décrite à différents niveaux d'abstraction et dans différentes formes. Nous distinguerons deux facettes pour caractériser les ressources pédagogiques, la facette « structure logique » et la facette « structure pédagogique ».
Modèle Apprenant : définit les caractéristiques de l'apprenant. Toutes ces connaissances sont utilisées dans le cadre de la construction de parcours individualisés. Nous caractérisons le modèle de l'apprenant selon deux points de vue : la nature des connaissances sur l'apprenant et le mode de gestion du modèle de l'apprenant.
Modèle d'apprentissage : spécifie à différents niveaux de détail les processus d'enseignement. En considérant que des processus différents peuvent être utilisés pour apprendre, un des objectifs des systèmes adaptatifs est de suggérer une démarche adaptée à chaque apprenant. Nous caractérisons le modèle d'apprentissage par deux facettes : le niveau d'abstraction auquel se situe la démarche d'apprentissage et l'approche de modélisation utilisée.

8.4. Les critères d'adaptation

La réalisation de l'adaptation se base sur plusieurs critères et aspects de l'apprenant. (Brusilovsky, 1998) [81] a identifié trois critères qui sont : les objectifs à atteindre par l'apprenant, ses connaissances concernant le ou les concepts d'un domaine donné, et ses préférences portant sur la présentation des contenus d'apprentissage.

8.4.1. Les préférences de l'apprenant

C'est la caractéristique prise en compte par un système e-learning adaptatif. Chaque apprenant a des préférences et des choix bien précis qui se manifestent par le choix d'un contenu plutôt que d'autre. Les préférences ne peuvent pas être déduites par le système, c'est à l'apprenant de les préciser.

8.4.2. L'objectif de l'apprenant

Pour tout système de e-learning adaptatif, déterminer l'objectif c'est répondre à la question suivante : «pourquoi l'apprenant utilise le système e-learning et quel est son objectif ?» [81]. L'objectif est une caractéristique liée à l'apprenant selon le contexte du domaine

d'enseignement. Ce paramètre est variable puisqu'il change d'une session d'apprentissage à une autre et peut même changer au cours d'une même session.

8.4.3. Les connaissances de l'apprenant

Les connaissances de l'apprenant constituent le critère le plus important pour les systèmes e-learning adaptatifs existants. D'ailleurs, la plupart des techniques de présentation adaptatives utilisées se basent sur les connaissances de l'apprenant comme principale source d'adaptation. Les connaissances de l'apprenant sont aussi variables puisqu'elles changent au fur et à mesure que l'apprenant avance dans son enseignement. Cependant, le système doit suivre le changement des connaissances de l'apprenant et mettre à jour le model de l'apprenant correspondant.

8.5. Les méthodes et techniques d'adaptation

Une méthode d'adaptation désigne une idée conceptuelle, une manière d'envisager l'adaptation, alors qu'une technique d'adaptation correspond aux moyens mis en œuvre pour implémenter une méthode (représentations des connaissances, algorithmes d'adaptation, etc.) [81]. Pour concevoir un système e-learning adaptatif, il est nécessaire de répondre à un certain nombre de questions (figure 1.13), afin de déterminer les méthodes et techniques nécessaires à cette conception [23] :

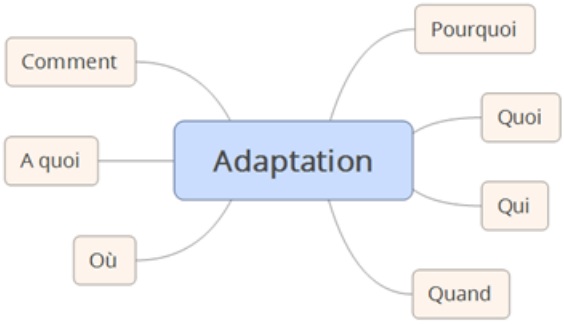

FIGURE 1.13 – Différents questions à posées lors du processus d'adaptation

Que peut-on adapter ? Les méthodes d'adaptation correspondent respectivement, soit à l'adaptation du contenu, soit à l'adaptation de la navigation entre les contenus. L'adaptation du contenu se décline en deux sous méthodes : l'adaptation de texte, ou l'adaptation de média. Pour le texte, l'adaptation consiste à utiliser des variantes des textes, à rajouter des explications, etc. L'adaptation de média est beaucoup moins développée. Certains systèmes permettent la substitution de média, mais les fichiers audio ou vidéo ne sont pas modifiés en fonction de l'utilisateur. L'adaptation de navigation consiste à aider l'utilisateur à se repérer en modifiant les liens qui lui sont proposés et en l'obligeant à utiliser certains liens plutôt que d'autres [82].
A quels éléments le système peut-il s'adapter ? Il existe quatre catégories de données auxquelles un système e-learning peut s'adapter : Premièrement, le système peut s'adapter

aux connaissances de l'utilisateur concernant un domaine. Deuxièmement, il peut s'adapter aux buts de l'utilisateur (que doit-il apprendre ?, quelle tâche souhaite-t-il réaliser ?). Troisièmement, il peut s'adapter aux expériences et aux compétences de l'apprenant. Finalement, le système peut s'adapter aux préférences de l'apprenant concernant la présentation d'un contenu (exemple, taille des caractères, couleurs, etc.).
Quelles méthodes et techniques d'adaptation peut-on employer ? L'adaptation peut s'appliquer soit aux contenus soit à la navigation suivant plusieurs méthodes et techniques que nous illustrons dans la figure 1.14.

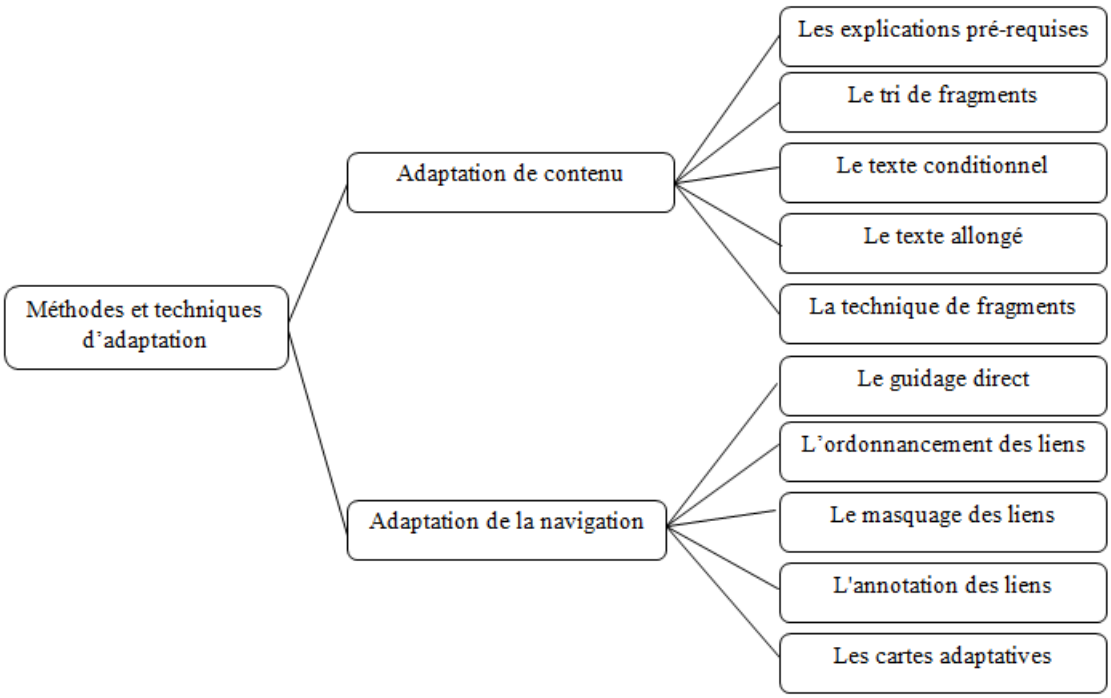

FIGURE 1.14 – Méthodes et techniques des hypermédias adaptatifs

Nous commençons par décrire les différentes méthodes d'adaptation des contenus :
◇ *Les explications pré-requises* : consistent à ajouter des explications introductives au début de chaque contenu présentant un sujet donné. Ces explications sont relatives aux prérequis nécessaires pour aborder le sujet en question.
◇ *Le tri de fragments* : consiste à trier les fragments composant un contenu selon leur pertinence par rapport au profil de l'utilisateur.
Pour chacune de ces méthodes d'adaptation du contenu, il existe des techniques pour les implémenter [83] :
◇ *Le texte conditionnel* : est une technique qui permet de proposer des informations supplémentaires. Ceci peut être réalisé par l'association de conditions aux informations supplémentaires. Ces conditions expriment généralement des critères requis pour y accéder. Par comparaison avec les valeurs affectées à l'utilisateur pour ces critères, le système décide de montrer ou non l'information supplémentaire.
◇ *Le texte allongé* : est une technique basée sur un principe d'expansion ou de réduction d'un texte dans un document hypermédia. Une partie du texte est associée à une information additionnelle, qu'il est possible de faire apparaître. Le système choisit ensuite de dévoiler

ou non l'information additionnelle en fonction des spécifications données par l'utilisateur.
◇ *La technique de fragments* : est basée sur le choix de contenus alternatifs. Ceci est réalisable de deux façons différentes : La première consiste à créer plusieurs versions d'un contenu. Au moment de l'affichage d'un contenu, le système sélectionne la version qui correspond le mieux à l'utilisateur ; La seconde technique adopte un principe similaire mais à un niveau de granularité plus fin, en créant différentes versions de fragments du contenu. Une sélection de la version adéquate est opérée pour construire une page à présenter à l'utilisateur.

De même, la navigation peut être adaptée suivant plusieurs techniques :
◇ *Le guidage direct* : est la technique la plus utilisée car elle est simple à mettre en œuvre. Ainsi, elle est basée sur l'ajout d'un lien hypertexte, nommé suivant, qui permet d'accéder à la page en adéquation avec les objectifs de l'utilisateur. Pour être réellement efficace, cette technique est utilisée avec au moins une des techniques décrites dans ce qui suit.
◇ *L'ordonnancement des liens* : consiste à afficher les liens hypertextes suivant un ordre, définissant l'intérêt ou l'importance des contenus cibles. Les liens les plus adéquats sont disposés de manière à réduire le nombre d'actions à effectuer pour atteindre l'information souhaitée.
◇ *Le masquage des liens* : consiste à limiter les possibilités de navigation en supprimant des liens hypertextes qui sont en inadéquation avec les objectifs de l'utilisateur.
◇ *L'annotation des liens* : part du principe que l'utilisateur doit savoir où il va arriver avant d'activer un lien. Il faut donc joindre à chaque lien des explications, textuelles ou graphiques, en fonction du profil de l'utilisateur.
◇ *Les cartes adaptatives* : permettent de présenter à l'utilisateur, l'organisation de l'hyperespace, à l'aide de liens, soit sous forme textuelle (une représentation hiérarchique de l'hyperespace), soit sous forme graphique. En effet, il est possible de présenter à l'utilisateur une organisation plus ou moins simplifiée, en fonction de son profil.

8.6. Les algorithmes évolutionnaires et l'adaptation

8.6.1. Généralités

Les algorithmes évolutionnaires [84], sont une famille d'algorithmes qui s'inspirent de la théorie de l'évolution pour résoudre des problèmes divers. Ils font ainsi évoluer un ensemble de solutions à un problème donné, dans l'optique de trouver les meilleurs résultats. Ce sont des algorithmes stochastiques [85], car ils utilisent itérativement des processus aléatoires.

Historiquement, trois grandes familles d'algorithmes ont été développées indépendamment, entre les années 1960 et 1970. Les premières méthodes sont les stratégies d'évolution, proposées par Rechenberg en 1965 [86], pour résoudre des problèmes d'optimisations continus. L'année suivante, Fogel, Owens et Walsh concoivent la programmation évolutionnaire comme une méthode d'intelligence artificielle pour la conception d'automates à états finis [87]. Enfin, en 1965, J.Holland propose les premiers algorithmes génétiques, pour l'optimisation combinatoire [88]. La publication en 1989 du livre de David Goldberg sur les algorithmes génétiques rendra ceux-ci particulièrement populaires [89].

Ces différentes approches ont beaucoup évolué et se sont rapprochées, pour finir par être regroupées sous le terme générique d'algorithmes évolutionnaires. Aujourd'hui, la littérature sur le sujet est extrêmement abondante, et ces algorithmes sont considérés comme un domaine de recherche très fertile. La grande majorité de ces méthodes sont utilisées pour résoudre des problèmes d'optimisation.

8.6.2. Les algorithmes génétiques

Les algorithmes génétiques sont inspirés de la théorie de l'évolution et des processus biologiques qui permettent à des organismes de s'adapter à leur environnement. Ils ont été inventés dans le milieu des années 60 par Holand [90]. Ces derniers sont nés des réflexions darwiniennes relatives à la théorie de l'évolution des espèces [91]. L'idée clé de cette théorie est que, sous les contraintes imposées par l'environnement, les espèces d'êtres vivants se sont progressivement auto-modifiées dans le but de s'adapter à leurs milieux naturels.

La mise en œuvre des algorithmes génétiques nécessite plusieurs étapes à détailler. La première est le codage d'un individu représenté par un chromosome. La seconde est le calcul de la performance. La troisième est de définir les opérateurs de reproduction.

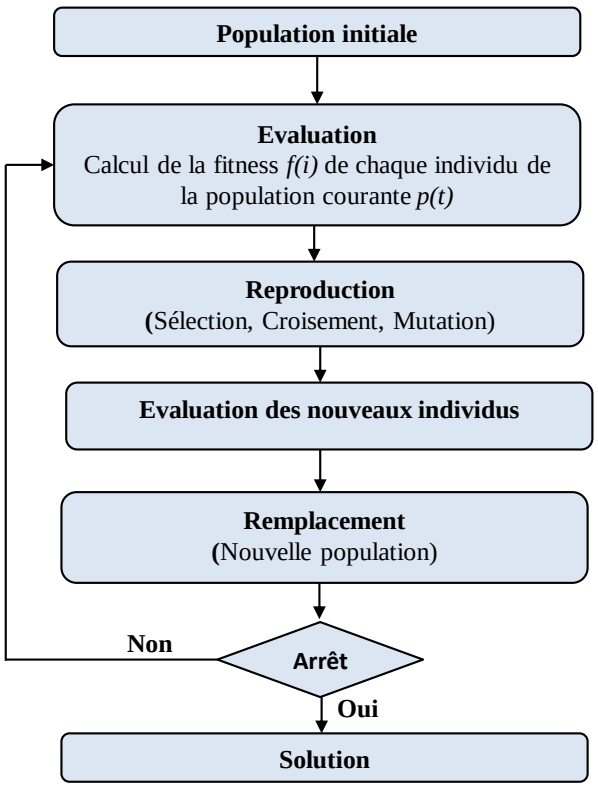

FIGURE 1.15 – Fonctionnement général d'un algorithme génétique

La figure 1.15 illustre le fonctionnement général d'un algorithme génétique. Il s'agit de

simuler l'évolution d'une population à laquelle nous appliquons différentes opérations :

- Individu et population : Le premier pas dans l'implantation des algorithmes génétiques est de créer une population d'individus initiaux. Chaque individu ou chromosome exprimé par un génotype, est constitué d'un ensemble fixe de gènes représentant chacune de ses caractéristiques. Le décodage d'un individu produit son phénotype. Un gène identifié par sa position appelée locus, peut prendre plusieurs valeurs dénommées allèles constituant ainsi l'alphabet de l'individu. Initialement, nous avons adopté particulièrement la représentation binaire, ce qui correspond à l'alphabet minimal {0,1} ; nous parlons alors de la version canonique des algorithmes génétiques.
Les codages réels sont désormais largement utilisés, notamment dans les domaines applicatifs pour l'optimisation de problèmes à variables réelles.

- Fonction d'adaptation : Les algorithmes génétiques s'inspirent de l'évolution des êtres vivants, en considérant que celle-ci tend à produire des organismes plus adaptés à leur environnement. Ils font évoluer un ensemble (une population) de solutions (les individus) ; par une mesure de leur adaptation ou capacité à résoudre le problème posé (la fitness) [92]. Les algorithmes sont conçus de façon à ce que plus la fitness d'un individu est élevée, plus il doit avoir de chances de transmettre son génotype au sein de la population.
La fonction d'adaptation ou fonction objectif, est un élément de réflexion fondamental lors de la modélisation d'un algorithme génétique car elle définit les contours de l'environnement dans lequel évolue la population d'individus. Cette fonction doit être capable de favoriser la sélection d'individus dans la direction de l'optimum qui est, à priori, inconnue.

- L'opérateur d'initialisation : Cet opérateur est utilisé pour générer la population initiale de l'algorithme génétique. La population initiale doit contenir des chromosomes qui soient bien répartis dans l'espace des solutions pour fournir à l'algorithme génétique un matériel génétique varié. La façon la plus simple est de générer aléatoirement les chromosomes.

- L'opérateur de sélection : La sélection tend à augmenter l'importance des bonnes solutions par rapport aux mauvaises. Les bonnes solutions sont supposées être les plus prometteuses pour la génération de descendants. Plusieurs méthodes existent pour sélectionner des individus destinés à la reproduction, les plus connues sont :
Premièrement, *la sélection par la roulette* [92] : La sélection des individus par le principe de la roulette s'inspire des roues de loterie. À chacun des individus de la population est associé un secteur d'une roue. L'angle du secteur étant proportionnel à la qualité (fitness) de l'individu qu'il représente. Vous tournez la roue et vous obtenez un individu. Les tirages des individus sont ainsi pondérés par leur qualité. Les meilleurs individus ont plus de chance d'être croisés et de participer à l'amélioration de la population.
Deuxièmement, *la sélection par tournoi* [93] : Le principe de la sélection par tournoi augmente les chances pour les individus de participer à l'amélioration de la population. Le principe est très rapide à implémenter. Un tournoi consiste en une rencontre entre plusieurs individus pris au hasard dans la population. Le vainqueur du tournoi est l'individu de meilleure qualité.

- L'opérateur de croisement (crossover) : Le croisement est mis en place pour que les nouveaux chromosomes gardent la meilleure partie des chromosomes anciens. Néanmoins, il est important qu'une partie de la population survive à la nouvelle génération. À partir de deux individus, nous obtenons deux nouveaux individus (enfants) qui héritent de certaines caractéristiques de leurs parents. Pour les algorithmes génétiques, nous distinguons deux types de croisement :
Le croisement à un point : C'est l'opérateur de croisement le plus simple. Il consiste à sélectionner un point de coupure, puis à subdiviser le génotype de chacun des parents. Ainsi, la figure 1.16 présente le cas où les parents A et B produisent alors les enfants A' et B'.

FIGURE 1.16 – Le croisement à un point

Les croisements multipoints : Ils reprennent le mécanisme de la méthode de croisement à un point en généralisant l'échange à 3 ou 4 sous chaines (figure 1.17).

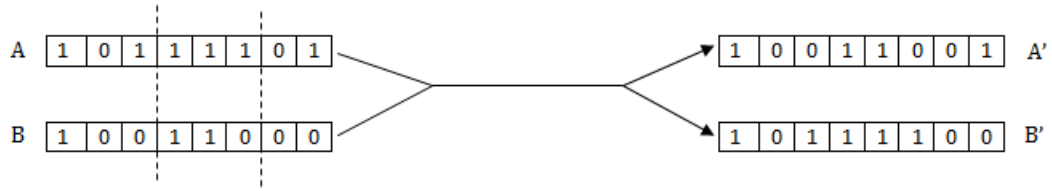

FIGURE 1.17 – Le croisement multipoints

- L'opérateur de mutation : La mutation sert à éviter une convergence prématurée de l'algorithme. En effet, un gène peut au sein d'un chromosome être substitué à un autre. Ainsi, la mutation permet d'introduire une certaine information dans la population en définissant un taux de mutation lors des changements de population qui est généralement compris entre 0,001 et 0,01. Il est nécessaire de choisir pour ce taux une valeur relativement faible pour conserver le principe de sélection et d'évolution (figure 1.18).

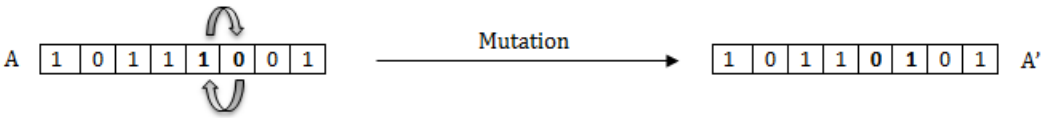

FIGURE 1.18 – L'opérateur de mutation

- Remplacement : Cette dernière étape du processus itératif consiste en l'incorporation des nouvelles solutions dans la population courante. Les nouvelles solutions sont ajoutées à la

population courante en remplaçant (total ou partiel) des anciennes solutions. Généralement, les meilleures solutions remplacent les plus mauvaises ; il en résulte une amélioration de la population.

En conséquence, il faut dire que à chaque étape de l'algorithme est associé un *opérateur*, qui décrit la façon de manipuler les individus. Après avoir initialisé une première population d'individus, nous itérons un nombre fini de fois, jusqu'à atteindre un critère d'arrêt (par exemple un nombre maximum de générations). La première étape de sélection permet de séparer les individus qui participeront à la reproduction de ceux qui n'y participeront pas. Les individus sélectionnés (les parents) se reproduisent (on dit aussi que l'on effectue des croisements), donnant un ensemble d'≪ enfants ≫ partageant une partie des caractéristiques de leurs ascendants. Ces enfants subissent alors une étape de mutation, qui modifie aléatoirement leur génotype. Les nouveaux individus sont alors évalués (on met à jour leur valeur en faisant appel à la fonction de fitness). Enfin, nous choisissons un nombre d'individus déterminé parmi l'ensemble $\{parents, enfants\}$, pour former la génération suivante.

Données : n, T
Résultat : g, génome ayant la meilleure fitness à la génération T
1 Créer une population P_0 de n génomes aléatoires : P_0=g_1^0, ... , g_n^0;
2 **pour** *(t=1 ;... ; T)* **faire**
3 | calculer la fitness f_i de chaque génomes aléatoires g_i ;
4 | sélectionner avec remplacement (i.e., un génome peut être séléctionné plusieurs fois) n génomes dans P_t, ces individus sont $g_1^{'}$, ... , $g_n^{'}$;
5 | Chaque paire ($g_1^{'}$,$g_2^{'}$),($g_3^{'}$,$g_4^{'}$), ..., ($g_{n-1}^{'}$,$g_n^{'}$) se reproduit pour donner les paires (g_1^t,g_2^t),(g_3^t,g_4^t), ..., (g_{n-1}^t,g_n^t) ;
6 | P_0 :=g_1^t, ... , g_n^t;
7 **fin**
8 Renvoyer l'individu ayant la meilleure fitness dans P_T

Algorithme 1 : Pseudo-code d'un algorithme génétique

Un algorithme génétique va faire évoluer une population dans le but d'en améliorer les individus. Le déroulement d'un algorithme génétique peut être résumé aux opérations suivantes :
1) Initialiser la population initiale P.
2) Evaluer P.
3) TantQue (Pas Convergence) faire :
- P' = Sélection des Parents dans P.
- P' = Appliquer Opérateur de Croisement sur P'.
- P' = Appliquer Opérateur de Mutation sur P'.
- P = Remplacer les Anciens de P par leurs Descendants de P'.
- Evaluer P.

FinTantQue

Le critère de convergence peut être de nature diverse, par exemple :
◇ Un taux minimum qu'on désire atteindre d'adaptation de la population au problème,
◇ Un certain temps de calcul à ne pas dépasser,
◇ Une combinaison de ces deux points.

Finalement, les algorithmes génétiques sont utilisés dans différents domaines d'applications. D'ailleurs, il faut dire qu'ils fournissent d'excellentes performances à faibles coûts. En effet, les algorithmes génétiques sont utilisés pour effectuer des optimisations sur des problèmes complexes afin d'obtenir une solution proche de l'optimal.

Conclusion

Dans ce chapitre, nous avons présenté l'état de l'art sur les systèmes e-learning dont l'objectif est l'utilisation des technologies de l'information et de la communication pour faciliter et améliorer la qualité de l'apprentissage humain. Nous avons aussi expliqué l'importance du tutorat en ligne lors d'une activité d'apprentissage. En outre, nous avons mis en évidence le rôle du tuteur lors du travail collaboratif en ligne. Au cours de ce dernier, le tuteur favorise les relations humaines et les contacts entre apprenants pour rompre avec l'isolement, le travail solitaire et l'absence d'échanges sur ce qui tourne autour des apprentissages.

Cependant, assister l'apprenant dans sa formation est l'objectif majeur de e-tutorat, c'est pourquoi nous avons présenté une étude détaillée des différents indicateurs d'analyse automatique des interactions dont le but est le suivi des différentes activités des apprenants et des groupes d'apprenants. Un bref état de l'art de e-learning adaptatif a été aussi exposé pour avoir une idée claire sur l'actualité de la recherche dans ce domaine ainsi que les méthodes et les techniques utilisées pour adapter un contenu pédagogique au profil d'un apprenant. Nous avons aussi introduit le principe des algorithmes génétiques que nous allons utiliser dans notre approche d'adaptation.

En résumé, le e-learning a connu une évolution croissante au cours de ces dernières années, les travaux de recherches en cours tentent de résoudre les problématiques rencontrés en utilisant des normes qui relèvent du domaine du web sémantique ou des algorithmes de l'intelligence artificielle.

Chapitre 2

Web sémantique et E-learning

"Le sens commun n'est pas si commun qu'on le pense"

- Proverbe latin,

Introduction

Le web sémantique est une extension du web qui vise à enrichir les ressources disponibles sur internet avec des descriptions sémantiques [94]. Son objectif est de rendre les informations du web actuel compréhensibles par les machines afin que les utilisateurs seront déchargés d'une bonne partie de leurs tâches de recherche, de construction et des combinaisons de résultat. Ces tâches seront accordées à des machines ayant la capacité d'accéder aux contenus des ressources et d'effectuer des raisonnements.

Cependant, les applications du web sémantique sont en cours de développement dans toutes les disciplines, notamment le e-learning. Ainsi, le monde de l'apprentissage a connu un changement perpétuel grâce aux sources de connaissances qui se multiplient et grâce aux formes de communication qui se diversifient. D'où la nécessité de profiter des opportunités offertes par le web sémantique.

En outre, il faut souligner que le web sémantique est une technologie prometteuse pour l'implémentation de e-learning, vu qu'elle est en pleine expansion ces dernières années. L'apprentissage en ligne peut bénéficier de cette vue globale du web sémantique, où toutes les ressources sont disponibles et annotées pour être mieux découvertes et mieux indexées.

Dans ce qui suit, nous présentons d'abord un bref historique des générations du web. Nous enchainons ensuite avec les couches de l'architecture du web sémantique, puis nous discutons les avantages d'utilisation des ontologies pour la description du contenu pédagogique.

1. Historique des générations du web

Cette dernière décennie a connu une évolution considérable et gigantesque des ressources disponibles sur Internet. Une évolution marquée par la croissance permanente des données exploitées à travers les technologies web [95]. Pour mieux comprendre les enjeux et les différentes phases de cette évolution, nous présentons dans la suite une synthèse des différentes générations du web [96], qui devrait nous fournir quelques éléments clés de compréhension (figure 2.1).

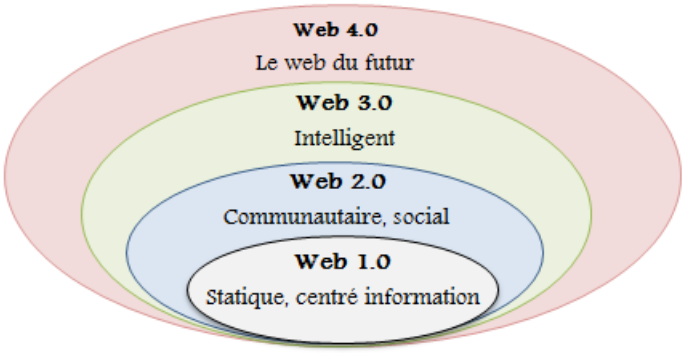

FIGURE 2.1 – Les générations du web 1.0 / 2.0 / 3.0/ 4.0

✓ **Le web 1.0** : c'est le web traditionnel, apparu au début des années 90, il est constitué de pages web liées entre elles par des hyperliens [95]. Le grand souci pour les développeurs de cette génération est comment faire pour afficher le site avec un bon design. Les langages de programmation à l'époque cherchent seulement à éditer le contenu et en aucun cas le sens que portent ce contenu. L'internaute a donc une attitude passive et peut seulement consulter les pages web sans pouvoir interagir avec d'autres acteurs.

✓ **Le web 2.0** : Appelé aussi web social [97], change totalement les perspectives. Il privilégie la dimension de partage et d'échange d'informations et de contenus (textes, vidéos, images ou autres). Il aperçoit l'émergence des réseaux sociaux, des forums, des blogs, des wikis, etc. À travers cette génération, le web se démocratise et se dynamise. L'avis du consommateur est sollicité en permanence. Toutefois, la reproduction de contenus énormes engendre une immense quantité de données difficile à vérifier.

✓ **Le web 3.0** : Nous parlons ainsi du web sémantique comme le qualifie certains auteurs. Les experts sont encore en train d'en débattre. Pour certains, ce sont des technologies qui adaptent en temps réel le contenu et la navigation d'un site internet en fonction du profil, des désirs et du comportement de l'internaute [98]. Il vise à organiser la masse des informations disponibles en fonction du contexte et des besoins de chaque utilisateur, en tenant compte de sa localisation, de ses préférences, etc. C'est un web qui tente de donner un sens aux données. Avec cette nouvelle génération, nous passons à un web plus intelligent et plus intuitif. Le web 3.0 s'approche plus de l'internaute et de ses préférences en visant sa satisfaction. Celle-ci

est concrétisée par la mise en valeur de la pertinence des services offerts. En fait, les utilisateurs se bénéficient d'un contenu web contextuel qui convient avec leurs profils.

✓ **Le web 4.0** : Dans la continuité du web 3.0, on entre dans un environnement d'interconnexion dont les machines sont connectées entre elles. L'environnement internet se transforme peu à peu en un véritable écosystème informationnel dans lequel nous serons complètement immergés [96]. Avec le web 4.0, internet sera toujours avec nous, nous serons informés en continu selon nos centres d'intérêts et des opportunités à saisir au cours de tous nos déplacements.

2. Le web sémantique

La vision du web sémantique a été présentée la première fois par Tim Berners-Lee comme suit : "*Le web sémantique est une extension du web actuel dans lequel les ordinateurs deviennent capables d'analyser toutes les données du web : le contenu, les liens, et les transactions entre personnes et ordinateurs. Avec le web sémantique, les mécanismes journaliers du commerce, de l'administration et de nos vies quotidiennes seront traités par des machines dialoguant avec d'autres machines*" [99].
L'expression web sémantique, attribuée le plus souvent à Tim Berners-Lee, regroupe un ensemble de programmes de recherche et de travaux variés. Leur objectif commun est de permettre aux machines de « comprendre » et de répondre aux demandes complexes de l'utilisateur en fonction du sens de ces demandes. La réalisation de cet objectif repose sur l'existence de données, accessibles par le web, structurées ou semi-structurées, représentées dans un formalisme autorisant des traitements automatisés allant au-delà des traitements liés à la présentation des données et mettant en œuvre des mécanismes d'inférence puissants.

L'idée principale est de parvenir à un web intelligent, où les informations ne seraient plus stockées mais comprises par les ordinateurs, pour apporter à l'utilisateur ce qu'il cherche vraiment. Le web sémantique permettra donc à rendre le contenu sémantique du web interprétable non seulement par l'homme, mais aussi par les machines. En convertissant les données écrites avec des mots clés en notions ou concepts.

Les termes *aller, apprendre, enseigner,etc* pour une machine sont des mots clés (suite de lettres) dénués de sens, éparpillés dans un document. Par exemple si vous tapez sur un moteur de recherche la requête suivante "*Citer tous les documents qui parle du Sahara marocain*". Le moteur de recherche traite cette phrase comme une suite de lettre et par conséquent va nous retourner tous les documents contenant les mots clés de la requête. Cependant, avec le web sémantique la machine cherche à comprendre le sens de la phrase saisie et à la base du sens de la phrase, elle retourne des documents qui contiennent la réponse, car elle a compris la question. En effet lorsqu'on transforme les mots en concepts, la machine devient capable de comprendre leurs sens dans leurs contextes.

En résumé l'enjeu cherché avec le web sémantique est de rendre la machine intelligente et capable d'analyser les données comme le ferait un cerveau humain pour traiter le sens

des informations.
L'évolution des travaux de recherche réalisés dans le cadre du web sémantique est marquée par différents niveaux de complexité. Ces derniers reposent sur une architecture en couches qui sera l'objet de la section suivant.

3. Le modèle en couche du web sémantique

L'architecture du web sémantique se compose d'un ensemble de couches, généralement représentées sous la forme d'une pyramide. Chaque niveau repose sur les résultats définis au niveau inférieur, c'est-à-dire que chaque niveau est progressivement plus spécialisé et plus complexe que son niveau précédent. D'autre part, tout niveau est indépendant des niveaux supérieurs afin qu'il puisse être développé et rendu opérationnel de manière autonome par rapport aux développements des niveaux supérieurs. Cette pyramide des langages illustrée dans la figure 2.2, a été initialement présentée par Tim Berners-Lee en 2001 [99].

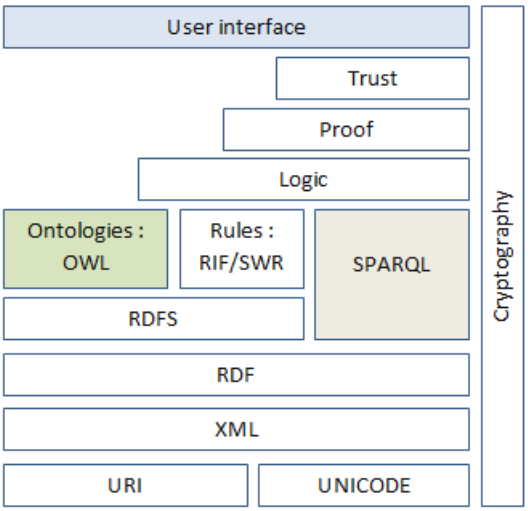

FIGURE 2.2 – Architecture en couches du web sémantique

Le schéma de la figure 2.2 récapitule les standardisations faite par le W3C [1] au sujet du web sémantique. Dans ce qui suit, nous présenterons plus en détail les différents niveaux de cette architecture :

3.1. URI (Uniform Ressource Identifier)

Un URI (Uniform Resource Identifier) veut dire en français identificateur uniforme de ressources [100]. D'une façon générale, le terme identificateur désigne une clé capable de

1. Le World Wide Web Consortium, abrégé par le sigle W3C, est un organisme de standardisation à but non lucratif, fondé en octobre 1994 chargé de promouvoir la compatibilité des technologies du World Wide Web

référencer un objet ayant une identité d'une manière unique. Dans le cas du web sémantique, l'URI est une séquence de caractères avec une syntaxe restreinte, qui permet d'identifier toute ressource utilisée dans le cadre d'une application web sémantique et dont la syntaxe respecte une norme d'internet mise en place par le W3C.
Par ailleurs, il est à noter que les données sont toujours encodées avec le jeu de caractères Unicode pour un maximum d'interopérabilité. C'est pourquoi cet élément figure dans cette couche de bas niveau, au même titre que l'URI.

3.2. XML (eXtensible Markup Langage)

XML (eXtensible Markup Language) est un métalangage proposé par le W3C permettant de représenter un document textuel de manière arborescente en utilisant un système de balisage [101]. Il a été élaboré pour faciliter l'échange, le partage et la publication des données à travers le web. Les langages proposés pour le web sémantique sont exprimés en XML. Ce dernier permet de structurer un document en définissant des balises selon les besoins des développeurs. Le choix de ces balises ne tient pas compte de la signification de la structure et des systèmes informatiques chargés de l'exploiter. Le langage XML est un langage de structuration et non de représentation de données, pour cela le W3C a proposé le langage XSL (eXtensible StyleSheet Language) [102] pour effectuer la représentation des données des documents XML.

3.3. Schéma XML

Avant de commencer à organiser les informations dans un document XML, il est impératif de définir la structure de ce dernier, afin de permettre notamment de vérifier sa validité. Le Schéma XML a été publié comme recommandation par le W3C en mai 2001. C'est un langage de description de format de document XML permettant de définir la structure et le type de contenu d'un document XML [101]. Il permet notamment de vérifier la validité de ce document.
Il est également possible, après une validation, de savoir avec quelle règle une information particulière a été testée : il s'agit du jeu de validation post-schema, ou PSVI (post-schema-validation infoset) [103].

3.4. RDF (Resource Description Framework)

RDF (Resource Description Framework) est un modèle de graphe destiné à décrire de façon formelle les ressources web et leurs métadonnées, de façon à permettre le traitement automatique de telles descriptions [104]. Développé par le W3C, RDF est le langage de base du web sémantique.
Un document structuré en RDF est un ensemble de triplets. Comme la montre la figure 2.3, un triplet RDF est une association (*sujet, prédicat, objet*) :

- Le *sujet* représente la ressource à décrire ;
- Le *prédicat* représente un type de propriété applicable à cette ressource ;

- L' *objet* représente une donnée ou une autre ressource : c'est la valeur de la propriété.

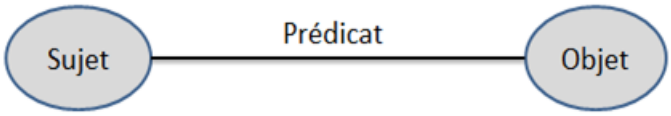

FIGURE 2.3 – Le triplet RDF

Le sujet et l'objet sont des ressources qui peuvent être identifiés par des URIs. Le prédicat est nécessairement identifié par une URI.
Les documents RDF peuvent être écrits en différentes syntaxes, y compris en XML. Il est possible d'avoir recours à d'autres syntaxes pour exprimer les triplets comme N-Triples, Turtle et N3 [105].
Une déclaration RDF est souvent schématisée par un graphe RDF, comme le montre la figure 2.4, où les sujets et les objets sont représentés par des ellipses, qu'on appelle nœuds, et le prédicat par une flèche qui va du sujet vers l'objet. Les rectangles, quant à eux, représentent des instances d'une ressource.

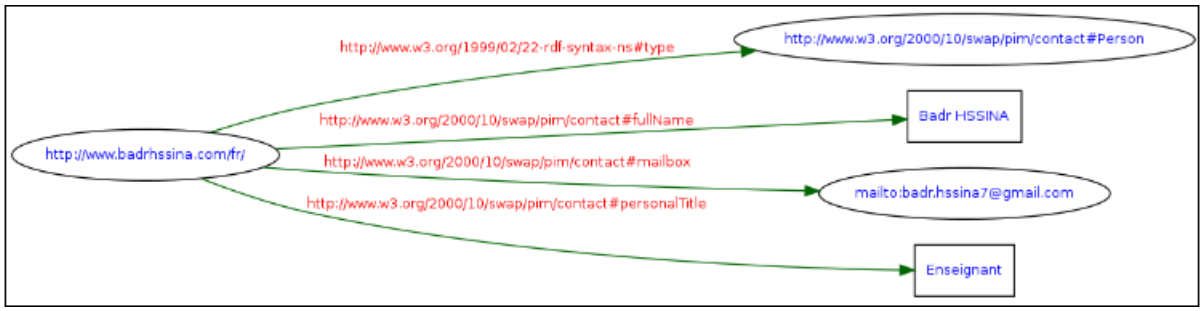

FIGURE 2.4 – Représentation schématique d'un graphe RDF

La figure 2.5 représente le code source de la description XML/RDF du graphe de la figure 2.4.

```
1: <?xml version="1.0" encoding="ISO-8859-1"?>
2: <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3:    xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
4:    <contact:Person rdf:about="http://www.badrhssina.com/fr/">
5:      <contact:fullName>Badr HSSINA</contact:fullName>
6:      <contact:mailbox rdf:resource="mailto:badr.hssina7@gmail.com"/>
7:      <contact:personalTitle
xml:lang="fr">Enseignant</contact:personalTitle>
8:    </contact:Person>
9: </rdf:RDF>
```

FIGURE 2.5 – Extrait du code de la description XML/RDF

3.5. RDF-Schéma

RDFS (Resource Description Framework Schema) est un langage extensible de représentation des connaissances [106]. Il appartient à la famille des langages du web sémantique publiés par le W3C. Un schéma est un vocabulaire de base pour décrire les déclarations RDF, au même titre que le schéma XML pour le langage XML. Il ajoute à RDF la possibilité de définir des hiérarchies de classes et de définir les genres et les propriétés des ressources, d'assigner des contraintes spécifiques sur la nature des documents et de fournir des informations sur l'interprétation des déclarations RDF. Les schémas RDF permettent donc de garantir qu'un document RDF est sémantiquement consistant.

3.6. RDF-attribut

RDFa (Resource Description Framework in attributs) est une recommandation du W3C définissant une syntaxe permettant d'ajouter des données structurées dans une page HTML ou n'importe quel document XML [107]. Ainsi formellement décrites, les données peuvent alors faire l'objet de traitements automatisés complexes, via des outils adaptés. Le code RDFa est invisible pour l'internaute et n'affecte pas ce qui est affiché.
RDFa est un ensemble d'éléments et d'attributs. Les données décrites en RDFa peuvent donc être facilement transformées en données RDF. À ce titre, RDFa est une technique permettant de mettre en œuvre le web sémantique.

3.7. OWL (Ontology Web Language)

OWL (Ontology Web Language) est un langage de représentation des connaissances construit sur le modèle de données de RDF [108]. Il fournit les moyens pour définir des ontologies web structurées. OWL a été recommandé par le W3C afin d'enrichir RDFS en définissant un vocabulaire plus complet pour la description des ontologies complexes. Il est enrichi par rapport à RDFS en lui ajoutant de nouvelles notions telles que : l'équivalence des classes, l'équivalence des relations, la symétrie et la transitivité des relations, la cardinalité, etc.

OWL permet, grâce à sa sémantique formelle basée sur une fondation logique largement étudiée, de définir des associations plus complexes des ressources ainsi que les propriétés de leurs classes respectives. Comme la montre la figure 2.6, OWL définit trois sous-langages, du moins expressif au plus expressif : OWL-Lite [109], OWL-DL [110] et OWL-Full [111].

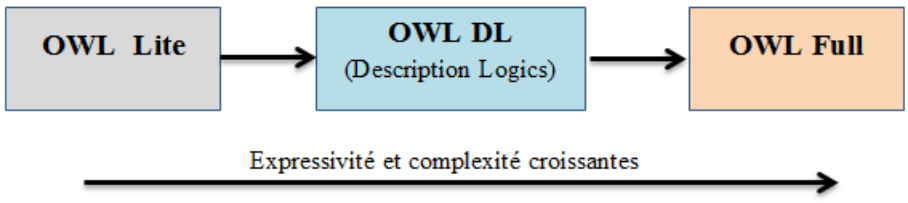

FIGURE 2.6 – Les sous langages d'OWL du moins expressive au plus expressive

OWL-Lite : Est la version la plus simple du langage OWL. Il correspond à la logique de descriptions. Sa simplicité lui permet d'avoir une théorie de complexité faible, et de garantir que les questions qu'on peut poser à un moteur d'inférence sur une base de données travaillant avec ce standard ont toujours une réponse et que cette réponse est calculable en un laps de temps réduit.
OWL-DL : Sigle pour (Ontology Web Language Description Logics) supporte les utilisateurs qui demandent un maximum d'expressivité tout en maintenant la complétude (garantie de calcul de toutes les conclusions) et la décidabilité (tous les calculs doivent finir en un temps fini). OWL-DL contient tous les constructeurs du langage OWL mais ils sont utilisables avec des restrictions.
OWL-Full : Plus flexible que OWL-DL destiné aux utilisateurs qui demandent un maximum d'expressivité avec la liberté syntaxique de RDF sans aucune garantie de calculs. Par exemple, une classe peut être traitée comme une collection d'individus et en même temps peut être vue comme un seul individu. OWL Full permet aussi de préciser d'avantage le sens du vocabulaire prédéfini.

3.8. Le langage de requête SPARQL

SPARQL (Protocol And Rdf Query Language) [112] est un langage de requête et un protocole qui permet de rechercher, d'ajouter, de modifier et de supprimer des données RDF disponibles à travers internet. SPARQL est l'équivalent de SQL car comme en SQL, on accède aux données d'une base de données via ce langage de requête alors qu'avec SPARQL, on accède aux données du graphe RDF. Cela signifie qu'en théorie, on pourrait accéder à toutes les données du web avec ce standard. L'ambition du W3C est d'offrir une interopérabilité non seulement aux niveaux des services, comme avec les services web, mais aussi aux niveaux des données structurées ou non structurées, qui sont disponibles à travers le web.
Dans la section suivante, nous présentons une description de la forme des requêtes SPARQL (figure 2.7)[112] :

PREFIX : indique l'adresse (espace de noms) d'un schéma pouvant être exploité ensuite dans la construction de la requête.
SELECT... [FROM]... retourne les ressources qui sont associées aux variables liées dans la clause :
WHERE : engendre un nouveau graphe qui complète le graphe interrogé.
UNION : graphes alternatifs (correspond à au moins un des graphes précisés).
FILTER : rajouter des conditions devant être satisfaites.
DESCRIBE : retourne une description des ressources satisfaisant la requête.
OPTIONAL : pattern(s) de graphe optionnel(s).
ASK : évalue si la requête va retourner un ensemble de ressources ou bien l'ensemble vide.

Une requête SPARQL se résume généralement par : "Je veux ces éléments d'information à partir du sous-ensemble des données qui répond à ces conditions."

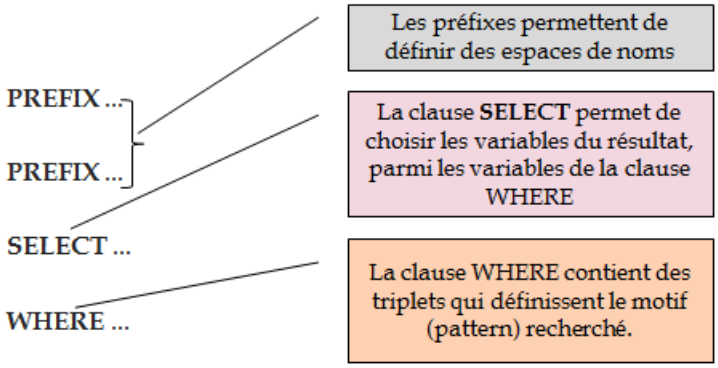

FIGURE 2.7 – La structure d'une requête SPARQL usuelle

Ce standard a été créé par le groupe de travail DAWG (RDF Data Access Working Group) du W3C. SPARQL est considéré comme l'une des technologies clés du web sémantique.

3.9. Les moteurs d'inférence

Une inférence (verbe *inférer* qui signifie *déduire*) est un processus de raisonnement qui s'appuie sur des connaissances acquises, et qui s'articule autour de règles fondamentales pour permettre d'obtenir de nouvelles informations [113].
Un moteur d'inférence permet aux systèmes experts de conduire des raisonnements logiques et de dériver des conclusions à partir d'une base de faits et d'une base de connaissances. Le moteur d'inférence va rajouter de nouvelle informations à partir d'informations existantes. C'est lui qui vient donner tout son sens au reste de la pyramide des technologies du web sémantique (URI, XML, RDF, OWL), en interprétant une base de connaissance.

Actuellement, plusieurs moteurs d'inférences gratuits ou commerciaux tels que Racer [114], Pellet [115], Fact [116], Fact++ [117], Surnia [118] et F-OWL [119] existent. La plupart de ces moteurs sont conçus pour raisonner sur les logiques de description, mais acceptent en entrée des fichiers OWL. Certains moteurs d'inférence ne peuvent raisonner qu'au niveau terminologique (c'est-à-dire au niveau des concepts et des propriétés) alors que des moteurs comme Pellet et Racer permettent de raisonner aussi sur les instances de concepts.

3.10. Les règles : RIF

Le RIF (Rule Interchange Format) [120] est une recommandation du W3C. Il existe différents langages de règles, cette couche a pour objectif de normaliser la représentation des règles RDF. Elle comporte deux langages de règles : SWRL (Semantic Web Rule Langage) [121] et RIF. SWRL est une extension d'OWL qui normalise la représentation des règles au format RDF. RIF ne repose pas directement sur RDF mais sur XML et il permet plutôt de faciliter l'utilisation et l'échange de règles entre les formats déjà existants.

3.11. La couche logique

En général, nous utilisons la couche logique pour exprimer les règles d'inférences. Une large variété de logiques a été conçue jusqu'à présent. Étant le formalisme le mieux apprécié dans la représentation de la connaissance, la logique descriptive est celle qui est, généralement, la plus adoptée pour la représentation des règles d'inférences. La logique descriptive est définie comme étant « *une famille de formalismes de représentation de la connaissance basée sur la logique. Elle est conçue pour représenter et raisonner sur la connaissance d'un domaine d'application d'une manière structurée et bien comprise. Elle dérive des réseaux sémantiques* » [122]. En effet, La logique de description se rapporte à la description de concepts utilisés pour spécifier les propriétés des objets et des individus qui se trouvent dans un domaine. La plupart des logiques de description divisent la connaissance en deux parties :

- Les informations terminologiques : définition des notions basiques ou dérivées et de la façon dont elles sont reliées entre elles. Ces informations sont *génériques* ou *globales*, vraies dans tous les modèles et pour tous les individus.
- Les informations sur les individus : ces informations sont *spécifiques* ou *locales*, vraies pour certains individus particuliers.

Toutes les informations connues sont alors modélisées comme un couple <T,A> où T est un ensemble de formules relatives aux informations terminologiques (la T-Box) et A est un ensemble de formules relatives aux informations sur les assertions ou les individus (la A-Box).
Une autre manière de voir la séparation entre ces informations est d'associer la T-Box aux règles qui régissent notre monde (par exemple la physique, la chimie, la biologie, etc.), et d'associer les individus de notre monde à la A-Box (par exemple Philip, Marie, un chien, etc.). La figure 2.7 illustre l'architecture des systèmes de représentation de connaissances selon Franz Baader [123].

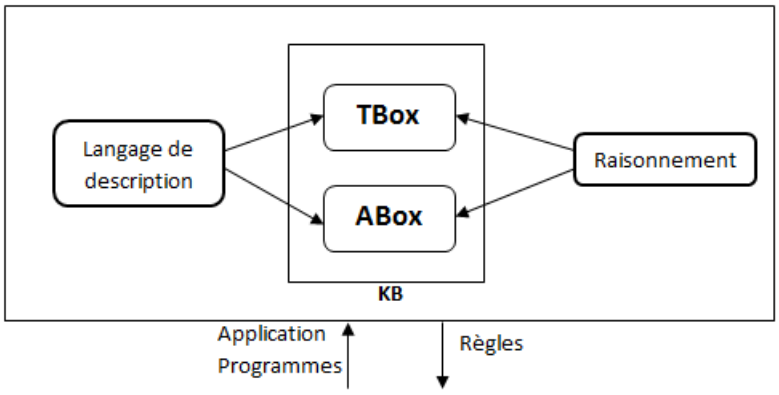

FIGURE 2.8 – Architecture des systèmes de représentation de connaissances basés sur la logique de description

Une base de connaissances KB (Knowledge Base) basée sur la logique de description comprend deux composants, la T-Box et la A-Box. Le vocabulaire consiste en des concepts, qui dénotent des ensembles d'individus, et des rôles qui dénotent des relations binaires entre

des individus.

3.12. La couche preuve

La couche *preuve* a pour but de prouver la pertinence de l'information retournée par les couches de plus bas niveau et des déductions obtenues à partir des inférences. Une des façons de le faire est de garder trace des sources d'information et des raisonnements effectués [124]. Malheureusement, à l'heure où nous rédigeons cette thèse, il n'existe encore aucun langage de preuves standardisé par le W3C. Un langage de preuve constitue un moyen simple pour prouver si une déclaration est juste ou pas. Une instance de ce dernier consiste en général en une liste de toutes les étapes d'inférence par lesquelles a transité l'information en question [125].

3.13. La couche confiance et cryptographie

Le web est un environnement très ouvert et dynamique. De ce fait, toute personne est donc en mesure d'éditer et de publier des informations de façon très simple. La couche confiance(Trust), dans l'architecture proposée par Tim Bernes-Lee, a pour objectif d'évaluer la fiabilité de l'information et des raisonnements [126].
La couche cryptographie repose sur les signatures numériques, le cryptage des données et sur la fiabilité des sources d'information [127]. Le web ne pourra atteindre son plein potentiel que si les utilisateurs aient confiance dans les transactions et la qualité de l'information fournie.

3.14. La couche utilisateur

C'est la dernière couche de la pyramide des couches du web sémantique. Elle permet à l'utilisateur d'exploiter les applications utilisants les technologies du web sémantique. Elle n'est pas spécialement placée au sommet, mais plutôt au-dessus de la dernière couche développée et elle évolue en même temps que le reste des technologies.

L'apport du web sémantique est d'une importance capitale pour la géstion des ressources web. Il est perçu depuis son apparition comme la technologie du futur d'après Berners-Lee. Cela explique d'ailleurs l'intérêt majeur que manifeste le W3C à son égard, notamment à travers les nombreuses recherches et publications effectuées dans ce domaine.

4. Notion d'ontologie

La notion de l'ontologie est emprunté de la philosophie où il fait référence à la science qui « *étudie l'être en tant qu'être* ». Avec l'émergence de l'ingénierie des connaissances et du web sémantique, ce terme a pris une tout autre tournure pour désigner la problématique de représentation et de manipulation des connaissances dans un système informatique [128]. En

effet, les ontologies permettent de décrire la structure et la sémantique des données. Quand des données sont présentées ou annotées par des ontologies, les logiciels peuvent mieux comprendre leurs sémantiques, ce qui facilite la localisation et l'intégration des données pour des objectifs divers.

4.1. Définition

De point de vue de la technologie, une ontologie est définie comme une "*spécification* ***formelle*** *et* ***explicite*** *d'une* ***conceptualisation partagée***" [129], sachant que :

- Le terme **formelle** se rapporte au fait que l'ontologie devrait être compréhensible par une machine ;
- Le terme **explicite** signifie que le type de concepts utilisés, et les contraintes sur leur utilisation sont explicitement définis ;
- Le terme **conceptualisation** signifie qu'un modèle abstrait des phénomènes est identifié par des concepts appropriés à ces phénomènes ;
- Le terme **partagée** reflète que l'ontologie devrait capturer la connaissance consensuelle admise par les communautés.

Une ontologie n'est en fin de compte qu'une modélisation du monde réel en concepts et relations entre ces concepts (figure 2.9). Elle est donc la manifestation d'une compréhension partagée d'un domaine de connaissance pour plus d'interopérabilité, de réutilisation et du partage [130].

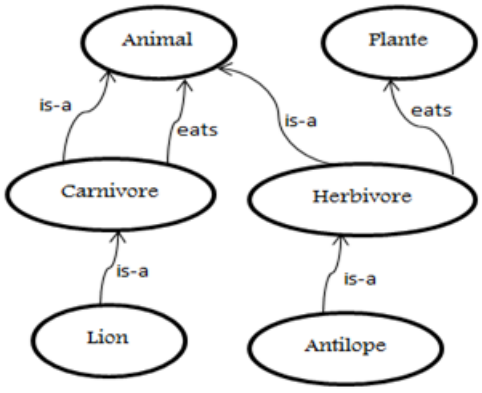

FIGURE 2.9 – Un simple exemple illustrant une partie d'une ontologie

Le but principal d'intégration d'ontologie dans les systèmes informatiques n'est pas seulement de servir d'argument déductif pour définir une réalité, mais de permettre une explication des termes et des significations pour définir une base consensuelle pour l'interopérabilité dans un domaine.
Les principales composantes d'une ontologie que nous pouvons distinguer sont donc l'objet de la section suivante.

4.2. Les composants d'une ontologie

D'après [129] et [130], une ontologie fournit le vocabulaire d'un domaine et définit le sens des termes et les relations qui les relient. La connaissance dans une ontologie est formalisée en utilisant cinq composantes : les concepts, les relations, les fonctions, les axiomes et les instances [131].

- **Concepts** : Un concept peut représenter un objet matériel, une notion, une idée ; les concepts sont aussi appelés termes ou classe, constituent les objets de base manipulés par les ontologies. Ils sont représentés dans le langage OWL [108] par *owl : Class*.
- **Relations** : Traduisent les interactions existant entre les concepts présents dans le domaine traité. Ces relations incluent la relation de spécialisation (subsomption), la relation de composition (méronymie), la relation d'instanciation, etc. Elles permettent de capturer, la structuration ainsi que l'interaction entre les concepts, ce qui permet de représenter une grande partie de la sémantique de l'ontologie. Elles sont représentées dans OWL par *owl : ObjectProperty*.
- **Fonctions** : sont des cas particuliers de relations dans lesquelles le n-ième élément de la relation est défini de manière unique à partir des n-1 éléments précédents.
- **Axiomes** : Permettent de modéliser des assertions toujours vraies, à propos des abstractions du domaine traduites par l'ontologie. Ils permettent de combiner des concepts, des relations et des fonctions pour définir des règles d'inférences et qui peuvent intervenir, par exemple, dans la déduction, la définition des concepts et des relations, ou alors pour restreindre les valeurs des propriétés ou les arguments d'une relation.
- **Instances** : ou individus constituent la définition extensionnelle de l'ontologie. Elles représentent des éléments singuliers véhiculant les connaissances à propos du domaine du problème.

4.3. Classification des ontologies

Selon Guarino [132], les ontologies sont classifiées en quatre catégories (figure 2.10). Cette classification se base sur le degré de généralité ou du niveau de dépendance d'une tâche :

- **Les ontologies de haut niveau** qui décrivent des concepts très généraux et fournissent des notions générales sous lesquelles tous les termes racines dans les ontologies existantes devraient être liés. Le principal problème est qu'il y a plusieurs ontologies de haut niveau et elles diffèrent sur les critères suivis pour classifier les concepts les plus généraux de la taxonomie.
- **Les ontologies de domaine** qui sont réutilisables dans un domaine spécifique donné. Ces ontologies fournissent les vocabulaires sur les concepts et leurs relations dans un domaine, sur les activités qui ont lieu dans ce domaine, et sur les théories et les principes élémentaires régissant ce domaine. Il y a une frontière claire entre les ontologies de domaine et les ontologies de haut niveau. Les concepts dans les ontologies de

domaine sont habituellement des spécialisations des concepts déjà définis dans les ontologies de haut niveau, et le même principe pourrait se produire avec les relations. Exemple : commerce, médecine, sciences de l'ingénieur (modèles mathématiques), entreprise, chimie, etc.

- **Les ontologies d'activités ou des tâches** qui décrivent le vocabulaire relié à une tâche générique ou à une activité (comme diagnostiquer, programmer, vendre, etc.) en spécialisant les termes dans les ontologies de haut niveau. Les ontologies de tâche fournissent un vocabulaire systématique des termes utilisés pour résoudre les problèmes liés aux tâches qui peuvent appartenir au même domaine. Ce sont des ontologies dépendantes des applications. Les ontologies de tâche de domaine (Domain-Task ontologies) sont des ontologies de tâches réutilisables dans un domaine donné.
- **Les ontologies d'applications** où les concepts dépendent à la fois d'un domaine et d'une activité particulière. Elles donnent les définitions des concepts et des relations appropriés appliquées pour spécifier un processus de raisonnement afin de réaliser une tâche donnée. Elles contiennent toutes les définitions nécessaires pour modéliser la connaissance requise pour une application particulière. Les ontologies d'application étendent et spécialisent souvent le vocabulaire des ontologies de domaine et de tâche pour une application donnée.

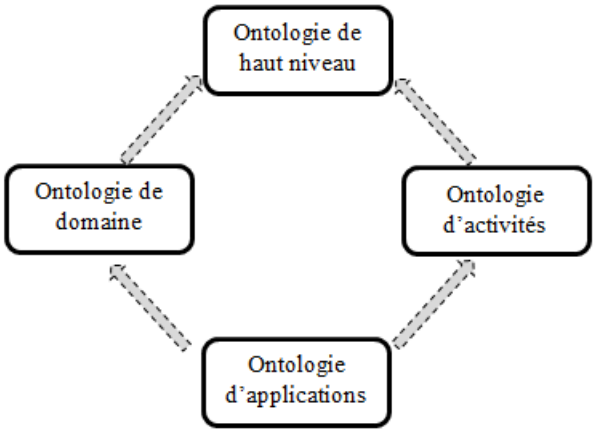

FIGURE 2.10 – Classification des ontologies

D'après [133], une ontologie est classifiée selon la richesse des structures utilisées. Donc, on peut distinguer différentes catégories d'ontologies :

- **Les ontologies terminologiques** qui sont utilisées pour spécifier les termes du vocabulaire d'un domaine de connaissances.
- **Les ontologies d'information** qui spécifient la structure/le schéma d'une base de données pour permettre le stockage d'informations.
- **Les ontologies qui modélisent de la connaissance** proposent des structures internes plus riches et qui sont davantage définies en fonction de leurs utilisations comme par exemple le partage d'informations.

- **Les ontologies génériques** qui sont utilisées pour représenter la connaissance commune (consensuelle) réutilisable dans les domaines. Ces ontologies incluent le vocabulaire lié aux objets, aux événements, au temps, à l'espace, à la causalité, au comportement, à la métrologie, etc.
- **Les ontologies de représentation** qui visent à expliciter les conceptualisations sous-jacentes aux formalismes de représentation des connaissances. Ces concepts des ontologies de représentation peuvent être utilisés dans les ontologies génériques ou les ontologies de domaine.

4.4. Le cycle de vie des ontologies

Étant donné que les ontologies sont destinées à être utilisées comme des composants logiciels dans des systèmes informatiques répondant à des objectifs opérationnels différents, leur développement doit s'appuyer sur les mêmes principes que ceux appliqués en génie logiciel. En particulier, elles doivent être considérées comme des objets techniques évolutifs et posséder un cycle de vie spécifique [134].

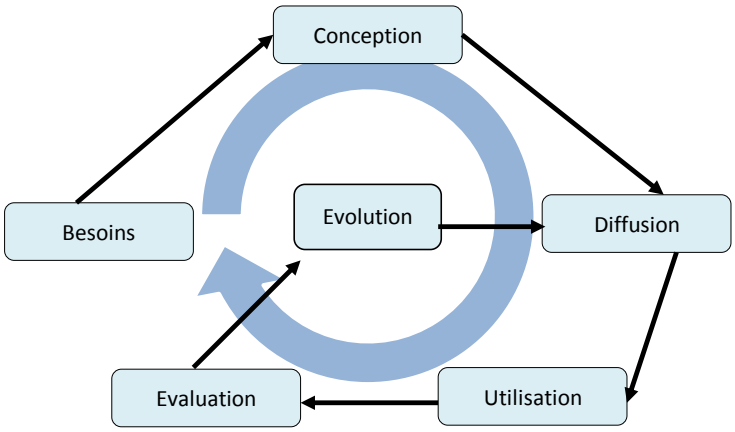

FIGURE 2.11 – Cycle de vie d'une ontologie

Ce cycle de vie (figure 2.11), comprend une étape initiale d'évaluation des besoins, une étape de construction, une étape de diffusion, et une étape d'utilisation. Après chaque utilisation significative, l'ontologie et les besoins sont réévalues et l'ontologie peut être étendue et, si nécessaire, en partie reconstruite [135]. Le cycle de vie par évolution de prototypes permet à l'otologiste de retourner de n'importe quel état à n'importe quel autre si une certaine définition manque ou est erronée. Ainsi, ce cycle de vie permet l'inclusion, le déplacement ou la modification de définitions n'importe quand durant le cycle de vie de l'ontologie. L'acquisition, la documentation et l'évaluation de connaissances sont des activités de support qui sont effectuées pendant la majorité de ces états.

4.5. Environnement et outils de modélisation

De nombreux outils de construction d'ontologies utilisant des formalismes variés et offrant différentes fonctionnalités ont été développés. Seuls les plus dominants sont cités içi :

ONTOLIGUA [136] : C'est un serveur d'édition d'ontologies au niveau symbolique, l'ONTOLINGUA utilise des classes, des relations (elles peuvent contenir des contraintes nécessaire et suffisantes), des fonctions, des objets (instances) et des axiomes pour décrire une ontologie.

ODE (Ontology Design Environment) [137] : développé au laboratoire d'Intelligence Artificielle de l'Université de Madrid, permet de construire des ontologies au niveau connaissance. La formalisation avec ODE s'effectue avec un langage de frames, tandis que l'opérationnalisation utilise des formalismes de type Ontolingua ou Flogic.

PROTEGE 2000 [138] : Interface modulaire permettant l'édition, la visualisation, le contrôle (vérification des contraintes) d'ontologies, l'extraction d'ontologies à partir de sources textuelles, et la fusion semi-automatique d'ontologies. Le modèle de connaissances sous-jacent à PROTEGE 2000 est issu du modèle des frames et contient des classes (concepts), des slots (propriétés) et des facettes (valeurs des propriétés et contraintes), ainsi que des instances des classes et des propriétés.

4.6. Le rapport entre ontologie et web sémantique

L'ontologie est un facteur clé qui facilite l'interopérabilité dans le web sémantique [129]. Les ontologies sont le noyau du web sémantique parce qu'elles permettent aux applications de communiquer en utilisant des termes partagés. L'ontologie facilite donc la communication en fournissant des notions précises qui peuvent être employées pour composer et échanger des messages (questions, réponses, etc.). Grâce au web sémantique, l'ontologie a trouvé un formalisme standard à l'échelle mondiale et s'intègre dans de plus en plus d'applications web, sans même que les utilisateurs ne le sachent. Cela se fait au profit des logiciels qui, à travers les ontologies et les descriptions qu'elles permettent, peuvent proposer de nouvelles fonctionnalités.

Cependant, il n'existe pas d'ontologie universelle partagée, adoptée par tous les utilisateurs d'un domaine donné. Les problématiques et les tentatives d'amélioration de l'interopérabilité du système comptent donc sur la réconciliation des différentes ontologies utilisées dans un domaine par des systèmes différents. Cette réconciliation est souvent réalisée par l'intégration manuelle ou semi-automatisée des ontologies. Elle consiste à identifier les liens de correspondance entre les ontologies, on parle alors de Mapping d'ontologies [131]. Les ontologies font partie intégrante des normes du W3C pour le web sémantique, dans lequel elles sont utilisées pour spécifier la norme des vocabulaires conceptuels et dans lequel l'échange de données entre différents systèmes a pour but de fournir :

- des services pour répondre aux requêtes ;
- de publier des bases de connaissances réutilisables ;

- d'offrir des services pour faciliter l'interopérabilité entre plusieurs systèmes hétérogènes et bases de données.

Le rôle clé des ontologies par rapport au web sémantique est de spécifier une représentation de modélisation des données à un niveau d'abstraction au-dessus des schémas d'une base de données classique(logique ou physique), afin que les données puissent être exportées, traduites, interrogées et unifiées pour tous les systèmes développés de manière indépendante.

5. Les ontologies et e-learning

L'ingénierie ontologique est devenue un thème propageant et d'actualité au sein des travaux de recherche menés dans le domaine de l'apprentissage en ligne [139]. Il y a de plus en plus de projets mettant en œuvre des ontologies traitant différents aspects de e-learning. Par exemple, les ontologies sont utilisées pour la recherche et l'indexation des différentes ressources pédagogiques sur le web [140] [141]. La création des documents d'apprentissage à l'aide de l'annotation ou les ontologies étaient le sujet de plusieurs travaux. En effet, le travail de Ricardo et Amorim [142] propose une ontologie pour faciliter la découverte et la réutilisation des objets d'apprentissage stockés dans des entrepôts. Il a défini les ontologies pour marquer la structure des objets d'apprentissage afin de permettre aux machines de les manipuler. Cependant, le travail de Suzanne Kabel [143] utilise l'indexation des objets d'apprentissage. Dans ce travail, en plus d'utiliser la norme LOM [28], les auteurs ont utilisé un vocabulaire d'indexation spécifiques et un vocabulaire d'indexation structurée pour les objets d'apprentissage plus élémentaires pour soutenir les tâches de l'indexation des objets d'apprentissage. Les auteurs de [144] ont proposé une approche basée sur l'utilisation de l'indexation des documents en utilisant l'ontologie de domaine pour améliorer la pertinence de la recherche d'information. Également, ils ont utilisé des liens sémantiques entre les documents ou fragments de documents afin de permettre la recherche de tous les documents pertinents.

Dans ce contexte, et dans le cadre de nos travaux de recherche nous avons conçu une ontologie pour évaluer la performance des apprenants sur une plate-forme e-learning en se basant sur les technologies du web sémantique [145]. En effet, superviser les activités des apprenants sur une plateforme e-learning est un défi majeur pour les enseignants et les tuteurs. Notre approche est basée sur des standards qui relèvent du domaine du web sémantique comme les ontologies, l'API JENA [146], et le langage de requête SPARQL dont l'objectif est d'aider un tuteur dans le suivi des activités réalisées par les apprenants. L'idée principale derrière notre approche est qu'une ontologie peut être non seulement utile comme un instrument d'apprentissage, mais peut également être utilisée pour évaluer les compétences des apprenants.

En résumé, les ontologies ont un rôle principal à tenir pour le partage, la réutilisation et la diffusion des connaissances sur une plateforme e-learning. Le contexte scientifique des travaux proposés est celui de l'étude des apports des ontologies pour la construction de modèles et d'outils à partir desquels pourront être développés différents services pour les acteurs d'un e-learning. L'objectif est de doter le système de capacités permettant d'adapter

les ressources et les services proposés à un apprenant en fonction de ses préférences d'apprentissage.

6. Conclusion

Dans ce chapitre nous avons mis l'accent sur le potentiel du web sémantique qui peut être traité comme une solution adéquate pour implémenter un système e-learning, du moment qu'il fournit tous les moyens pour le développement d'une ontologie d'apprentissage.

Dans le chapitre suivant, nous allons présenter une contribution à un environnement éducatif qui permet de calculer la similarité sémantique entre des documents textes. Le but de notre contribution est de recommander aux apprenants des documents pédagogiques similaires à leurs choix antérieurs. Ces documents sont indexés sémantiquement pour faciliter le processus de recommandation qui est basé sur le contenu. Cette approche que nous allons présenter soutient l'apprentissage personnalisé des apprenants hétérogènes.

Chapitre 3

Vers une recommandation sémantique des documents sur une plateforme e-learning

"C'est ce que nous pensons déjà connaître qui nous empêche souvent d'apprendre"

- Gaston Bachelard, *Philosophe français*

Introduction

La problématique élaborée dans ce chapitre se situe dans le cadre général des systèmes de recommandation. En effet, la recommandation sémantique des documents est un domaine de recherche prometteur, car elle garantit un accès rapide et ciblé à l'information. Nous nous intéressons particulièrement à la description du contenu des documents par des descripteurs sémantiques, plus précisément, l'indexation sémantique d'un corpus (collection de documents). En effet, la recommandation sémantique d'un document s'effectue par le calcul de la similarité sémantique entre un document et chaque document du corpus.

De plus, le filtrage des informations dans le domaine de e-learning est crucial car il permet de cibler d'une manière efficace les documents que les apprenants doivent apprendre tout au long de leur parcours pédagogique. Pour répondre à cette problématique, nous proposons un système novateur de recommandation sémantique des documents textuels qui s'appuie sur l'intégration d'une description sémantique des documents. Notre approche s'établit sur l'adéquation de la recommandation aux besoins des apprenants.

En somme, les développements réalisés dans le cadre de ce chapitre se composent de trois axes principaux. Nous présentons d'abord, les systèmes de recommandation en particulier la recommandation basée sur le contenu. Ensuite, nous introduisons la notion de l'indexation sémantique et le calcul de la similarité sémantique en se basant sur un corpus textuel. Enfin, nous terminons par les critères d'évaluation des systèmes de recommandation sémantique exploités dans un environnement d'apprentissage en ligne.

1. Les systèmes de recommandation

Les systèmes de recommandation sont une forme spécifique de filtrage de l'information visant à présenter les éléments d'information (films, musique, livres, actualités, images, pages web, etc) qui sont susceptibles d'intéresser l'utilisateur. Généralement, un système de recommandation permet de comparer le profil d'un utilisateur à certaines caractéristiques de référence, et cherche à prédire son besoin.

1.1. Introduction aux systèmes de recommandations

Un système de recommandation est "un système capable de fournir des recommandations personnalisées et de guider l'utilisateur vers des ressources intéressantes ou utiles au sein d'un espace de données important" [147]. Les systèmes de recommandation ont prouvé ces dernières années qu'ils sont un bon moyen pour faire face au problème de surcharge cognitive [148]. En effet, pour résoudre ce problème, un système de recommandation met en œuvre des items inconnus qui peuvent être pertinents pour les utilisateurs.

Ainsi, un système de recommandation se base sur des connaissances variées (profil de l'utilisateur, contexte de consommation, items disponibles, historique des transactions, feedbacks d'autres utilisateurs sur l'item, etc.). L'utilisateur peut alors parcourir les recommandations et peut fournir un feedback implicite ou explicite. Par exemple, dans une plateforme e-learning un feedback peut être des notes ou des avis que les apprenants peuvent attribuer aux contenus. Toutes les actions et les feedbacks des utilisateurs peuvent être enregistrés dans la base de données du système de recommandations, et utilisés ensuite pour générer de nouvelles recommandations.

Aujourd'hui, la raison pour laquelle les gens pourraient être intéressés à utiliser un système de recommandation est qu'ils ont tant d'éléments à choisir dans une période limitée de temps et ils ne peuvent pas évaluer toutes les items possibles. De manière générale, pour atteindre cet objectif, les utilisateurs doivent fournir leur propre profil et ils vont recevoir une image réduite et personnalisée de ces informations. À première vue, ils ressemblent aux moteurs de recherches d'informations. Cependant, la différence est que les moteurs de recherche permettent de renvoyer tous les éléments qui correspondent à la requête, classés par degré de pertinence. Alors que, l'objectif de la recommandation est de retourner un contenu personnalisé, intéressant et utile aux utilisateurs.
Dans les sections suivantes, nous décrivons brièvement ces différentes approches de recommandation.

1.2. Les approches de la recommandation

Il est possible de classer les systèmes de recommandation de différentes manières. La classification la plus connue est une classification selon trois approches : la recommandation à base de contenu, la recommandation à base d'utilisateurs et la recommandation hybride :

1.2.1. Recommandation à base de contenu

Nous parlons «d'approche basée sur le contenu» ou «content-based approach» [149] : Il s'agit de recommander des objets (ou contenus) en se basant sur les qualités et propriétés intrinsèques de l'objet lui-même et en les corrélant avec les préférences et les intérêts de l'utilisateur. Ce type de système va donc extraire un certain nombre de caractéristiques et d'attributs propres à un contenu, afin de pouvoir recommander à l'utilisateur des contenus additionnels possédant des propriétés similaires (figure 3.1). Cette approche permet de créer un profil pour chaque objet ou contenu, c'est-à-dire un ensemble d'attributs qui caractérisent l'objet.

La décision de recommandation ou non d'un document à un utilisateur peut se baser sur le contenu de celui-ci, c'est-à-dire sur une comparaison des thèmes abordés dans le document par rapport aux thèmes qui intéressent l'utilisateur. Pour décider la suggestion d'un document ou non, un système de recommandation peut se baser sur les mots-clés principaux du document et les comparer avec les mot-clés apparaissant dans d'autres documents que l'utilisateur a évalués positivement dans le passé.

Par exemple, dans le cas d'un site de vente de livre en ligne, nous allons se baser sur les caractéristiques du livre pour effectuer des recommandations, comme par exemple le sujet de l'ouvrage, son genre, son auteur, son éditeur, etc. Nous pourrions ainsi recommander le livre "*Les Misérables*", par exemple, à un utilisateur, si nous savons d'une part que ce livre est un roman de *Victor Hugo* et d'autre part que l'utilisateur aime les romans de *Victor Hugo*.

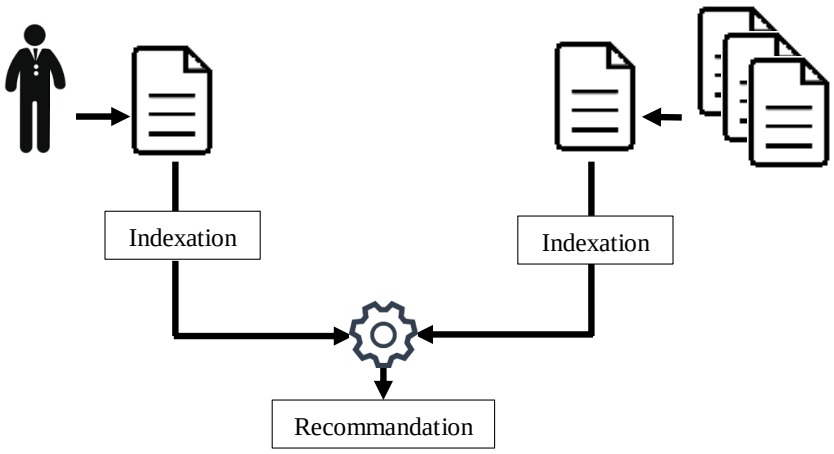

FIGURE 3.1 – Recommandation basée sur le contenu

Un système de recommandation pourra donc accomplir cette tâche seulement s'il a à sa disposition deux types d'information :
La description des caractéristiques du livre et un profil utilisateur qui décrit les intérêts antérieurs de celui-ci en termes de préférence de type de livre. La tâche de recommandation consiste donc à déterminer les livres qui correspondent le mieux aux préférences de l'utilisateur.

1.2.2. Recommandation à base d'utilisateurs

Nous parlons «d'approche sociale» ou «Collaborative Filtering» [150]. Il s'agit de recommander des choses sur la base du comportement passé des utilisateurs similaires, en effectuant une corrélation entre des utilisateurs ayant des préférences et intérêts similaires (figure 3.2). Dans ce type de recommandation, nous utilisons des méthodes qui collectent et analysent des données sur le comportement, les activités, les préférences des utilisateurs. Les algorithmes liés à ce genre de recommandation tentent de prédire ce que l'utilisateur préférera en cherchant des utilisateurs qui ont les mêmes comportements que l'utilisateur à qui l'on souhaite faire des recommandations. L'idée est de décider si une personne P1 a la même opinion (ou les même goûts) qu'une personne P2 sur un objet O1, alors la personne P1 a plus de chance d'avoir la même opinion que P2 sur un autre objet O2, plutôt que d'avoir la même opinion que quelqu'un choisi au hasard pour l'objet O2.

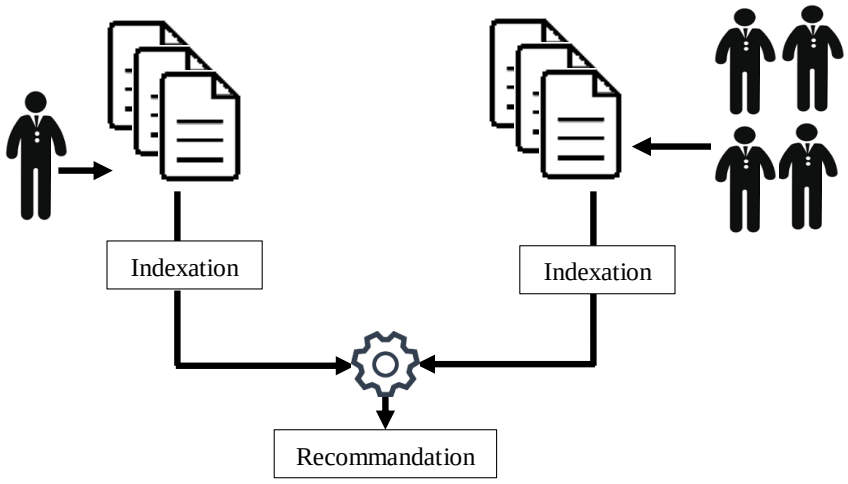

FIGURE 3.2 – Recommandation sociale

Le principe de base est donc de dire que si des utilisateurs ont partagés des mêmes intérêts dans le passé, il y a de fortes chances qu'ils partagent aussi les mêmes goûts dans le futur.

1.2.3. Recommandation hybride

La recommandation hybride [151] est une combinaison des deux approches ci-dessus. L'hybridation de ces deux techniques, afin de traiter les insuffisances de chaque technique et profiter de leurs points forts, a fait l'objet de plusieurs travaux de recherche [152] [153] [154] [155] [156]. Ce type de recommandation consiste à déterminer les items les plus proches des items appréciés par l'utilisateur en appliquant un filtrage sur le contenu, puis d'appliquer un filtrage collaboratif en se basant sur la qualité des items à partir des évaluations des utilisateurs. Le système Fab [157], compte parmi les premiers systèmes de recommandation

hybride, il combine le filtrage collaboratif et le filtrage basé sur le contenu afin de traiter le problème de sur-spécialisation et le problème du démarrage à froid pour les items. Dans ce système, pour qu'un item soit recommandé à l'utilisateur courant, deux critères doivent être satisfaits :
(i) Le contenu de l'item doit être similaire au contenu des items choisis par l'utilisateur, et (ii) l'item doit être apprécié par les voisins les plus proches de l'utilisateur courant [158].
La recommandation hybride présente un gain au niveau de la précision des recommandations par rapport à des approches purement collaboratives ou purement basées sur le contenu.

1.3. Les systèmes de recommandation pour e-learning

L'état de l'art montre un nombre important de systèmes de recommandation proposés dans le domaine de l'éducation. L'objectif, a toujours été d'élaborer et de mettre en place des outils susceptibles de rendre plus efficaces et pertinents les processus d'apprentissage. D'où l'intérêt des systèmes de recommandations qui peuvent jouer un rôle important, en proposant des ressources d'apprentissage (un item ou un ensemble d'items) pertinentes aux apprenants en fonction de leur contexte d'apprentissage. Une discussion sur les avantages et limites des différentes techniques appliquées dans ce contexte a été présentée par Drachsler dans [159].

Dans [160], Zaíane propose un système de recommandation dans un contexte e-learning basé sur un agent logiciel qui tente «intelligemment» de recommander des items à un apprenant sur la base des items des apprenants précédents. Dans le contexte des systèmes de recommandation e-learning, Bobadilla [161] affirme que les utilisateurs avec une plus grande connaissance (par exemple, ceux qui ont obtenu de meilleurs résultats dans les différents tests) ont plus de poids dans les listes des recommandations que les utilisateurs avec moins de connaissances. Pour conclure, les systèmes de recommandation offrent une approche prometteuse pour faciliter les tâches à la fois d'apprentissage et d'enseignement [162].

Par ailleurs, dans le cadre de cette thèse, nous nous sommes intéressés à la problématique de la recommandation par une contribution intitulée « Une approche hybride de calcul de similarité sémantique pour une recommandation basée sur le contenu des documents texte sur une plateforme e-learning » [163]. Le détail de cette contribution sera présenté profondément dans le chapitre 4.

Nous proposons dans la suite, les fondements principals de notre approche de recommandation basée sur le contenu. Les items qui vont être recommandés par ce système sont des documents texte et les utilisateurs du système sont des apprenants inscrits sur une plateforme e-learning. L'idée de cette approche est le calcul de la similarité sémantique entre les documents d'un corpus textuel.

2. Indexation d'un corpus textuel

L'évolution d'internet et des nouvelles technologies de stockage, de transfert et de traitement de l'information ont causé une forte augmentation du volume de documents numériques. Cette augmentation est accompagnée par une croissance des besoins des utilisateurs en information. Afin de satisfaire ses besoins d'utilisateurs qui tentent à avoir une information pertinente, les outils de gestion de l'information ont besoin d'extraire des descripteurs existantes déjà dans ces documents. Cependant, l'acquisition ou l'extraction de ces descripteurs est toujours un problème crucial et d'actualité. L'extraction des descripteurs d'une manière manuelle est une tâche lourde et coûteuse à cause de la masse et de la diversité des volumes de documents à traiter.

Le processus de représentation de ces documents est appelé le processus d'indexation ou tout simplement l'indexation [164]. L'indexation consiste à analyser les documents et la requête afin d'extraire un ensemble de descripteur [165]. Ces descripteurs sont des unités textuelles significatives dans le document. Dans une indexation classique, les descripteurs d'un document peuvent être des termes simples ou des termes composés.

Dans une indexation manuelle, chaque document du corpus est examiné par un documentaliste spécialisé dans le domaine afin d'identifier les descripteurs. À la fin de cette étape d'analyse des documents, une liste de descripteurs est établie. Ce type d'indexation est fiable et donne des bons résultats. En effet, suite au développement rapide des connaissances et des technologies, de nouveaux mots sont ajoutés aux langues d'une manière continue. Afin de gérer ces ajouts, les compétences des documentalistes et des spécialistes doivent être mises à jour continuellement. Ainsi, des méthodes et des outils d'indexation issus des Traitements Automatiques de la Langue Naturelle (TALN) [166] ont été proposés afin de rendre cette tâche entièrement automatique. Cependant, comparés aux résultats de l'indexation manuelle, les résultats obtenus par une indexation automatique sont souvent jugés insatisfaisants. Pour remédier à ce défaut, certains travaux [167] proposent d'exposer les résultats de l'indexation automatique à un documentaliste. Ce dernier sélectionne les descripteurs jugés valides parmi la liste des descripteurs exposés. Ce type d'indexation est appelé indexation semi-automatique ou indexation supervisée.

Que ce soit le processus d'indexation manuelle, supervisée ou automatique, un ensemble de descripteurs est associé à chaque document du corpus. L'ensemble des descripteurs permettant de représenter les documents du corpus constituent le langage d'indexation ou le jeu d'indexation [168]. Dans l'indexation manuelle et l'indexation semi-automatique, le jeu d'indexation est réduit à un ensemble de descripteurs jugés valides par l'expert.
En résumé, chaque descripteur extrait d'une manière automatique doit être validé par un spécialiste. On parle alors d'un langage d'indexation contrôlé. Contrairement à l'indexation manuelle et à l'indexation semi-automatique, en indexation automatique le jeu d'indexation est constitué de tous les descripteurs issus de l'analyse automatique des documents du corpus.

2.1. Notion de corpus

Un corpus est un ensemble de documents (textes, images, vidéos, etc.), regroupés dans une optique précise. On peut utiliser des corpus dans plusieurs domaines : études littéraires, linguistiques, scientifiques, philosophiques [169]. Les corpus sont des outils indispensables et précieux en traitement automatique du langage naturel. Ils permettent en effet d'extraire un ensemble d'informations utiles pour des traitements statistiques.
D'un point de vue méthodologique, ils apportent une objectivité nécessaire à la validation scientifique en traitement automatique du langage naturel. Il est donc possible de s'appuyer sur des corpus (à condition qu'ils soient bien formés) pour formuler et vérifier des hypothèses scientifiques.

2.2. Processus d'extraction des termes pertinents

Dans cette section, nous présentons la démarche proposée afin d'extraire automatiquement les termes pertinents à partir d'un corpus de documents. Les étapes de cette démarche sont :

a. La segmentation du texte [170],

b. La lemmatisation [171],

c. Racinisation (Stemming) [172]

d. L'extraction des mots vides,

e. Suppression des mots vides.

Le détail de chaque étape sera présenté dans ce qui suit.

a. La segmentation ou Tokenization :

Le prétraitement du corpus est l'étape préliminaire pour identifier les données lexicales à partir des textes des documents. Afin d'assurer l'adaptabilité de notre modèle à de nouveaux corpus, nous avons travaillé sur des textes bruts. Le prétraitement consiste à segmenter le texte en phrases puis en mots en se basant sur des délimiteurs.
Segmentation ou Tokenization est le découpage en groupes homogènes et distincts par un critère significatif. Cette étape permet d'extraire les mots à partir du document on jouant sur les espaces ou les parenthèses ou l'apostrophe pour la langue française. Un token (élément en français) est une séquence de caractères compris entre deux séparateurs. Un séparateur peut être un "blanc", une ponctuation, une parenthèse, etc. Plusieurs procédures d'analyse nécessitent une segmentation du texte comme un des premiers pas d'analyse.

De plus, la segmentation d'un texte en phrases se fait par l'utilisation des marqueurs de ponctuation : «.», « ?» et «!». Nous n'avons pas traité les cas particuliers avec la présence du point tels que : les adresses mail et les abréviations. En outre, la segmentation en phrases permet d'attribuer à chaque phrase du document un identifiant, son rang d'apparition dans le document. Ces identifiants des phrases sont utilisés dans les étapes ultérieures pour déterminer si deux termes occurrents dans la même phrase.

Pour la segmentation des mots permet de segmenter les phrases en une suite de mots à l'aide des caractères non-alphabétiques, «blanc», «tabulation», «.», «]», etc. Les dates et nombres ne sont pas pris en compte dans la segmentation en mots.

b. La lemmatisation :

La lemmatisation [171] désigne l'analyse lexicale du contenu d'un texte regroupant les mots d'une même famille. Chacun des mots d'un contenu se trouve ainsi réduit en une entité appelée lemme. La lemmatisation regroupe les différentes formes que peut revêtir un mot, soit : le nom, le pluriel, le verbe à l'infinitif, etc.
Une phase préalable importante à l'indexation des documents est la lemmatisation des mots. Le terme lemmatisation fait référence à la réduction des mots à leur forme canonique (racine) de sorte que, par exemple, différentes formes grammaticales ou déclinaisons de verbes soient identifiées et indexées comme une occurrence du même mot. Par exemple, le processus de lemmatisation garantit que les mots "voyage" et "voyagé" seront reconnu par le programme comme un seul et même mot.
Les mots (lemmes) d'une langue utilisent plusieurs formes en fonction de leur genre (masculin ou féminin), leur nombre (singulier ou pluriel), leur personne (moi, toi, eux...), leur mode (indicatif, impératif...) donnant ainsi naissance à plusieurs formes pour un même lemme.

c. La racinisation (Stemming) :

La racinisation ou désuffixation (stemming) est un procédé de transformation des flexions en leur radical ou racine (stem). La racine d'un mot correspond à la partie du mot restante une fois que l'on a supprimé son (ses) préfixe(s) et suffixe(s), à savoir son radical. Contrairement au lemme qui correspond à un mot réel de la langue, la racine ne correspond généralement pas à un mot réel. Par exemple, le mot « recommander » a pour radical « recommand » qui ne correspond pas à un mot réel. L'idée générale de la racinisation est de ramener les mots à leur racine par exemple les mots malade, malades, maladie, maladies, maladive devient après la racinisation malad.

Les techniques utilisées pour ce faire reposent généralement sur une liste d'affixes (suffixes, préfixes, postfixe, antéfixes) de la langue considérée et sur un ensemble de règles de racinisation/désuffixation construites a priori qui permettent, étant donné un mot de trouver sa racine.

Un programme informatique de racinisation est appelé un racinisateur (stemmer). Les algorithmes les plus connus ont été développés par Julie Beth Lovins (1968) [173] et Martin Porter (1980) [174]. La racinisation est un procédé fréquent dans les applications de traitement automatique du langage naturel, par exemple dans la traduction automatique, la recherche d'information (reconnaissance d'entités) et l'indexation des moteurs de recherche.

d. L'extraction des mots vides candidats :

Un mot vide candidat est un mot susceptible d'être un mot vide. Dans cette étape nous affectons à chaque mot une catégorie : vide ou pleine. Les mots vides sont des mots qui

sont communs à tous les textes dans une même langue. Ils ont une utilité fonctionnelle. En français, les mots vides évidents pourraient être « le », « la », « de », « du », « ce », « ça », etc. Dans un contexte monolingue où tous les documents du corpus sont rédigés dans une même langue, les mots vides sont principalement des mots caractéristiques de cette langue tels que les pronoms, les prépositions, les articles, etc. dans ce contexte les mots vides ont dits encore mots grammaticaux. Alors il est inutile de les indexer ou de les utiliser dans un processus de recherche d'information. Dans un texte, un mot vide est un mot non significatif contrairement à un mot plein.

e. La suppression des mots vides :

Les mots vides (Stop words en anglais) sont les mots qui se répètent fréquemment dans tous les documents et qui n'ont aucun pouvoir discriminant lors du processus de la catégorisation de texte. La listes des mots vides contient les pronoms personnels, les prépositions, les articles, etc. Si le mot apparu est un mot vide alors le système doit le supprimer. Pour réaliser ce travail nous avons un document qui contient tous les mots vides de la langue anglaise et lors de l'indexation si un mot d'un document est un mot vide nous ne le prenons pas en considération.

2.3. Les différents modèles de représentation d'un document texte

Dans cette section nous introduisons quelques modèles classiques de représentation des documents texte à savoir le modèle booléen, probabiliste et vectoriel. Ces modèles ont en commun le vocabulaire d'indexation et se basent sur le formalisme des mots clés. Le vocabulaire d'indexation est constitué des mots qui apparaissent dans les documents.

2.3.1. Le modèle booléen

Ce modèle est le plus ancien dans le domaine de recherche d'information. La simplicité de ce modèle a fait son succès. La requête est représentée sous forme d'une expression logique [165]. Dans cette expression, les descripteurs sont combinés entre eux en utilisant les opérateurs booléens ET, OU et XOR. Les documents satisfaisant l'expression logique représentant la requête sont considérés comme pertinents.
Dans le modèle booléen, la pertinence des documents est une variable booléenne ce qui ne permet pas de trier dans un ordre de pertinence les documents retournés [175]. L'utilisateur est donc obligé de consulter tous les documents de la réponse afin de trouver les documents recherchés. Pour remédier à cette limite, un modèle étendu a été proposé dans [176]. Le modèle booléen étendu affecte à chaque terme dans le document et dans la requête une pondération.

2.3.2. Le modèle probabiliste

Le modèle probabiliste [177] est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête. L'idée est de retrouver des documents qui ont en même temps

une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Les documents – et les requêtes – sont représentés par des vecteurs de booléens dans un espace à n dimensions.

$$d_j = (w_{1,j}, w_{2,j}, ..., w_{n,j}) \tag{3.1}$$

$$q = (w_{1,q}, w_{2,q}, ..., w_{n,q}) \tag{3.2}$$

avec $w_{k,j} \in \{0, 1\}$ et $w_{k,q} \in \{0, 1\}$. La valeur de $w_{k,j}$ (resp. $w_{k,q}$) représente le fait que le terme t_k apparaît dans le document d_j (resp. la requête q).
La similarité entre le document d_j et la requête q est alors calculée en fonction de ces deux probabilités de la manière suivante.

$$sim(d_j, q) = \frac{P(R/d_j)}{P(NR/d_j)} \tag{3.3}$$

avec $P(R/d_j)$ (resp. $P(NR/d_j)$) la probabilité que la réponse soit pertinente (resp. non pertinente) pour la requête q, étant donné le document d_j comme réponse.
Le modèle probabiliste se démarque des autres modèles classiques par l'utilisation des probabilités pour représenter explicitement le concept fondamental de la recherche d'information qu'est la pertinence.

2.3.3. Le modèle vectoriel

Le modèle vectoriel représente un document - ainsi qu'une requête - par un vecteur dans un espace dont chaque dimension correspond à un descripteur atomique. Chaque coordonnée dans cet espace dénote l'importance du descripteur dans le document considéré. Le traitement d'une requête est alors basé sur la comparaison des vecteurs documents et requête. Ce modèle vectoriel de représentation de documents est proposé par Gerard Salton dans les années 1970 [178]. Il est utilisé en recherche d'information, notamment pour la recherche documentaire, la classification ou le filtrage de données. Ce modèle concernait originellement les documents textuels et a été étendu à d'autres types de contenus.

L'ensemble de représentation des documents est un vocabulaire comprenant des termes d'indexation. Ceux-ci sont typiquement les mots les plus significatifs du corpus considéré : noms communs, noms propres, adjectifs... Ils peuvent éventuellement être des constructions plus élaborées comme des expressions ou des entités sémantiques. Chaque contenu est ainsi représenté par un vecteur $\overrightarrow{v}$, dont la dimension correspond à la taille du vocabulaire. Chaque élément v_i du vecteur $\overrightarrow{v}$ consiste en un poids associé au terme d'indice i. Un exemple simple est d'attribuer v_i au nombre d'occurrences du terme i dans l'échantillon de texte. La composante du vecteur représente donc le poids du mot i dans le document.

2.5.3.1. Les types de pondération

La représentation des documents dans un espace vectoriel consiste à voir un texte comme un sac de mots (il n'y a plus d'ordre). On associe à chaque mot, un poids (nombre réel), mesurant son *importance* dans le texte. Chaque coordonnée correspond au degré d'importance d'un mot donné dans le texte. En effet, les méthodes de pondération les plus largement

utilisées pour le texte sont :

2.5.3.1.1. Pondération binaire

La pondération binaire accepte deux états soit 0, soit 1. Si le terme est présent dans le document c'est l'état 1, sinon c'est l'état 0.
Exemple : soit les deux documents textes suivants (d1 et d2). Chaque document contient une phrase.
d1 :≪Learning online with Moodle is an art≫.
d2 :≪Moodle is an open-source platform that facilitates learning enormously to the learner≫.
Le tableau 3.1 illustre la représentation vectorielle des deux documents.

	Learning	Online	With	Moodle	Is	An	Art	Open-source	Platform	That	Facilitates	Enormously	To	The	learner
d1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
d2	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1

TABLE 3.1 – Exemple d'une pondération binaire

2.5.3.1.2. Pondération par nombre d'occurrences

Cette méthode de pondération consiste à compter le nombre d'occurrence (fréquence) d'un mot dans un document.
Exemple : soit les deux documents textes suivants (d1 et d2).
d1 :≪Learning online with Moodle is an art≫.
d2 :≪Moodle is an open-source platform that facilitates learning enormously to the learner. Indeed, learning online with Moodle is an art≫.
Le tableau 3.2 illustre la représentation vectorielle des deux documents.

	Learning	Online	With	Moodle	Is	An	Art	Open-source	Platform	That	Facilitates	Enormously	To	The	Learner	Indeed
d1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
d2	2	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1

TABLE 3.2 – Exemple d'une pondération par nombre d'occurrences

2.5.3.1.3. Méthode de pondération intelligente TF×IDF

Le TF×IDF (Term Frequency-Inverse Document Frequency) [179] est une méthode de pondération souvent utilisée en recherche d'information. Cette mesure statistique permet d'estimer l'importance d'un mot par rapport au document qui le contient, en tenant

compte du poids de ce mot dans le corpus complet. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Des variantes de la formule originale sont souvent utilisées dans des moteurs de recherche pour apprécier la pertinence d'un document en fonction des critères de recherche de l'utilisateur.

Si une requête contient le terme t, un document a d'autant plus de chances d'y répondre qu'il contient ce terme. Néanmoins, si le terme t est lui-même très fréquent au sein du corpus, c'est-à-dire qu'il est présent dans de nombreux documents (e.g. les articles définis : le, la, les), il est en fait peu discriminant. C'est pourquoi cette méthode propose d'augmenter la pertinence d'un terme en fonction de sa rareté au sein du corpus (fréquence du terme dans le corpus élevée). Ainsi, la présence d'un terme rare de la requête dans le contenu d'un document fait croître le "score " de ce dernier.

Le TF×IDF définit le poids d'un terme comme le produit de deux informations :
- La fréquence $tf(t, A)$ du terme t dans le document A :

$$tf(t, A) = \frac{n_{t,A}}{N_A}, \tag{3.4}$$

où $n_{t,A}$ est le nombre d'occurrence de t dans A et N_A la taille du document A.

- La fréquence inverse $idf(t, D)$ du terme t dans le corpus D :

$$idf(t, D) = \log(\frac{|D|}{|\{A \in D : t \in A\}|}), \tag{3.5}$$

où $|D|$: Le nombre total de documents dans le corpus.
$|A \in D : t \in A|$: Le nombre de documents où le terme t apparaît.
Il s'agit de diviser le nombre total de documents présents dans le corpus D par le nombre de documents contenant le terme t, et d'en calculer le logarithme. La fréquence inverse de document IDF(Inverse Document Frequency) est une mesure de l'importance du terme dans l'ensemble du corpus.

Finalement, le poids $tfidf(t, A, D)$ du terme t, appartenant au document A du corpus D s'obtient en multipliant les deux mesures($tfidf(t, A, D) = tf(t, A) \times idf(t, D)$). En outre, le TF×IDF vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants.

Dans cette section, nous avons présenté les traitements qui s'effectuent sur un corpus textuel à savoir le prétraitement, la segmentation, la lemmatisation et la suppression des mots vides. Cependant, nous avons commencé par la notion de corpus, les différents types de corpus à savoir les corpus comparables, les corpus parallèles et les corpus multilingues. Puis, nous avons cité les différents modèles de représentation d'un document texte à savoir le modèle booléen, le modèle probabiliste et le modèle vectoriel.

Dans la section suivante, nous allons découvrir la notion de l'indexation et les différentes méthodes de calcul de similarité.

3. Indexation sémantique et calcul de similarité

L'indexation automatique de documents est un concept qui relève du domaine des sciences de l'information et des bibliothèques et qui utilise des méthodes logicielles pour établir un index pour un ensemble de documents et faciliter l'accès à leur contenu.

L'indexation automatique de documents essaye de répondre à la question suivante : Comment organiser au mieux une collection de documents afin de pouvoir plus tard retrouver facilement celui qui nous intéresse ?

Une réponse classique consiste à annoter manuellement chaque document d'une série de métadonnées (titre, catégorie(s), date de parution, auteur etc.). Cette approche a l'avantage d'être facile à mettre en œuvre, et de fournir des renseignements de qualité (selon l'expertise de la personne chargée de l'annotation). Cependant, cette solution est ambiguë (un même document pouvant se décrire de plusieurs façons) et elle est coûteuse (puisqu'il faut payer un expert pour qu'il prenne en charge tout nouveau document dans notre collection).

Par conséquent, et pour remédier à ce types de problème, l'indexation sémantique est une nécessité. Elle se base directement sur le contenu, dans le but d'obtenir des résultats cohérents. Ce type d'indexation doit souvent refléter le sens du contenu des documents (nous pouvons parler ici d'une approche sémantique) [180]. En outre, l'indexation a d'abord comme but de marquer les documents et d'orienter le public vers les documents pertinents. En représentant les documents sous forme de vecteurs de descripteurs, il devient possible de les comparer et de mesurer leurs distances les uns des autres par des algorithmes de mesure de similarité.

La prochaine section présente les ressources linguistiques externes les plus exploitées dans l'indexation sémantique, ainsi que les approches de calcul de similarité entre les documents texte.

3.1. Les ressources sémantiques utilisées en indexation

Dans cette section, nous présentons quelques ressources sémantiques externes utilisées en indexation. Certains types de ces ressources sont populaires, nous reprenons leurs définitions dans le domaine des technologies de l'information :

- **Taxonomie** : Une taxonomie est un vocabulaire contrôlé organisé sous une forme hiérarchique simple. Les liens hiérarchiques dans une taxonomie correspondent à des liens de spécialisation affinant le sens d'un terme (Il existe donc un lien précis entre un terme du vocabulaire et ses enfants). La taxonomie, à la différence du thesaurus ne permet pas de parcourir la hiérarchie de manière connexe ni de restreindre ou de spécialiser le champ des connaissances [181].

- **Thésaurus** : Un thesaurus constitue un dictionnaire (vocabulaire contrôlé) hiérarchisé [182]. Ce vocabulaire est normalisé et présente les termes génériques ou spécifiques

à un domaine. Les termes y sont organisés dans une hiérarchie de concepts liés par des relations sémantiques. Un thesaurus peut également fournir des définitions. Les relations couramment présentes dans un thesaurus sont des relations taxonomiques (hiérarchie), d'équivalence (synonymie), d'association (proximité sémantique, proche de, relié-à, ...).

- **Ontologies** : Les ontologies permettent de décrire un ensemble de connaissances spécifiques ou générales, mais aussi de représenter des relations entre concepts complexes, ainsi que des axiomes et règles absentes des thesaurus. Cette formalisation permet aux outils informatiques de raisonner ce qui leur confère un meilleur niveau de compréhension [183].Il devient ainsi possible d'inférer de nouvelles connaissances à partir de celles déjà acquises.

- **Les réseaux sémantiques : WordNet** : Un réseau sémantique est un graphe orienté et étiqueté (ou, plus précisément, multi-graphe) [184]. Un arc lie (au moins) un nœud de départ à (au moins) un nœud d'arrivée. Les relations sont des relations de proximité sémantique de types partie-de, cause-effet, parent-enfant, etc. Les concepts sont représentés sous forme de nœuds et les relations sous forme d'arcs. Un des réseaux sémantiques les plus utilisés est Wordnet [185], largement utilisé pour la recherche d'information. Wordnet est un réseau lexical et sémantique qui a été initialement élaboré pour la langue anglaise. Les noms, verbes, adjectifs et adverbes sont organisés en ensembles de synonymes (appelés synsets), chaque ensemble représente un concept lexical. De plus, différentes relations lient les ensembles de synonymes.

Plusieurs travaux ont utilisés des ressources sémantiques dans le processus d'indexation. En effet, le vocabulaire contrôlé ne contient que des termes d'indexation pour faciliter l'accès et l'indexation des documents, alors que la taxonomie classifie, organise et ajoute des relations hiérarchiques entre groupes de termes. Le thésaurus possède en plus des relations sémantiques entre termes d'autres relations comme la causalité, l'association, etc. L'ontologie, quant à elle, est plutôt la représentation formelle des informations qui sont définies sous la forme des concepts avec leurs relations.

3.2. Les méthodes de calcul de similarité entre documents textes

Dans cette section de ce chapitre, nous présentons certaines approches permettant de comparer des textes. Ainsi, nous ne prétendons pas donner une liste exhaustive de toutes les méthodes existantes, mais nous allons tenter de donner un aperçu sur les méthodes les plus utilisées pour le calcul de similarité entre documents textes.

3.2.1. Introduction

Les approches présentées dans ce qui suit consistent à donner une vision sur les mesures de similarité entre documents afin de prendre la meilleure décision de développement

possible. Il s'agit de comparer des textes de petites tailles. De plus, les résultats devront donc être pertinentes et être obtenus le plus rapidement possible. Ainsi, nous souhaitons déterminer, pour un texte donné T, un ensemble de textes T_{min} similaire, c'est-à-dire les textes qui sont plus proches de T.

La similarité entre documents textuels est une problématique importante de plusieurs disciplines comme l'analyse de données textuelles, la recherche d'information ou l'extraction de connaissances (Text Mining), la recommandation de documents, etc. Dans chacun de ces domaines, les similarités sont utilisées pour différents traitements :

- En analyse de données textuelles, les similarités sont utilisées pour la description et l'exploration de données ;
- En recherche d'information, l'évaluation de la similarité entre une requête et un ensemble de documents est utilisée pour identifier les documents pertinents par rapport à la requête de l'utilisateur ;
- En Text Mining, les similarités sont utilisées pour des représentations synthétiques de vastes collections de documents.
- En recommandation des documents, la similarité est calculée entre le document choisi et les autres documents qui peuvent intéressés l'utilisateur.

Pour le calcul de la similarité entre documents textuels nous distinguons deux types de similarité, la similarité syntaxique et la similarité sémantique :

3.2.2. Calcul de similarité syntaxique

Une mesure permettant de comparer des documents textuels, consiste à comparer les caractères constituants ces documents. C'est une métrique qui mesure la similarité ou la dissimilarité entre deux chaines de caractères. Par exemple "banquier" et "banquière" peuvent être considérés comme très proches alors que "banquier" et "financier" pourront être considérés comme très différents.

Nous présentons dans ce qui suit les mesures de similarité syntaxique les plus utilisées en exploitant la représentation vectorielle d'un document.

3.2.2.1. La similarité cosinus

Étant donnée une représentation vectorielle d'un corpus de documents, nous pouvons introduire la notion mathématique de proximité entre documents, en exploitant une mesure très utilisée qu'est la similarité cosinus [186], qui consiste à quantifier la similarité entre deux documents en calculant le cosinus entre leurs vecteurs.

La similarité cosinus (ou mesure cosinus) permet de calculer la similarité entre deux vecteurs (deux document A et B) en déterminant le cosinus de l'angle entre eux. Cette métrique est fréquemment utilisée en fouille de textes.
Soit deux vecteurs $\overrightarrow{A}$ et $\overrightarrow{B}$, la similarité $sim_{cosinus}(\overrightarrow{A}, \overrightarrow{B})$ s'obtient par le produit scalaire et la norme des vecteurs (figure 3.3) :

$$sim_{cosinus}(\overrightarrow{A}, \overrightarrow{B}) = \frac{\overrightarrow{A} \times \overrightarrow{B}}{\|\overrightarrow{A}\| \times \|\overrightarrow{B}\|} \tag{3.6}$$

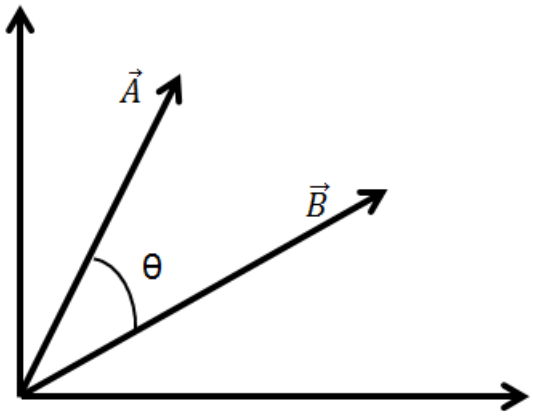

FIGURE 3.3 – La proximité des deux documents A et B est représentée par l'angle θ

En particulier, une valeur nulle indique que la requête est strictement orthogonale au document. Physiquement, cela traduit l'absence de mots en commun entre A et B. Comme la valeur cosθ est comprise dans l'intervalle [-1,1], la valeur -1 indiquera des vecteurs résolument opposés (aucune similarité entre les deux documents), 0 des vecteurs orthogonaux et 1 des vecteurs similaires. Les valeurs intermédiaires permettent d'évaluer le degré de similarité.

3.2.2.2. La distance Euclidienne

La distance euclidienne $\|\overrightarrow{AB}\|$ est définie comme la racine carrée de la somme des différences à la puissance deux entre les attributs de même rang des deux vecteurs $\overrightarrow{A}$ et $\overrightarrow{B}$.

$$d(\overrightarrow{A}, \overrightarrow{B}) = \sqrt{\sum_{j=1}^{n}(A_j - B_j)^2} \tag{3.7}$$

Pour obtenir un indice de la similarité existant entre $\overrightarrow{A}$ et $\overrightarrow{B}$, puisque plus la distance est faible, plus la similarité est censée augmenter, on peut prendre l'inverse de la distance euclidienne. Pour éviter d'avoir un dénominateur nul, on ajoute 1 :

$$sim(\overrightarrow{A}, \overrightarrow{B}) = \frac{1}{1 + \sqrt{\sum_{j=1}^{n}(A_j - B_j)^2}} \tag{3.8}$$

De ce fait, on a $0 < sim(\overrightarrow{A}, \overrightarrow{B}) \leq 1$ et plus $sim(\overrightarrow{A}, \overrightarrow{B})$ s'approche de 1, plus la similarité entre les documents A et B est grande.

3.2.2.3. Le coefficient de Jaccard

Le coefficient de Jaccard [187] est le rapport entre la cardinalité (la taille) de l'intersection des ensembles considérés et la cardinalité de l'union des ensembles. Il permet d'évaluer la similarité entre les ensembles. Les documents A et B sont donc représentés, non pas comme des vecteurs, mais comme des ensembles de terme. La similarité obtenue $sim_{jaccard}(A, B) \in [0, 1]$.

$$sim_{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.9}$$

Il est aussi possible d'utiliser la représentation vectorielle.

$$sim_{jaccard}(\overrightarrow{A}, \overrightarrow{B}) = \frac{\overrightarrow{A} \times \overrightarrow{B}}{\|\overrightarrow{A}\|\|\overrightarrow{B}\| - \overrightarrow{A} \times \overrightarrow{B}} \tag{3.10}$$

3.2.2.4. L'indice de Dice

L'indice de Dice [188] mesure la similarité entre deux documents A et B en se basant sur le nombre des termes communs entre A et B.

$$sim_{dice}(A, B) = \frac{2N_c}{N_1 + N_2}, \tag{3.11}$$

où N_c est le nombre de termes communs à A et B, et N_1 (respectivement N_2) est le nombre de termes de A (respectivement B).

3.2.2.5. Coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson [189] calcul la similarité entre deux documents A et B comme le cosinus de l'angle entre leurs représentations vectorielles centrées-réduites. La similarité obtenue $sim_{pearson}(A, B) \in [-1, 1]$.

$$sim_{pearson}(A, B) = sim_{cosinus}(A - \overline{A}, B - \overline{B}), \tag{3.12}$$

où $\overline{A}$ (respectivement $\overline{B}$) représente la moyenne de A (respectivement B)

3.2.2.6. Avantages et inconvénients

Nous essayons de lister quelques avantages et inconvénients liées aux approches syntaxiques. D'après [190] [191] les performances de la similarité cosinus, Indice de Dice, coefficient de Jaccard et coefficient de Person sont très proches et qu'elles sont significativement meilleures que celles de la distance euclidienne. Cependant, [192] affirme que plus le document est de petite taille, plus les résultats obtenus avec la distance euclidienne sont meilleurs, tandis qu'ils sont plus mauvais avec la similarité cosinus ou avec le coefficient de Jaccard.
Avantages :
- Les techniques basées sur l'approche syntaxique ne laissent pas de place aux exceptions ;
- Elles sont faciles dans leur mise en œuvre.
Inconvénients :
- Les techniques basées sur l'approche syntaxique ne prennent pas en compte la sémantique. Par exemple, il est difficile de trouver une forte similarité entre « j'ai une pomme » et « j'ai

un fruit ». En effet, la prise en compte de la sémantique semble importante.

3.2.3. Calcul de similarité sémantique

La similarité sémantique est un concept selon lequel nous évaluons la ressemblance d'un ensemble de documents ou de termes par une métrique basée sur la signification (contenu sémantique) de ces derniers. Concrètement, cela peut être réalisé en définissant une similitude topologique, par exemple, en utilisant des ontologies pour définir une distance entre les mots, ou en définissant une similitude statistique, par exemple en utilisant un modèle d'espace vectoriel pour corréler les termes et les contextes à partir d'un corpus de texte approprié. Parmi de telles mesures de similarité, citons par exemple, Resnik [193], LSA (Analyse sémantique latente) [194], ESA (Analyse sémantique explicité) [195], etc.

Pourtant, la distance sémantique peut être de deux types : la similarité sémantique et la parenté sémantique. La première est un sous-ensemble de la seconde, mais les deux termes peuvent être utilisés indifféremment, ce qui rend encore plus important d'être conscient de leur distinction. Deux concepts sont considérés comme sémantiquement similaires s'il y a une synonymie, hyponymie, antonymie, ou toponymie entre eux.
Les mesures de similarité de textes ont été utilisées dans de nombreux domaines. Par exemple, pour la classification de textes [196], la désambiguïsation du sens [197], la traduction automatique [198]. Outre cela, les approches de calcul de similarité sémantique sont classifiées en deux catégories, Approche basée sur la distance et approche basée sur le contenu informationnel :

3.2.3.1. Approche basée sur la distance

Cette similarité est évaluée par la distance qui sépare les objets dans une ontologie. Ces mesures se servent de la structure hiérarchique de l'ontologie pour déterminer la similarité sémantique entre les concepts. Le calcul des distances dans l'ontologie est basé sur un graphe de spécialisation des objets. Dans chaque graphe, la distance de l'ontologie doit être caractérisée par le plus court chemin qui fait intervenir un ancêtre commun ou le plus petit généralisant, connectant potentiellement deux objets à travers des descendants communs. Parmi les travaux classifiés sous cette optique nous pouvons citer :

- La méthode de Rada

La méthode de Rada [199] est basée sur le concept de l'ontologie. Cette dernière est représentée par un graphe dont les nœuds sont des concepts, les arcs sont les liens entre concepts. En effet, la similarité entre deux concepts peut être calculée à partir du nombre de liens qui séparent les deux concepts. Plusieurs variantes existent en fonction du chemin pris en compte pour calculer la distance entre les concepts. La mesure proposée par Rada et al. (1989) utilise une métrique $dist(c_1, c_2)$ qui indique le nombre d'arcs séparant les deux concepts c_1 et c_2 par le plus court chemin dans la hiérarchie.

$$sim_{rada}(c_1, c_2) = \frac{1}{1 + dist(c_1, c_2)}, \tag{3.13}$$

où $dist(c_1, c_2)$ est la distance entre $c_1 etc_2$

- La méthode de Wu et Palmer

Wu et Palmer (1994) [200] ont proposé une autre méthode basée sur le plus petit généralisant commun, c'est-à-dire le généralisant commun à c_1 et c_2 le plus éloigné de la racine.

$$sim_{wupalmer}(c_1, c_2) = \frac{2 \times depth(c)}{depth(c_1) + depth(c_2)}, \tag{3.14}$$

où $depth(c_i)$ correspond au niveau de la profondeur du concept c_i (son emplacement dans l'hiérarchie des concepts), et c représente le plus petit ancêtre commun à c_1 et c_2 (figure 3.4).

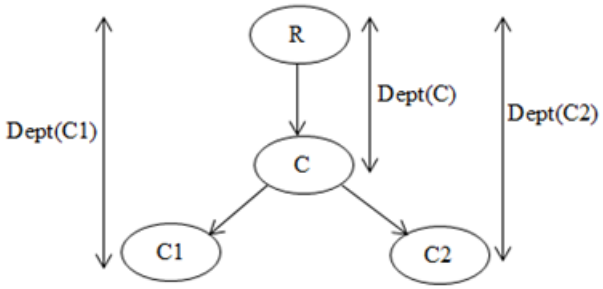

FIGURE 3.4 – Un extrait d'une hiérarchie de concepts

Ainsi, l'approche basée sur la distance présente l'inconvénient que la similarité dépend de l'organisation des concepts dans la hiérarchie. La valeur de la similarité dépend des choix pris lors de la construction de la hiérarchie des concepts.

- La méthode de de Hirst et St.Onge

L'idée de base de la méthode de de Hirst et St.Onge [201] est que si deux concepts sont reliés entre eux par un chemin très court et qui « ne change pas la direction » alors les deux concepts sont similaires. Le calcul de la similarité est basé sur le poids du chemin le plus court qui mène d'un concept à un autre et le nombre de changements de directions.

$$sim_{Hist}(c_1, c_2) = T - PCC - K \times D \tag{3.15}$$

Où : T et K sont des constantes, PCC est la distance du plus court chemin en nombre d'arc et D le nombre de changements de direction.

3.2.3.2. Approche basée sur le contenu informationnel

Cette approche prend en considération le contenu informatif [202] des concepts de l'ontologie. La similarité est alors calculée à partir de l'information partagée par les concepts. Cette approche adopte une nouvelle mesure qui est la mesure entropique de la théorie de

l'information.

$$E(c) = -\log(p(c)) \tag{3.16}$$

La probabilité d'un concept c est calculée en divisant le nombre des instances de c par le nombre total des instances. En associant des probabilités aux concepts d'une taxonomie, il est possible d'éviter le manque de fiabilité des distances des arcs. Cette caractéristique quantitative de l'information fournit une nouvelle façon de mesurer la similarité sémantique. Plus l'information est partagée par deux concepts, plus ils sont similaires. Parmi les travaux, recensés dans la littérature, sous cette bannière on peut citer :

- La méthode de resnik

La notion du contenu informationnel (CI) a été initialement introduite par Resnik [202] qui a prouvé que la similarité sémantique entre deux concepts est mesurée par la quantité de l'information qu'ils partagent. Pour évaluer la pertinence d'un objet, il faut calculer le contenu informationnel. Le contenu informationnel est obtenu en calculant la fréquence de l'objet dans le corpus (Wordnet). La mesure de similarités proposée par Resnik est donnée par la formule suivante :

$$sim_{Resnik}(c_1, c_2) = Max[-\log(P(CS(c_1, c_2)))], \tag{3.17}$$

où $CS(c_1, c_2)$ représente le concept le plus spécifique (maximise la valeur de similarité) qui subsume (situé à un niveau hiérarchique plus élevé) les deux concepts c_1 et c_2 dans l'ontologie.
La probabilité d'un mot quelconque à être une instance du concept c (figure 3.5) :

$$P(c) = \frac{\sum_{w \in Words(C)} count(w)}{N} \tag{3.18}$$

où $Words(c)$ est l'ensemble de mots englobés par le concept c et N le nombre de mots dans le corpus.

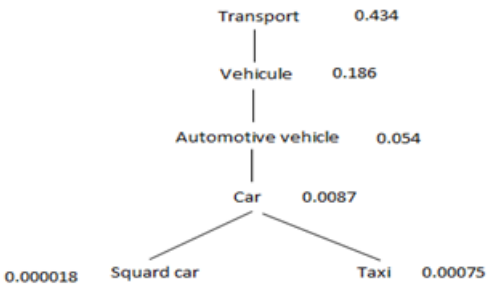

FIGURE 3.5 – Un fragment de la hiérarchie de WordNet montrant la probabilité p(c) attaché à chaque concept

En fait, pour mesurer la similarité entre deux concepts, la façon la plus simple pour exploiter le contenu informationnel de ces concepts est la méthode de Resnik.

- La méthode de Lin

Lin (1998) [203] a étendu l'intuition de Resnik par la mesure de la quantité d'informations en commun entre c_1 et c_2. Par exemple, il a souligné que plus il y a de différences entre c_1 et c_2, moins ils sont similaires.
La mesure de Lin est donnée par la formule suivante :

$$sim_{Lin}(c_1, c_2) = \frac{2 \times \log(P(CS(c_1, c_2)))}{\log(P(c_1)) + \log(P(c_2))} \tag{3.19}$$

Cette mesure utilise une approche hybride qui combine deux sources de connaissances différentes (Thesaurus, corpus).
En plus, elle représente la similarité comme degré probabiliste de chevauchement des concepts descendants de c_1 et c_2.

3.2.3.3. Approche mixte

Cette approche est fondée sur un modèle qui combine l'approche basée sur les arcs (distances) et l'approche basée sur le contenu informationnel.

- La méthode de Jiang et conrath

Le principe des mesures mixtes consiste à considérer le plus court chemin reliant deux concepts dans l'ontologie et de pondérer ces liens à partir de leur poids sémantique.
Le poids sémantique des liens prend notamment en compte le contenu informationnel des concepts. La mesure de Jiang et Conrath (1997) [204] est donnée par la formule suivante :

$$sim_{jcn}(c_1, c_2) = \frac{1}{dist_{jcn}}, \tag{3.20}$$

où $dist_{jcn}$ est calculée par :

$$dist_{jcn}(c_1, c_2) = E(c_1) + E(c_2) - (2 \times E(CS(c_1, c_2))), \tag{3.21}$$

où $E(c_i$ est le contenu informationnel du concept c_i.

- La méthode de Leacock et Chodorow

Une autre méthode qui combine l'approche de comptage des arcs et l'approche du contenu informationnel est la mesure proposée par Leacock et Chodorow [205], qui est basée sur la longueur du plus court chemin entre deux synsets de Wordnet. Les auteurs ont limité leur attention à des liens hiérarchiques « is-a » ainsi que la longueur de chemin par la profondeur globale de la taxonomie. La formule est définie par :

$$sim_{LC}(c_1, c_2) = -\log(\frac{cd(c_1, c_2)}{2 \times M}), \tag{3.22}$$

où M est la longueur du chemin le plus long qui sépare le concept racine, de l'ontologie, du concept le plus en bas. On dénote par $cd(c_1, c_2)$ la longueur du chemin le plus court qui

sépare c_1 de c_2.

4. Évaluation d'un système de recommandation

Nous allons étudier dans cette partie comment nous pouvons évaluer la performance d'un système de recommandation pour s'assurer de sa capacité à satisfaire les besoins des utilisateurs. Le choix d'une mesure, doit être dépendant du type de données à traiter, et des intérêts des utilisateurs [206]. Comme le domaine de la recommandation découle de celui de la recherche d'information, il est donc souvent normal d'utiliser des mesures d'évaluation de la recherche d'information [207]. Certaines de ces mesures ont été ajustées aux besoins du domaine de la recommandation.

L'évaluation de la performance d'un système de recommandation dans la littérature est souvent limitée au calcul de la précision de la prédiction [206]. La précision mesure, en général, la différence entre les valeurs des itèmes prédites par le système de recommandation et les valeurs réellement fournies par les utilisateurs. Par contre, le rappel détermine la capacité d'un système à retourner tous les documents pertinents pour une demande utilisateur.
Ces deux mesures sont données par les formules suivantes :

- Le rappel est le rapport entre le nombre d'items pertinents sélectionnés et le nombre total d'items pertinents.

$$Rappel = \frac{N_{ps}}{N_p} \tag{3.23}$$

- La précision est le rapport entre le nombre d'items pertinents sélectionnés et le nombre total d'items sélectionnés.

$$Precision = \frac{N_{ps}}{N_s} \tag{3.24}$$

L'objectif est de mesurer la fréquence des bons et mauvais jugements portés par le système de recommandation à l'égard des items.

La $F - mesure$ [206] a été proposée la première fois par Cleverdon en 1968 pour les systèmes de recherche d'information. Cette mesure est un compromis entre le rappel et la précision :

$$F - mesure = \frac{2 \times Rappel \times Precision}{Rappel + Precision} = \frac{2 \times N_{ps}}{N_s + N_p} \tag{3.25}$$

La figure 3.6 représente la répartition des documents suite à une interrogation utilisateur. À partir de ces ensembles de documents les deux mesures précision et rappel sont calculées.

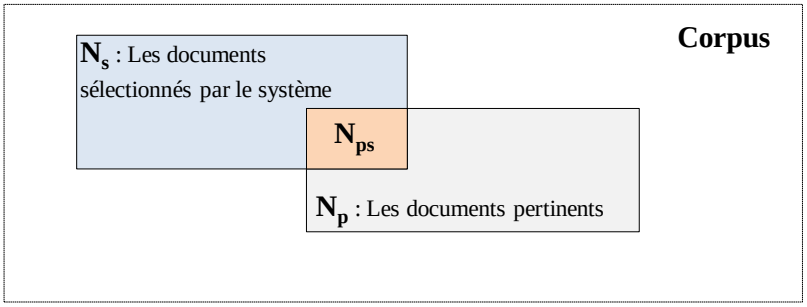

FIGURE 3.6 – Répartition des documents d'un corpus suite à une interrogation

Supposons que dans un cas idéal, un système de recommandation est capable de ramener tous les documents pertinents du corpus et de rejeter tous les documents non pertinents.

$$Rappel = Precision = 1 \tag{3.26}$$

Pour ce système idéal, la valeur précision est égale à la valeur du rappel. Cette valeur est égale à 1.

Conclusion

Dans ce chapitre, nous avons décrit les différents types de systèmes de recommandation, en l'occurrence la recommandation à base de contenu, la recommandation à base d'utilisateurs et la recommandation hybride. Ensuite, nous avons abordez la recommandation basée sur le calcul de la similarité. Ainsi, nous avons identifié deux grandes catégories de calcul de similarité : la similarité syntaxique et la similarité sémantique. Ces deux approches présentent néanmoins des caractéristiques complémentaires.

Nous avons présenté également les concepts fondamentaux de la recommandation sémantique qui est caractérisée par l'utilisation des ressources sémantiques et des mesures de similarités sémantique entre les concepts. Nous avons cité également, quelques travaux de la recommandation sémantique des ressources dans le domaine de e-learning. Finalement, nous avons introduit des mesures d'évaluation de la performance des systèmes de recommandation, dont la précision et le rappel sont les plus populaires.
La conception et la réalisation de nos contributions seront présentées en détail dans les chapitres suivants du présent document.

Chapitre 4

Contribution à l'apprentissage adaptatif et à la recommandation sémantique

"La vie est l'adaptation continue de relations internes à des relations externes"

- Herbert Spencer,

Introduction

L'objectif de ce chapitre est de présenter notre contribution à l'apprentissage adaptatif et à la recommandation sémantique, afin de répondre à une des problématiques soulevées dans les chapitres précédents. Nous présentons, en premier lieu, notre démarche pour le développement d'une plateforme e-learning adaptatif qui permet de générer des parcours d'apprentissage adaptés aux profils des apprenants et selon les objectifs pédagogiques fixés par l'enseignant. Ainsi, nous étudions la problématique de l'adaptation du profil de l'apprenant aux objectifs pédagogiques comme un «problème d'optimisation», en utilisant les algorithmes génétiques pour chercher un parcours optimal.

Nous proposons ensuite un système de recommandation, considéré comme un sous-ensemble d'un système e-learning adaptatif. Ce système innovant de recommandation sémantique permet de retourner des documents susceptibles d'intéresser l'apprenant. Cette recommandation est basée sur une méthode hybride de calcul de similarité sémantique qui combine entre l'approche basée sur les arcs (distances) et la représentation vectorielle des documents. Les résultats obtenus pour chaque expérience sont présentés et discutés afin de tirer des conclusions sur l'efficacité du système de recommandation réalisé, ainsi que les difficultés et les défis qui doivent être relevés.

1. Générer un parcours d'apprentissage adapté au profil Apprenant

Le besoin des plateformes e-learning à assurer une formation évoluée et adaptée nécessite l'introduction de nouvelles approches pour la résolution des problématiques rencontrées. À ce propos, l'adaptabilité des systèmes de formation devient une caractéristique recherchée. L'utilisation des algorithmes génétiques permet d'automatiser la recherche de parcours adaptés au profil de l'apprenant. En effet, nous attribuons des parcours pédagogiques différents aux apprenants appartenant à une même classe.

Dans cette section de ce chapitre, nous allons présenter l'architecture de notre plateforme e-learning adaptatif. Nous présentons aussi l'approche utilisée et la méthodologie de conception du système (modélisation UML) qui répond aux objectifs de notre recherche. Enfin, nous concluons par une expérience et évaluation de notre approche basée sur les algorithmes génétiques.

1.1. Architecture de notre système e-learning adaptatif basé sur les algorithmes génétiques

La figure 4.1 illustre notre système adaptatif conçu pour générer des parcours pédagogiques qui sont adaptés au profil de l'apprenant et à l'objectif pédagogique fixé par l'enseignant. Nous avons étudié la problématique de l'adaptation comme un «problème d'optimisation», en utilisant des algorithmes génétiques. Notre système cherche un parcours optimal à partir du profil de l'apprenant en passant par des objectifs pédagogiques intermédiaires.

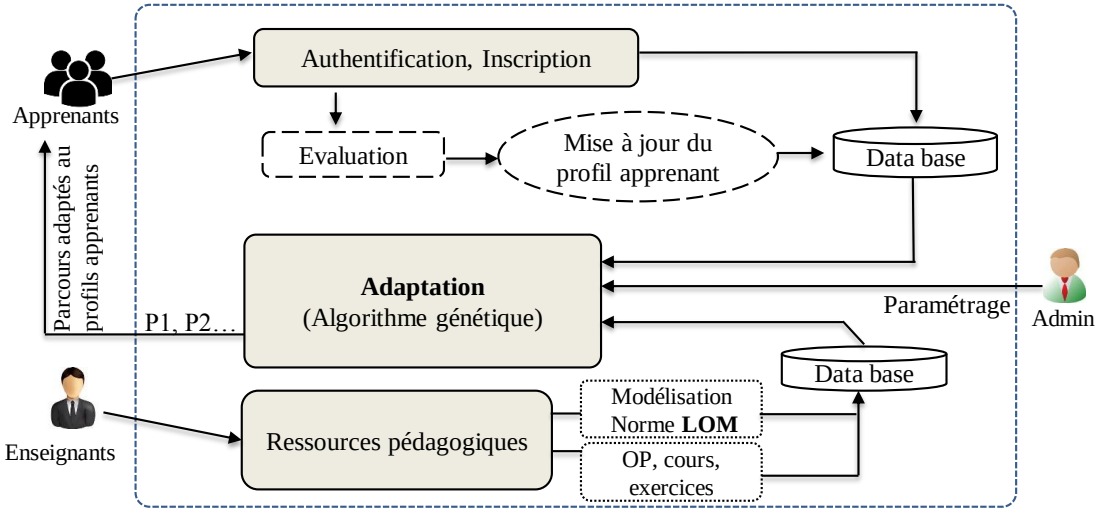

FIGURE 4.1 – Architecture générale de notre système e-learning adaptatif

Ainsi, nous nous focalisons sur l'adaptation du contenu du cours, et plus précisément à la proposition des séquences d'unités de cours convenables au profil de l'apprenant. Ce profil est obtenu selon les résultats d'une évaluation des acquis de l'apprenant. Une fois que

la définition du profile établie, nous passons à la personnalisation du contenu (ressources pédagogiques).
Pour préparer les ressources pédagogiques à l'adaptation, le système crée une fiche descriptive pour ces ressources, en respectant la norme LOM, pour les stocker dans la base de données.

1.2. La conception de notre système (Modélisation UML)

Dans notre système, nous proposons de présenter les connaissances acquises par l'apprenant sous forme d'un ensemble de concepts. Pour atteindre notre objectif, nous aurons besoin de présenter le profil de l'étudiant, l'objectif pédagogique à atteindre et les ressources pédagogique selon des formats compatibles.

Les tâches que notre système e-learning adaptatif doit assurer sont :
- La préparation de la base de données à utiliser pour rechercher le parcours adapté. Cela suppose que le système doit permettre l'ajout des ressources pédagogiques. Ces ressources peuvent appartenir à différents modules et visent à acquérir différents objectifs.
- L'adaptation de la formation au profil de l'apprenant. Le système doit pouvoir affecter à chaque apprenant, un profil associé et permettre sa mise à jour à travers l'évaluation des connaissances (les quiz).

1.2.1. Diagramme de cas d'utilisation

Le diagramme de cas d'utilisation représente la structure des grandes fonctionnalités nécessaires aux utilisateurs du système. Il donne une vue du système dans son environnement extérieur et définit la relation entre l'utilisateur et les éléments que le système met en œuvre. Les figures 4.2, 4.3 et 4.4 représentent respectivement les diagrammes de cas d'utilisation des acteurs enseignant, apprenant et administrateur.

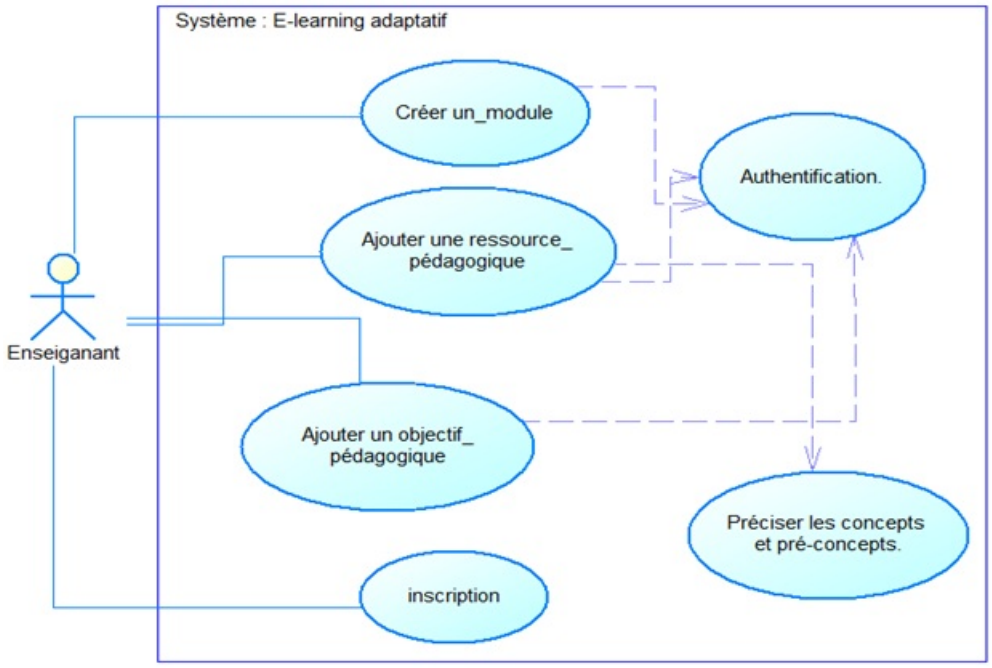

FIGURE 4.2 – Diagramme de cas d'utilisation de l'acteur enseignant

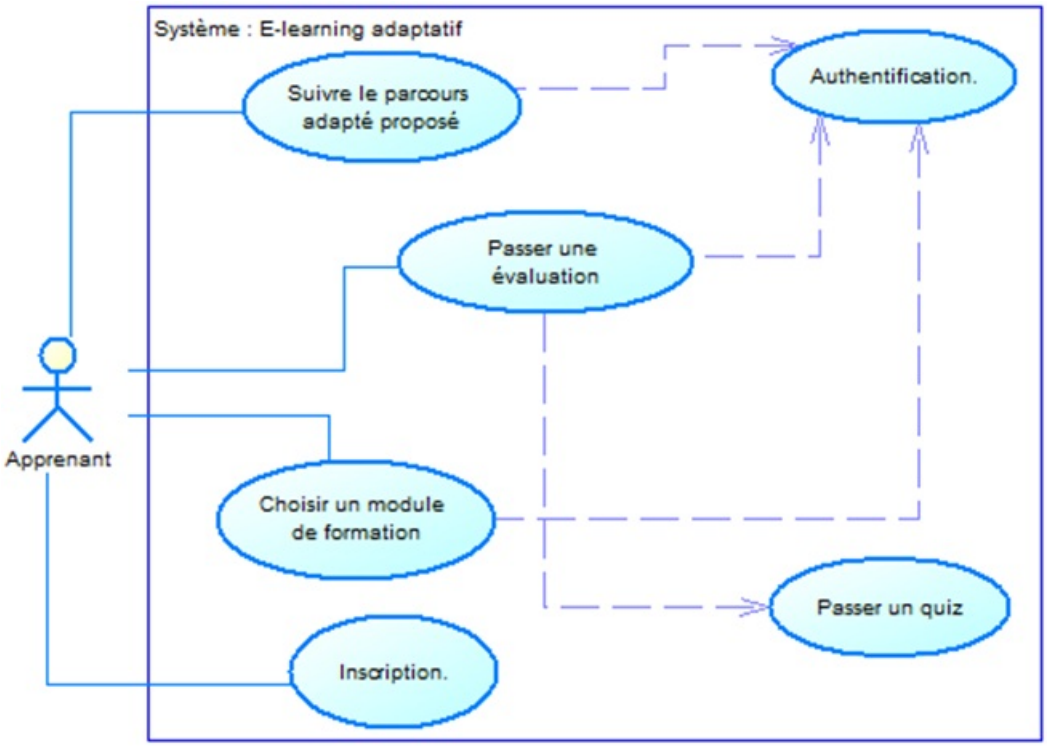

FIGURE 4.3 – Diagramme de cas d'utilisation de l'acteur apprenant

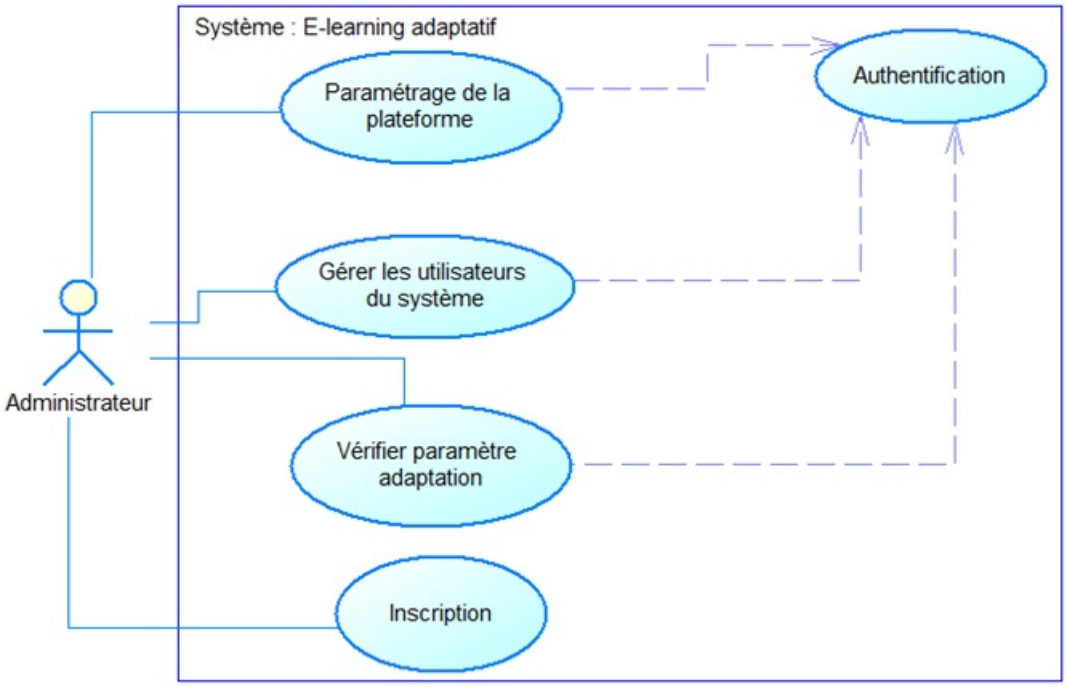

FIGURE 4.4 – Diagramme de cas d'utilisation de l'acteur administrateur

1.2.2. Diagramme de séquence

Les principales informations contenues dans un diagramme de séquence sont les messages échangés entre les lignes de vie, présentés dans un ordre chronologique. Les figures 4.5 et 4.6 représentent respectivement les diagrammes de séquences *inscription_Apprenant* et *tâches_Enseiganant*.

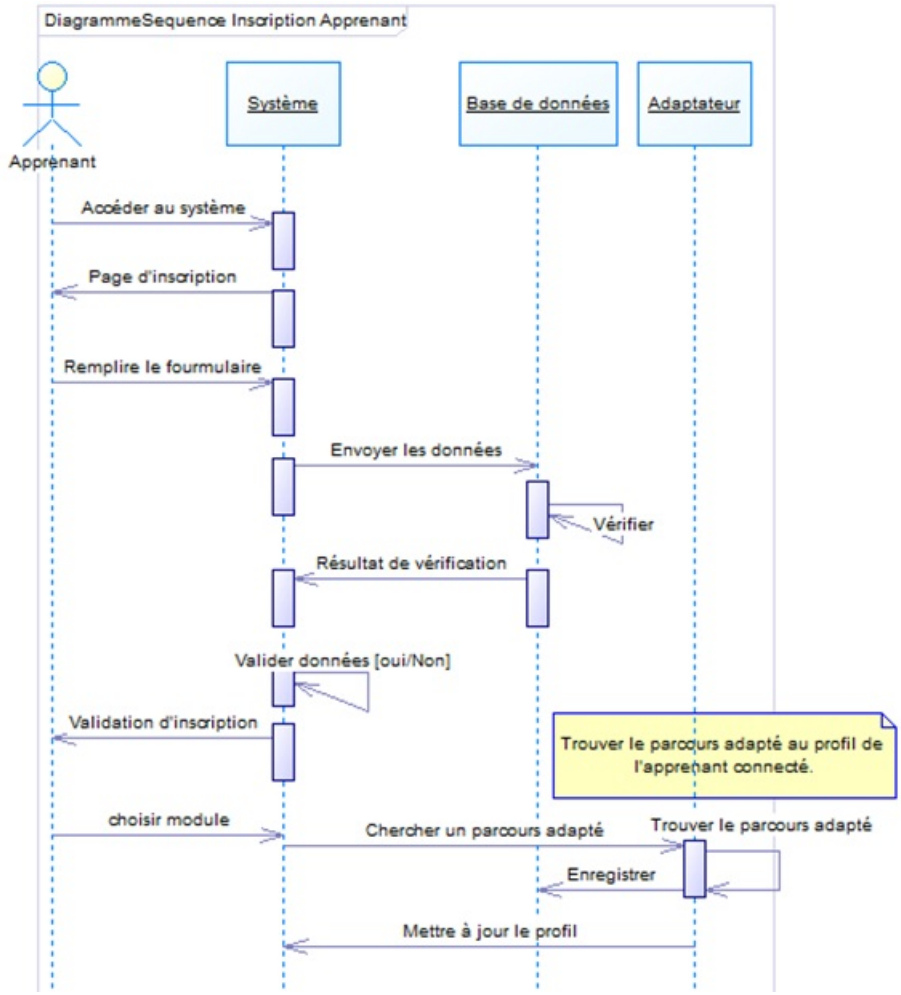

FIGURE 4.5 – Diagramme de séquence inscription_Apprenant

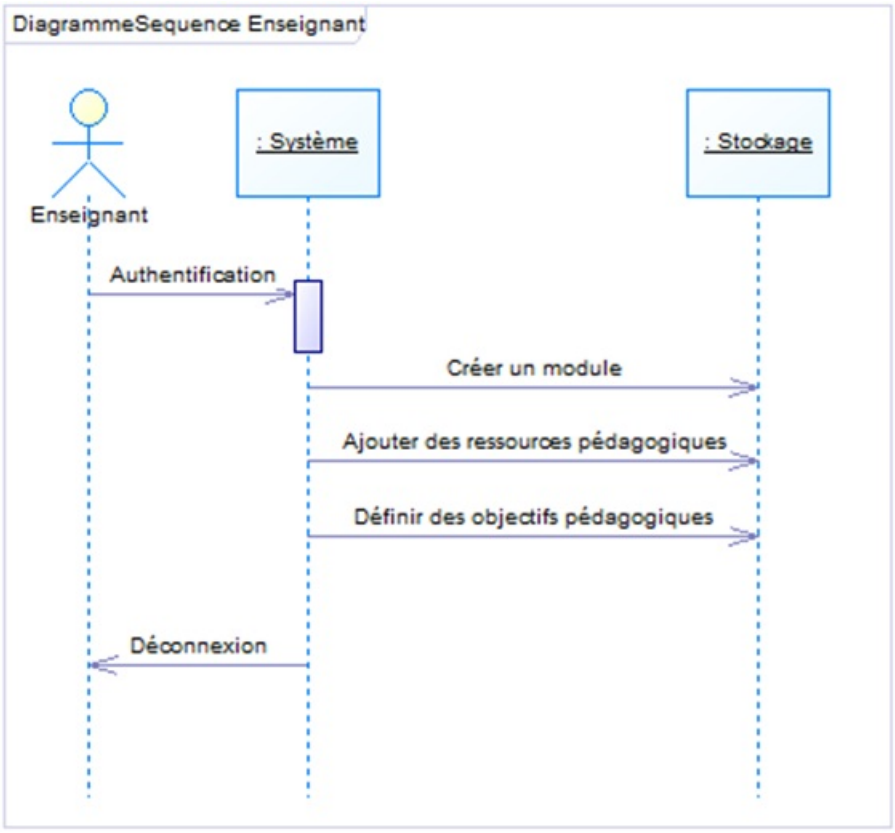

FIGURE 4.6 – Diagramme de séquence tâches_Enseiganant

1.2.3. Diagramme de classes

Le diagramme de classes est considéré comme le plus important de la modélisation orientée objet, il est le seul obligatoire lors d'une telle modélisation. Alors que le diagramme de cas d'utilisation montre un système du point de vue des acteurs, le diagramme de classes en montre la structure interne. Il permet de fournir une représentation abstraite des objets du système qui vont interagir ensemble pour réaliser les cas d'utilisation. La figure 4.7 représente le diagramme de classe de notre système.

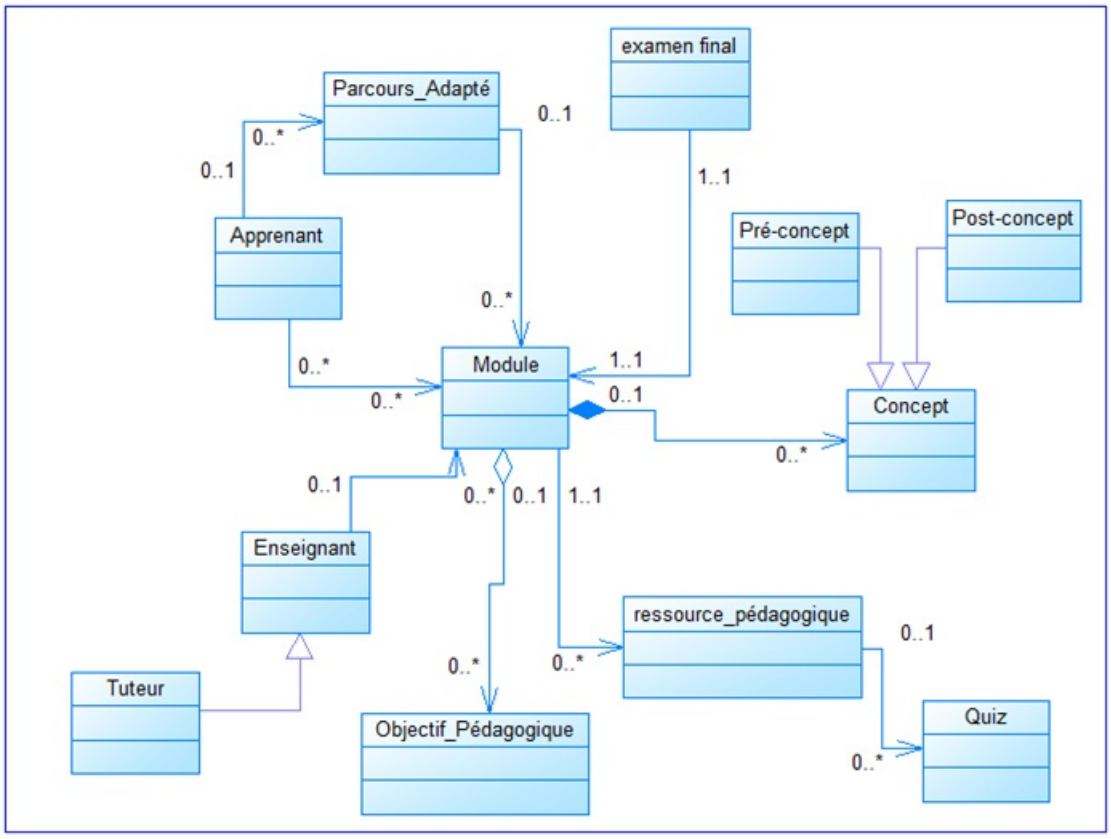

FIGURE 4.7 – Diagramme de classe de notre système e-learning adaptatif

1.3. Adéquation des algorithmes génétiques à notre approche d'adaptation

D'après l'état de l'art présenté dans le chapitre 1, les algorithmes génétiques sont capables de s'adapter à n'importe quel espace de recherche. Ils demandent une mesure de la qualité de la solution et nécessitent la définition de l'espace par un codage et des opérateurs qui lui permettent de le parcourir efficacement.

L'implémentation des algorithmes génétiques dans notre système, nécessite la définition des paramètres suivants :
– La taille du chromosome n représente le nombre de concepts du module de la formation en cours.
– La probabilité de croisement P_c utilisée est la probabilité d'acquisition des concepts pour l'apprenant.
– La probabilité de mutation P_m est la probabilité de déduction de concepts à partir d'autres pour l'apprenant.
Dans ce qui suit nous présentons les phases de l'algorithme génétique utilisé :
Initiation : La première étape est la génération aléatoire d'une population P1 de n individus. Ces individus représentent les solutions du problème. La population représente les états du profil possibles à l'issue des formations.
Fitness : Une évaluation de la fonction d'adaptation $f(s)$ de chacun des individus est ef-

fectuée. Dans notre cas on utilise la fonction cosinus comme fonction de fitness ; qui calcule la similarité entre les concepts acquis et les objectifs qui sont les individus de notre population.
Nouvelle population : Création d'une nouvelle génération $P(t+1)$ en fonction de la population $P(t)$. Cette itération de création se poursuit tant que les nouvelles solutions de $P(t+1)$ ne satisferont pas la fonction d'adaptation.
Sélection : La sélection va choisir une première population intermédiaire P_s de n solutions à partir de $P(t)$. La première population intermédiaire P_s est sélectionnée en effectuant n tirages aléatoires de solutions de $P(t)$. Le croisement fait l'échange d'informations entre deux solutions sélectionnées de P_s pour former la population P_c de n solutions. Pour sélectionner des couples, la population P_s est parcourue et chaque solution a une probabilité P_c d'être sélectionnée pour le croisement. Après la sélection, les couples choisis échangent de l'information (bits) selon l'opérateur de croisement.
Après la modification des chromosomes, les enfants forment la population $P(t+1)$. Ainsi la nouvelle génération P_2 peut recommencer le cycle, à moins que la condition d'arrêt soit satisfaite.
$C = C1, C2, \ldots Cn$ Ensemble de concepts, acquis et connaissances préliminaires de la formation à adapter.
$n =$ nombre de concepts.
$P = p1, p2, \ldots pn$ Vecteur profil où $pi = 1$, si Ci est acquis ; $pi = 0$ sinon.
$G = g1, g2, \ldots gn$ Vecteur objectif pédagogique de la formation. Avec $gi = 1$ si Ci doit être acquis, $gi = 0$ sinon.
$G(111011)$ signifie qu'à la suite de cette formation, l'apprenant doit acquérir les concepts 1, 2, 3, 5 et 6.
$P(110000)$ signifie que l'étudiant a déjà acquis les concepts 1 et 2.

L'objectif pédagogique définit par l'enseignant est reformulé à l'aide de l'ensemble des concepts pour représenter les connaissances que l'apprenant doit avoir acquis à l'issue de cette formation. Pour implémenter notre algorithme, nous avons utilisé la représentation ou le codage binaire, avec une probabilité de croisement égale à 0.7 et une probabilité de mutation égale à 0.02. C'est à dire que l'apprenant arrive à acquérir tous les concepts présents dans un cours ou à déduire quelques concepts à partir des autres concepts.

1.4. Implémentation

L'application a été conçue à l'aide d'outils gratuits sous forme de site web dynamique. Cela facilite son utilisation et réduit les finances. Pour intégrer les algorithmes génétiques au sein de notre système, nous avons utilisé le package JGAP. Ce dernier permet d'implémenter les algorithmes génétiques et la programmation génétique avec le langage de programmation JAVA. Il fournit les mécanismes génétiques de base qui peuvent être facilement employés pour résoudre des problématiques d'ordre évolutionnaires. L'utilisation des algorithmes génétiques nous a permis d'automatiser la recherche de cours adaptés à l'apprenant.

Pour comprendre l'intérêt de notre solution, imaginons un étudiant intéressé par l'apprentissage sur notre plateforme. Après son inscription, il doit passer une évaluation pour que

nous puissions définir son niveau de connaissance (concept acquis). Ensuite, notre système fournit à cet apprenant un parcours d'apprentissage personnalisé à son profil (niveau d'apprentissage). Cette adaptation s'appuie sur un algorithme génétique qui cherche le parcours le plus optimal c'est-à-dire, la liste des objectifs pédagogiques adéquats au profil de l'apprenant.

Après l'authentification de l'enseignant, la fenêtre illustrée sur la figure 4.8 s'affiche. Ce formulaire permet à l'enseignant de mettre à jour les informations de son compte.

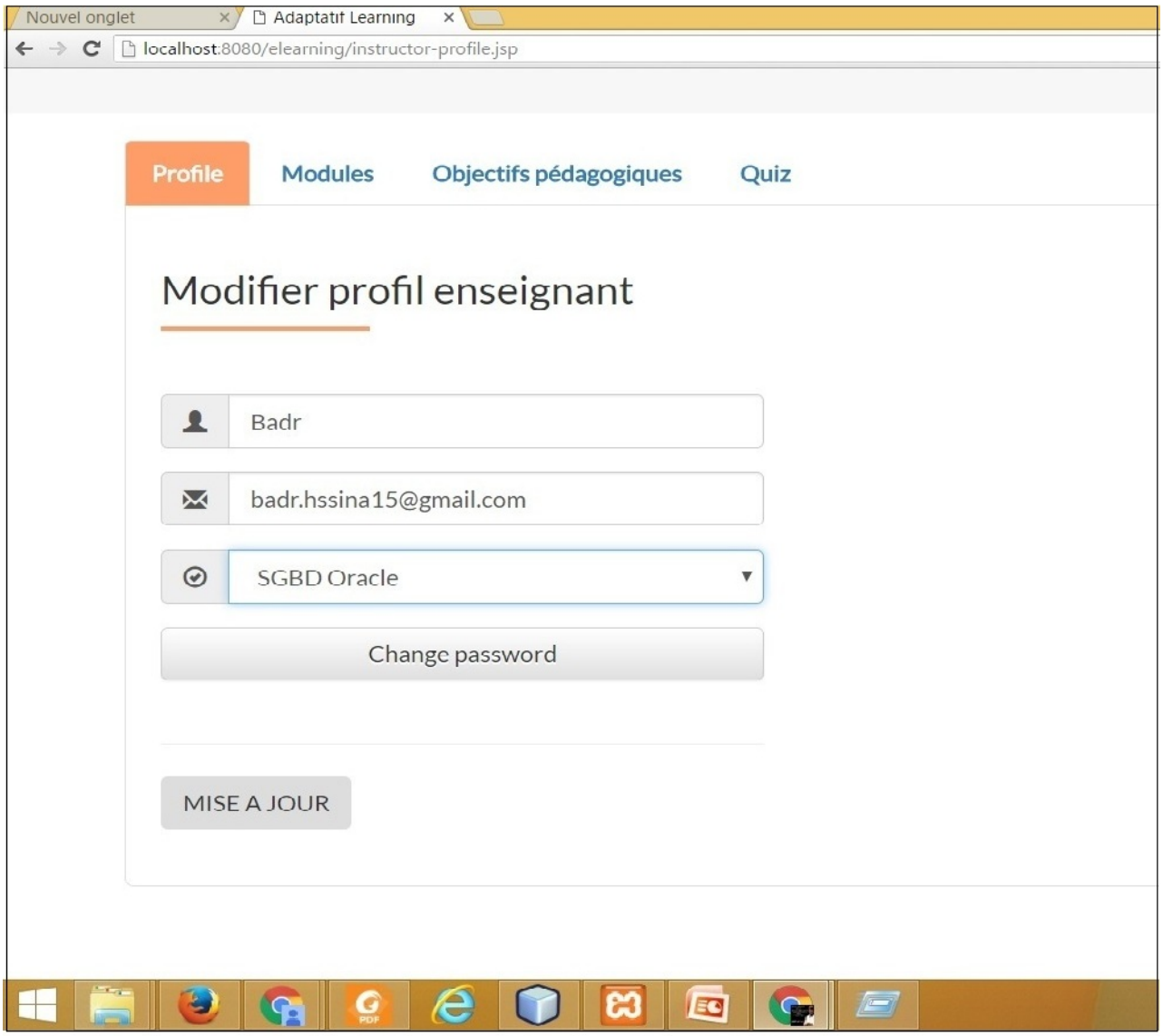

FIGURE 4.8 – Formulaire de mise à jour du profil enseignant

La fenêtre illustrée sur la figure 4.9 permet à l'enseignant d'ajouter un nouveau module, sa description et les concepts qui le constituent.

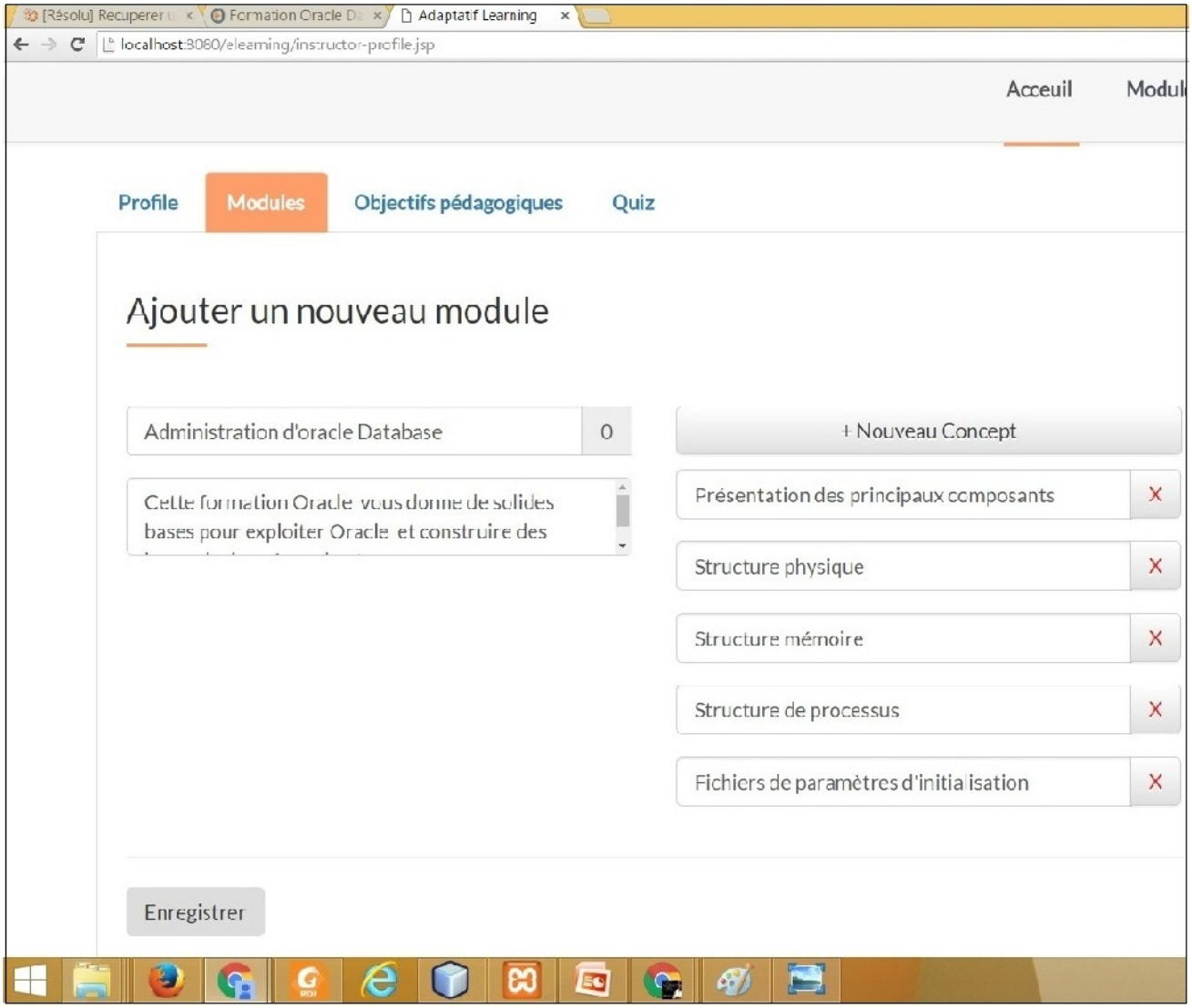

FIGURE 4.9 – Formulaire de création d'un nouveau module

La fenêtre illustrée sur la figure 4.10 permet à l'enseignant d'ajouter un nouvel objectif pédagogique, en spécifiant à quel module est lié et sur quels concepts est basé.

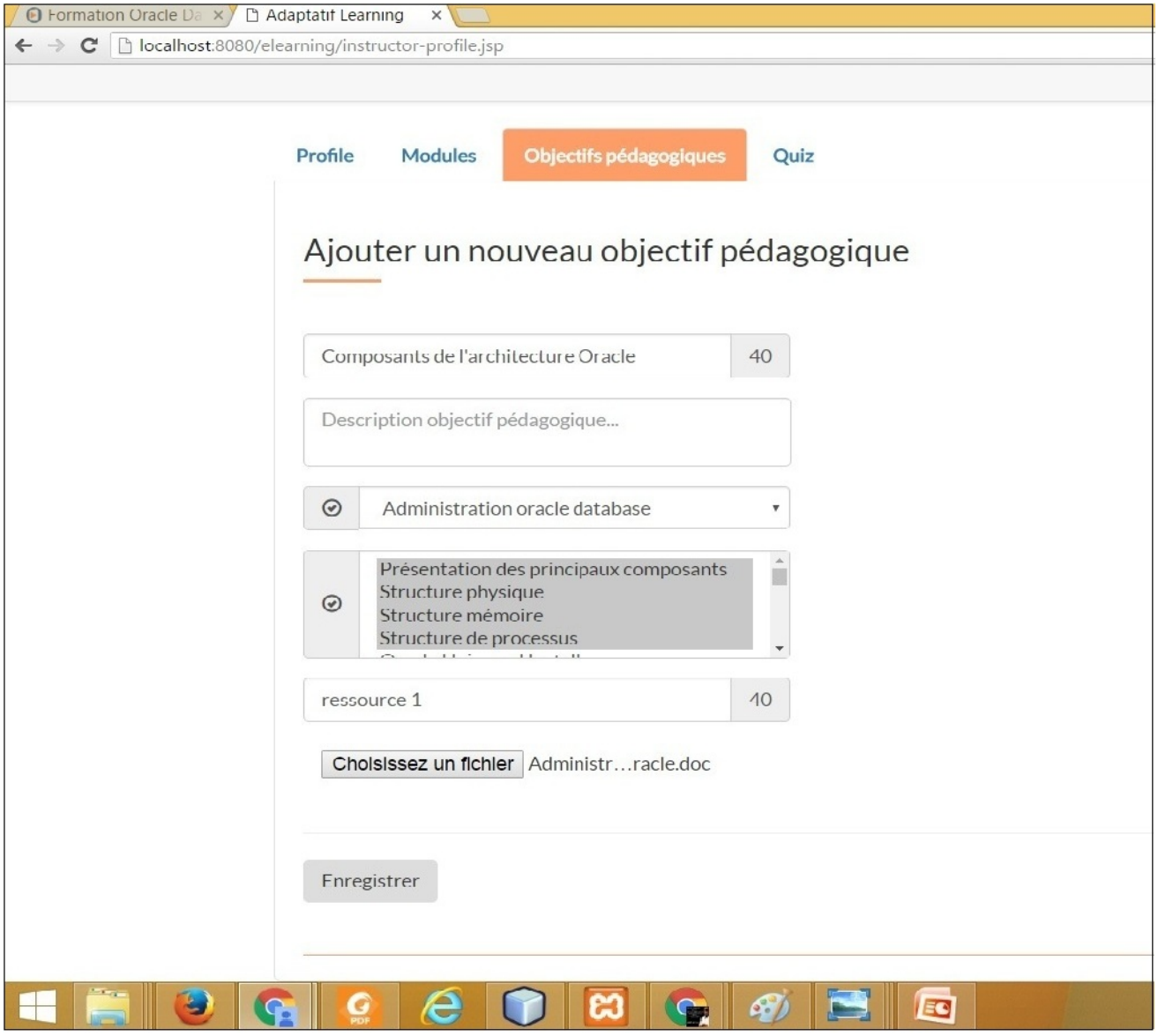

FIGURE 4.10 – Formulaire de création d'un nouvel objectif pédagogique

Par ailleurs, pour suivre une formation, l'apprenant doit avoir comme connaissances préliminaires certains concepts définis par l'enseignant (prérequis). Afin de déterminer ces derniers, un apprenant doit passer une évaluation.
La fenêtre illustrée sur la figure 4.11 permet à l'apprenant de passer une évaluation sous forme de quiz afin de construire son profil initial. Ce dernier sera constitué de concepts jugés aquis à partir des réponses correctes.

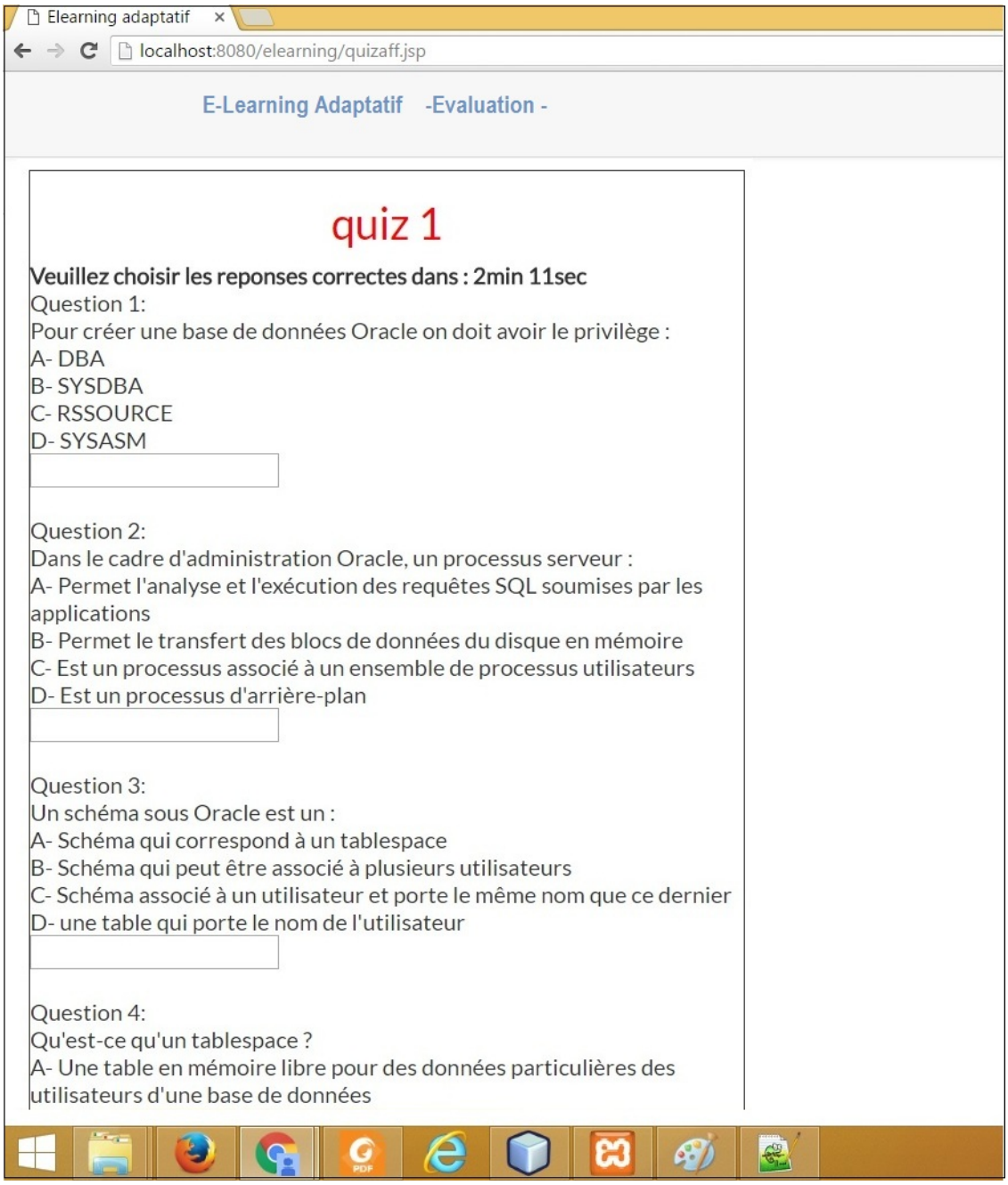

FIGURE 4.11 – Interface quiz de l'espace apprenant

L'adaptabilité est réalisée en implémentant les algorithmes génétiques, pour générer le parcours adapté à chaque apprenant.

1.5. Expérience et évaluation

Le point de départ est le profil de l'apprenant, le point d'arrivée est l'objectif pédagogique de la formation et les états intermédiaires sont les évolutions du profil après le suivi des cours disponibles.

Notre expérience est menée avec 29 apprenants.
Le module de la formation choisi est : Initiation à administration oracle.
Le nombre concept : 15
Le nombre d'objectif pédagogique : 12
Après l'exécution des algorithmes génétiques. Le système fournit le résultat illustré sur la figure 4.12. Le profil, l'objectif et la solution sont présentés sous format vectoriel. La représentation est binaire, ce qui signifie que si c'est égale à 0 alors le concept n'est pas acquis, et si c'est égal à 1 alors, le concept est acquis.

Id	Profil apprenant	Objectif choisi	Solution trouvée	Distance entre objectif et Solution
1	p(000000000000000)	G1(111100000000000)	S(110100000000000)	-0,25
2	p(100000000000000)	G1(111100000000000)	S(111100000000000)	0
3	p(110000000000000)	G1(111100000000000)	S(111111000000000)	0,5
4	p(111000000000000)	G2(111111000000000)	S(111110000000000)	-0,16
5	p(111010000000000)	G2(111111000000000)	S(111111000000000)	0
6	p(110000000000000)	G2(111111000000000)	S(111111000000000)	0
7	p(111100000000000)	G2(111111000000000)	S(111111100000000)	0,16
8	p(110100000000000)	G2(111111000000000)	S(111110100000000)	-0,16
9	p(110110000000000)	G2(111111000000000)	S(111110110000000)	-0,16
10	p(000000000000000)	G2(111111000000000)	S(111111010000000)	0,16
11	p(111000000000000)	G3(111110110100000)	S(111111111100000)	0,25
12	p(100000000000000)	G3(111110110100000)	S(111111110100000)	0,12
13	p(110000000000000)	G3(111110110100000)	S(111110111100000)	0,12
14	p(111010000000000)	G3(111110110100000)	S(111110110100000)	0
15	p(000000000000000)	G3(111110110100000)	S(111111110100000)	0,12
16	p(111100000000000)	G3(111110110100000)	S(111110111100000)	0,12
17	p(110100000000000)	G3(111110110100000)	S(111110100100000)	-0,12
18	p(100000000000000)	G4(111111111000000)	S(111111111100000)	0,11
19	p(000000000000000)	G4(111111111000000)	S(111111111000000)	0
20	p(110000000000000)	G4(111111111000000)	S(111111101100000)	-0,11
21	p(111010000000000)	G4(111111111000000)	S(111111011100000)	-0,11
22	p(111110000000000)	G4(111111111000000)	S(111111101100000)	0
23	p(110110000000000)	G4(111111111000000)	S(111111111010000)	0,11
24	p(111000000000000)	G4(111111111000000)	S(111111111011000)	0,22
45	p(111100000000000)	G4(111111111000000)	S(111111111110000)	0,22
26	p(111000000000000)	G5(111111011011000)	S(111111111011000)	0,1
27	p(100000000000000)	G5(111111011011000)	S(111111011111000)	0,1
28	p(110000000000000)	G5(111111011011000)	S(111111111111000)	0,2
29	p(111010000000000)	G5(111111011011000)	S(111111111010000)	0

FIGURE 4.12 – Les solutions trouvées pour le profil et l'objectif des 29 étudiants

La fenêtre illustrée sur la figure 4.13 affiche un exemple d'un parcours d'apprentissage adapté au profil de l'apprenant connecté.

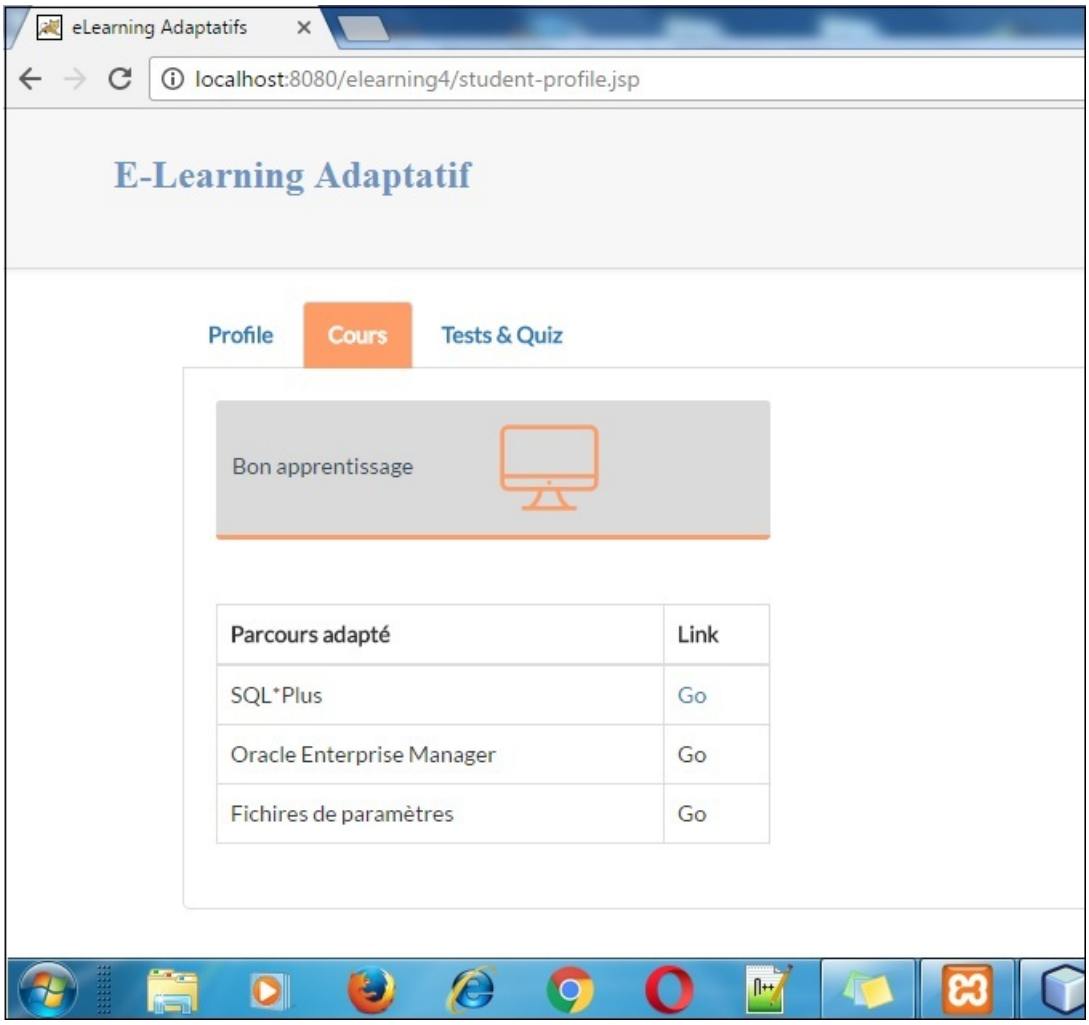

FIGURE 4.13 – Exemple d'un parcours adapté au profil de l'apprenant

Nous remarquons que tant qu'il y a une distance significative entre le profil et l'objectif, l'erreur diminue (le pourcentage de la distance entre la solution et l'objectif). C'est-à-dire, tant que le nombre de concept à acquérir est grand, tant que nous arrivons à trouver une solution plus adaptée. Si la distance entre l'objectif et le profil est plus petite la solution dépasse parfois l'objectif.
L'évaluation de notre algorithme génétique a montré sa convergence vers la solution optimal à partir de la génération 20 (4.14).

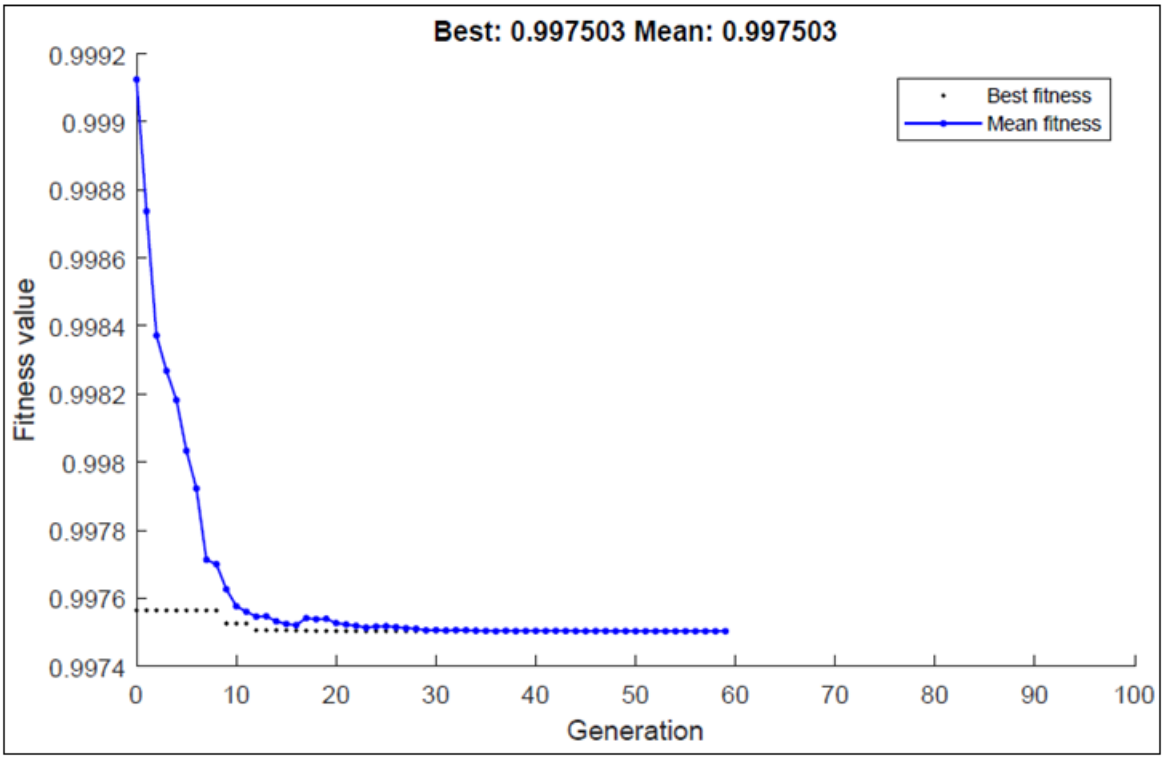

FIGURE 4.14 – La convergence de notre algorithme génétique vers la solution optimal

La figure 4.15 affiche la distance moyenne entre les individus (Objectifs pédagogiques).

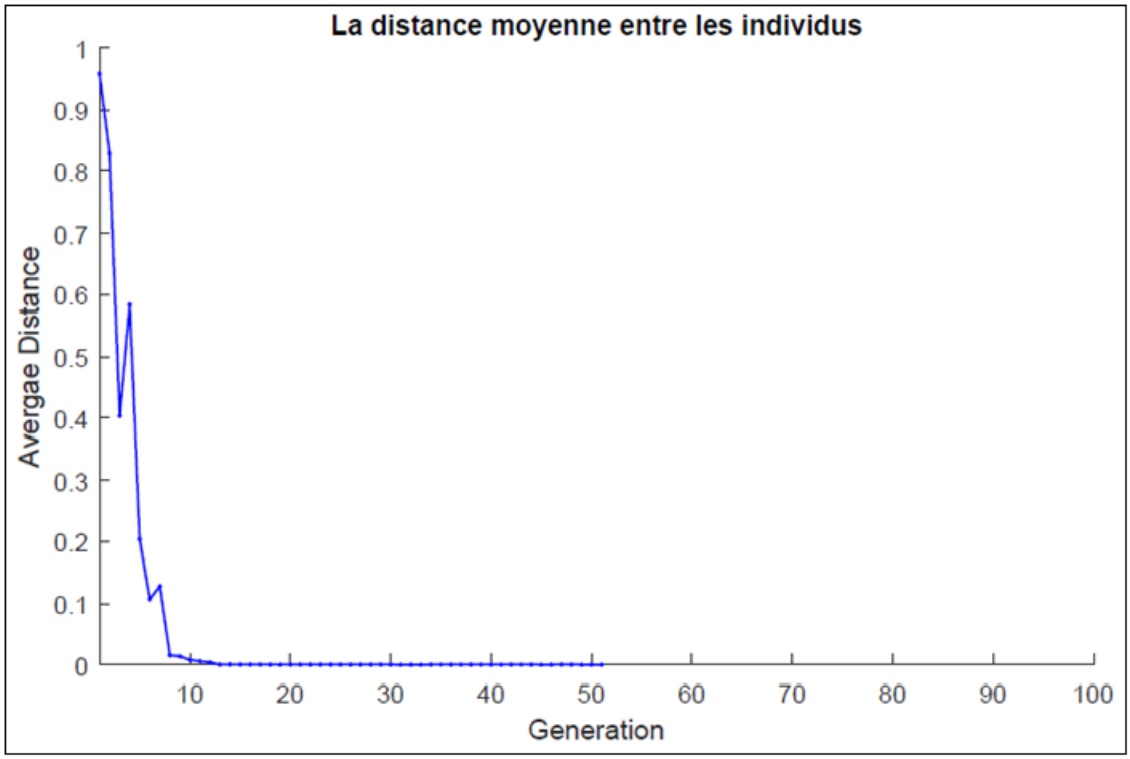

FIGURE 4.15 – La distance moyenne entre les individus

La figure 4.16 illustre les meilleurs et pires scores de la fonction de fitness de génération en génération.

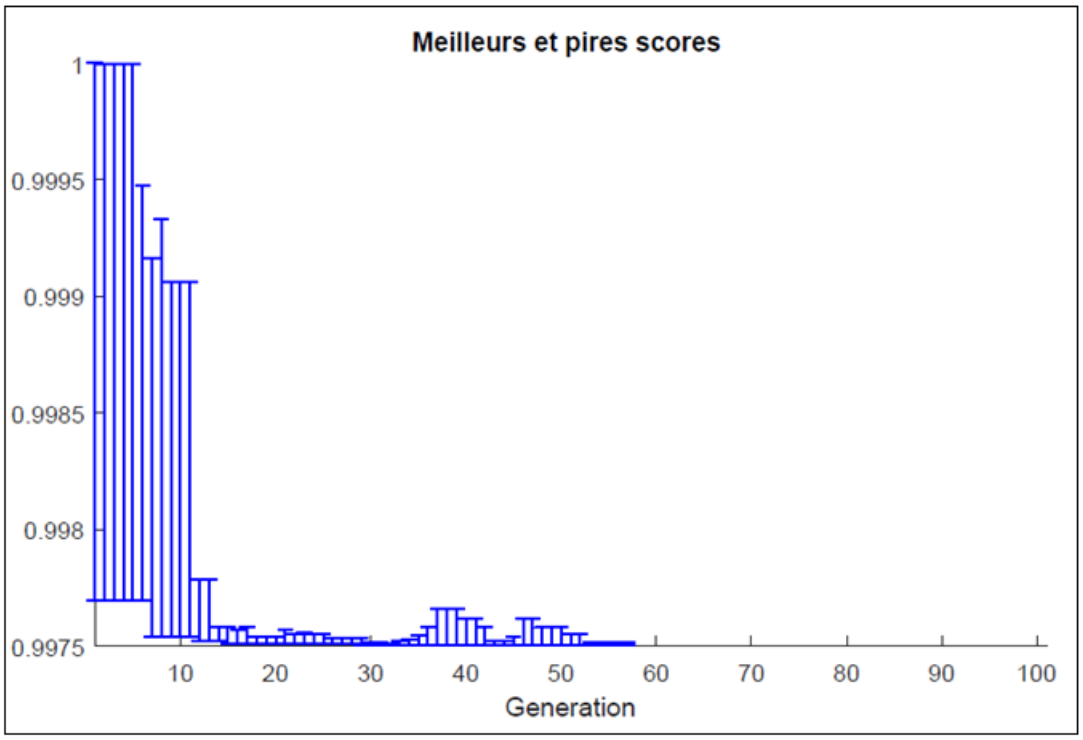

FIGURE 4.16 – Meilleurs et pires scores de la fonction de fitness

Enfin, nous estimons que notre système permet de contribuer à rendre l'apprentissage plus adaptatif, attractif et plus efficace. Dans la section suivante nous allons ajouter un aspect sémantique à cet environnement d'apprentissage en proposant un système de recommandation sémantique des documents textes aux apprenants.

2. Notre approche de recommandation basée sur le contenu et guidée par la sémantique

Nous allons introduire dans cette section notre approche hybride du calcul de la similarité sémantique, pour contribuer à la recommandation des documents textuels aux apprenants inscrits sur notre plateforme e-learning. Le but de notre approche est d'orienter les apprenants dans leurs parcours d'apprentissage, en les suggérant des ressources sur la base de leurs expériences d'apprentissage antérieures. L'évaluation de notre système de recommandation sémantique a montré des résultats très satisfaisants.

Tout d'abord, il faut dire que modèle le plus utilisé pour la représentation des documents textuels est le modèle vectoriel qui se base sur des mesures de similarités statiques. Le problème du modèle vectoriel c'est qu'il ne prend pas en considération la relation entre les composants du vecteur. C'est-à-dire, il n'utilise pas les relations sémantiques comme par exemple le sens entre les composants du vecteur. Pour combler ce manque, notre travail consiste à effectuer un enrichissement sémantique de la représentation vectorielle des documents en se basant sur l'utilisation des mesures de la similarité sémantique déjà cité dans le troisième chapitre.

Ainsi, notre contribution est un système de recommandation sémantique basé sur le contenu qui permet de retourner à partir d'un corpus, les documents qui peuvent intéressés un apprenant c'est-à-dire les documents qui sont similaires sémantiquement à un document choisi précédemment par un apprenant (figure 4.17).

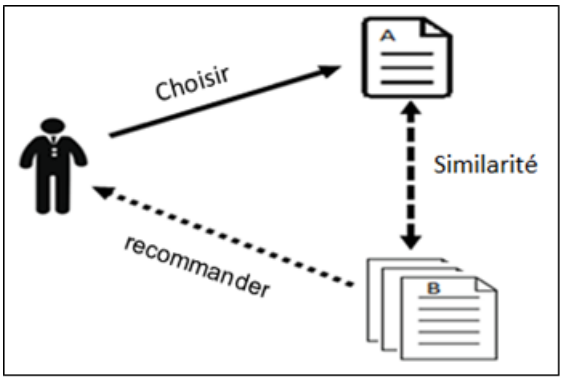

FIGURE 4.17 – Recommandation basée sur le contenu

La technique de la recommandation basée sur le contenu est fondée sur l'hypothèse suivant : si un utilisateur a apprécié un item, les items du contenu similaire seront également appréciés par le même utilisateur. Cette technique est basée sur l'analyse du contenu des items précédemment consultés par les utilisateurs et les items qui n'ont pas encore été consultés.

Dans cette partie, nous présentons les différentes étapes suivies dans notre approche afin de réaliser cette recommandation en se basant sur des mesures de similarité sémantique.

2.1. Description de notre approche de recommandation sémantique

L'objectif de notre travail est de réaliser un système de recommandation sémantique qui permet de chercher dans un ensemble de documents (corpus), le ou les documents qui peuvent intéresser un apprenant.

La figure 4.18 illustre l'architecture de notre système. La première phase est l'indexation puis l'enrichissement sémantique à partir de la base de données lexicale WordNet et après la pondération TF×IDF pour une représentation des documents par le modèle vectoriel. Toutes ces opérations doivent être validées avant que nous appliquons notre approche du calcul de la similarité. Enfin, nous recommandons aux apprenants les documents similaires à leurs choix antérieurs en se basant sur les valeurs de similarité obtenues.

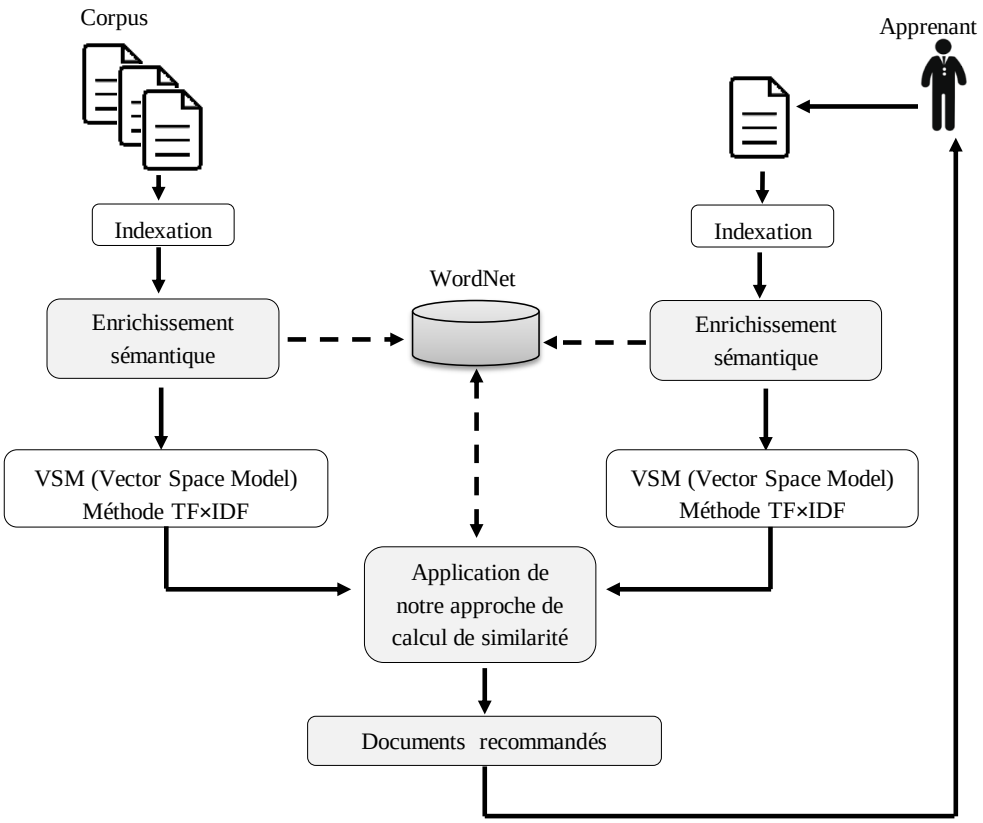

FIGURE 4.18 – Notre approche de calcul de similarité entre les documents texte pour une recommandation sémantique

L'enrichissement sémantique est une étape clé dans notre travail. Nous introduisons une relation sémantique qui est la synonymie en se basant sur WORDNET. L'idée est d'enrichir la représentation vectorielle par les synonymes. Nous recherchons des synonymes pour chaque mot de la représentation vectorielle du document, après que nous ayons cherché à savoir si ces synonymes sont dans le document, puis si elle est trouvée, nous ajoutons sa fréquence dans le vecteur représentant le document.

2.2. Méthode proposée pour l'extraction des termes à partir du corpus

Pour cela, il est nécessaire de disposer d'un ensemble de documents (un corpus) qui contient les documents que nous voulons chercher. Notre corpus est constitué d'un ensemble de documents écrits en anglais traitant plusieurs domaines comme algorithmique, réseaux informatiques, système informatique, etc.

L'indexation de chaque document du corpus est une phase très importante. En effet une bonne indexation donnera des bons résultats lors de la recommandation.

Les documents qui peuvent être recommandés aux apprenants sont représentés sous forme d'un vecteur :

$$\overrightarrow{d} = (w_1, w_2, ..., w_n) \tag{4.1}$$

de n composantes, où chaque composante représente le poids du mot dans le document (sa fréquence).
Soit $C = \{d_1,d_2,...,d_n\}$ dénotant un corpus de n documents et $T= \{t_1,t_2,...,t_n\}$ est le dictionnaire ou l'ensemble des mots du corpus. T est obtenu en appliquant des opérations de Traitement automatique du langage naturel (TALN), comme la tokenisation, la racinisation et l'élimination des mots vides. La figure 4.19 illustre notre solution proposée pour extraire les termes du corpus.

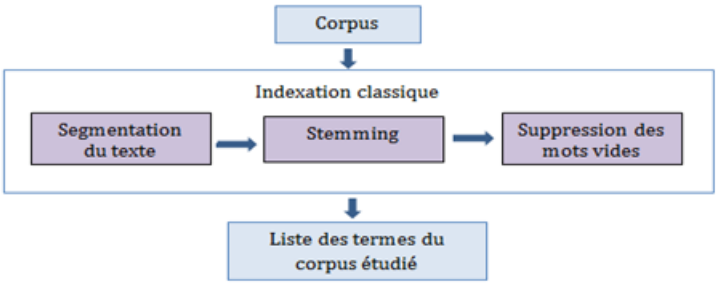

FIGURE 4.19 – Méthode proposée pour extraire les termes du corpus

Dans cette étape nous présentons chaque document par les termes qui le compose, cette étape est constituée des sous étapes suivantes :

◇ **Segmmentation ou Tokenization :** Cette étape permet d'extraire les mots à partir du document en jouant sur des séparateurs. Un séparateur peut être un "blanc", une ponctuation, une parenthèse, etc. Un token (élément en français) est une séquence de caractères compris entre deux séparateurs. De plus, le traitement automatique des langues nécessite une segmentation (tokenisation) du texte comme un des premiers pas de l'analyse.

◇ **Lemmatisation :** Une phase préalable importante à l'indexation des documents est la lemmatisation des mots. Le terme lemmatisation fait référence à la réduction des mots à leur forme canonique (racine) de sorte que, par exemple, différentes formes grammaticales ou déclinaisons de verbes soient identifiées et indexées (recensées) comme une occurrence du même mot. Par exemple, le processus de lemmatisation garantit que les mots "assimiler" et "assimilé" seront reconnu par le programme comme un seul et même mot.

La lemmatisation d'une forme d'un mot consiste à en prendre sa forme canonique. Celle-ci est dénie comme suit : s'il s'agit d'un verbe on le prend à l'infinitif, pour les mots on utilise le masculin singulier.
Par exemple l'adjectif petit existe sous quatre formes : petit, petite, petits et petites. La forme canonique (lemmatisation) de tous ces mots est petit. Ainsi, Il existe plusieurs formes

du verbe être : suis, es, est, sommes, était, été, etc. La forme canonique de fûmes est être.

◇ **Racinisation(Stemming) :** La racine d'un mot correspond à la partie du mot restante une fois que l'on a supprimé son (ses) préfixe(s) et suffixe(s), c'est son radical. Contrairement au lemme qui correspond à un mot réel de la langue, la racine ne correspond généralement pas à un mot réel. Par exemple, le mot marcher a pour radical march qui ne correspond pas à un mot réel.

◇ **Elimination des majuscules :** Une opération importante est l'élimination des majuscules c'est-à-dire nous procédons à une transformation de tous les mots en minuscules. En effet, le mot BOY et le mot boy vont être considérer différent alors qu'ils ont le même sens donc nous devons transformer les majuscules en minuscule.

◇ **Elimination des mots vides :** Les mots vides (stop words en anglais) sont les mots qui se répètent fréquemment dans tous les documents et qui n'ont aucun pouvoir discriminant lors du processus de la catégorisation de texte. La listes des mots vides contient les pronoms personnels, les prépositions, les articles, etc. exemple he, she, him, and, also, etc.
Si le mot apparu est un mot vide alors le système doit le supprimer. C'est la raison pour laquelle dans ce travail nous avons un document qui contient tous les mots vides de la langue anglaise et lors de l'indexation si le mot du document est un mot vide nous ne le prenons pas en considération.

2.3. Enrichissement sémantique de la représentation des documents

L'indexation sémantique des documents est née du problème de l'ambiguïté des mots de la langue naturelle utilisés pour l'indexation classique. L'indexation sémantique a pour objet de représenter les documents texte par le sens des mots, permettant ainsi de lever toute ambigüité, ce qui a pour conséquence d'améliorer les résultats de la recherche (dans notre cas les résultats de la recommandation).

Pour retrouver les sens corrects des mots dans un document, l'indexation sémantique requiert des techniques de désambiguïsation des sens des mots. Ces techniques se basent sur l'utilisation de ressources, telles que les corpus d'apprentissage, les ressources terminologiques et les ontologies.

Parmi les ressources ontologiques les plus exploités dans la désambiguïsation des mots, nous retrouvons WordNet. Il est à la base de nombreux travaux et projets récents en indexation sémantique qui visent l'accès aux textes par le sens.

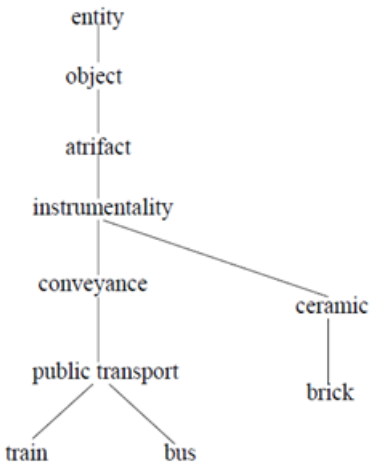

FIGURE 4.20 – Un fragment de la hiérarchie is-a de WordNet

WordNet peut être également considéré comme une ontologie pour les termes de la langue anglaise. Il contient environ 100 000 termes, organisés en hiérarchies taxonomiques. Les noms, verbes, adjectifs et adverbes sont regroupés en ensembles de synonymes (synsets). Les synsets sont liés à d'autres synsets plus ou moins élevés dans la hiérarchie par différents types de relations. Les relations les plus courantes sont l'hyponyme / Hypernym (c'est-à-dire les relations is-A) et les relations Meronym / Holonym (c'est-à-dire les relations Part-Of). La figure 4.20 ci-dessus illustre un fragment de la hiérarchie is-A de WordNet.

Nous avons exploitée dans ce travail la relation de synonymie pour enrichir la représentation vectorielle de nos documents en se basant sur WordNet. Tout d'abord, nous cherchons pour chaque mot de la représentation vectorielle du document les synonymes associés et après nous vérifions si ces synonymes sont dans le document, puis pour chaque synonyme trouvé nous ajoutons sa fréquence dans le document au poids du mot et ainsi de suite jusqu'à l'épuisement des mots.

2.4. Fréquence des mots dans les documents

Nous avons utilisé le VSM (Modèle d'Espace Vectoriel) pour la représentation des documents de notre corpus. En fait, le modèle vectoriel est une représentation mathématique du contenu d'un document, selon une approche algébrique. L'ensemble de représentation des documents est un vocabulaire comprenant des termes d'indexation. Ces derniers sont typiquement les mots les plus significatifs du corpus (noms communs, noms propres, adjectifs...). Chaque contenu est ainsi représenté par un vecteur $\overrightarrow{v}$, dont la dimension correspond à la taille du vocabulaire. Chaque élément v_i du vecteur v représente le poids du mot i dans le document. La méthode de pondération que nous avons utilisée dans notre approche est TF×IDF.
La matrice de la figure 4.21 représente la matrice des pondérations TF×IDF représentant un extrait du corpus qui contient 11 documents texte avec 20 termes.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
Action	0,0132	0,0091	0,0247	0,0206	0,0030	0,0292	0,0359	0,0088	0,0000	0,0351	0,0118
alternative	0,0740	0,0000	0,0231	0,0290	0,0340	0,0000	0,1010	0,0987	0,0083	0,0000	0,0111
Circuit	0,0196	0,0068	0,0061	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0029
Computer	0,0346	0,0024	0,0043	0,0027	0,0000	0,0000	0,0000	0,0277	0,0000	0,0000	0,0000
Device	0,0564	0,0292	0,0176	0,0221	0,0130	0,0313	0,0000	0,0564	0,0000	0,0125	0,0168
document	0,1198	0,0118	0,0107	0,0201	0,0039	0,0951	0,0000	0,0114	0,0038	0,0000	0,0051
Electronic	0,1851	0,0255	0,0231	0,0436	0,0085	0,0411	0,0000	0,0247	0,0333	0,0165	0,0000
Graph	0,1562	0,0000	0,0325	0,0408	0,0239	0,0579	0,0000	0,0347	0,0000	0,0000	0,0000
hardware	0,0439	0,0076	0,0412	0,0172	0,0101	0,0244	0,0000	0,0000	0,0049	0,0098	0,0066
Internet	0,1562	0,0180	0,0325	0,0408	0,0359	0,0579	0,0000	0,0694	0,0117	0,0000	0,0155
interpretation	0,0564	0,0195	0,0353	0,0111	0,0000	0,0313	0,0000	0,0376	0,0063	0,0251	0,0084
Keyboard	0,1562	0,0180	0,0325	0,0204	0,0120	0,1157	0,0000	0,0694	0,0000	0,0231	0,0000
Metal	0,2591	0,0255	0,0463	0,0436	0,0425	0,2057	0,0000	0,0247	0,0000	0,0000	0,0000
program	0,0218	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0010	0,0019	0,0039
Result	0,1129	0,0097	0,0176	0,0111	0,0130	0,0313	0,0000	0,0188	0,0127	0,0125	0,0168
software	0,1562	0,0539	0,0651	0,0817	0,0359	0,2893	0,0000	0,1389	0,0117	0,0000	0,0777
storage	0,0370	0,0000	0,0000	0,0000	0,0085	0,0823	0,0000	0,0987	0,0166	0,0165	0,0442
System	0,0658	0,0136	0,0165	0,0052	0,0091	0,0292	0,0000	0,0088	0,0030	0,0058	0,0039
Treatment	0,0083	0,0000	0,0000	0,0000	0,0005	0,0023	0,0000	0,0028	0,0009	0,0018	0,0006
writing	0,3645	0,0180	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0155

FIGURE 4.21 – La matrice représentant la pondération des termes avec la méthode TF×IDF d'un extrait de corpus

La mesure TF×IDF est une mesure statistique qui permet d'évaluer l'importance d'un mot dans un document faisant partie d'un corpus.

◇ **Exemple de prétraitement :**
Supposons qu'après la segmentation d'un corpus, nous aurons la liste des termes suivante :
- $S = \{master, learner, material, mathematical, Physics, science, natural, phenomena\}$
Supposons que ce corpus contient 4 documents :
- $D1 = \{natural, natural, master, phenomena, science, learner, learner, \}$
- $D2 = \{phenomena, science, material\}$
- $D3 = \{natural, science, Physics, mathematical, mathematical, phenomena, phenomena\}$
- $D4 = \{master, master, master, phenomena\}$
Le tableau 4.1 illustre le poids des mots dans chaque document (la représentation vectorielle) en utilisant la pondération TF×IDF :

$$w_{x,y} = tf_{x,y} \times \log(\frac{N}{df_x}), \tag{4.2}$$

avec $tf_{x,y}$ est la fréquence du mot x dans le document y et df_x est le nombre de document contenant x et N est le nombre total de documents dans le corpus. Ainsi, un terme qui a une valeur de $w_{x,y}$ élevée doit être à la fois important dans ce document, et aussi il doit apparaître peu dans les autres documents.

	Document1	Document2	Document3	Document4
master	0.099	0	0	0.519
learner	0.285	0	0	0
mathematical	0	0	0.285	0
physics	0	0	0.142	0
science	1.143	0.333	1.143	0
natural	0.285	0	0.099	0
phenomena	1.143	0.333	0.285	0.250
material	0	0.333	0	0

TABLE 4.1 – La représentation vectorielle des documents après la phase du prétraitement

2.5. Comparaison des différentes mesures de similarité sémantique utilisées dans l'analyse textuelle

Une mesure de similarité sémantique entre des documents est un concept selon lequel un ensemble de métriques sont donnés en se basant sur la similitude de leur signification (contenu sémantique).
Plus précisément, cela peut être réalisé en définissant une similitude topologique, par exemple, en utilisant des ontologies pour définir une distance entre les mots, ou en définissant une similarité statistique, par exemple le modèle d'espace vectoriel pour trouver une corrélation entre les termes et les contextes à partir d'un corpus.
Ces mesures se servent de la structure hiérarchique de l'ontologie (WordNet) pour déterminer la similarité sémantique entre les concepts. Compte tenu de deux mots, leur similarité peut être estimée à partir de leur position relative dans la hiérarchie de WordNet.

Le tableau 4.2 illustre les mesures de similarité de Wu & Palmer [200], Resnik [193], Jiang Conrath [204] et Lin [203] calculées en se basant sur WordNet entre deux mots extraits de notre corpus :

Terme1	Terme2	Wu&Palmer	Resnik	JiangConrath	Lin
Treatment	operations	0.6315	2.6043	0.0695	0.2660
system	processing	0.4615	0.6143	0.0662	0.0752
computer	equipment	0.7777	3.4450	0.1128	0.4373
Electronic	automatic	0.0	0.0	0.0	0.0
Device	screen	0.75	2.4933	0.1048	0.3433
Keyboard	computer	0.8421	4.3675	0.0822	0.4181
Metal	magnetic	0.0	0.0	0.0	0.0
interpretation	implementation	0.4444	1.7798	0.0595	0.1748
Processor	circuit	0.2105	0.0	0.0475	0.0
Memory	storage	0.4210	1.7798	0.0675	0.1938
Program	software	0.3529	0.7794	0.0615	0.0875
Text	document	0.8	4.6933	0.3750	0.7787
Chart	graph	0.8333	6.8174	0.1817	0.6909
Sequence	Action	0.4	0.7794	0.1029	0.1383
Value	Result	0.2857	0.0	0.0722	0.0
Assigning	writing	0.6315	2.6043	0.0674	0.2599
Sequential	alternative	0.0	0.0	0.0	0.0
code	program	0.4285	0.7794	0.0722	0.1012
network	Internet	0.2352	0.0	0.0526	0.0
Gateway	hardware	0.6315	2.4933	0.0573	0.2222

TABLE 4.2 – Mesures de similarité sémantique selon WordNet entre deux mots extraits de notre corpus

Pour choisir la meilleure mesure nous allons calculer la similarité sémantique entre deux documents identiques (ce calcul est répété sur plusieurs documents) puis nous choisissons la mesure qui donne des bons résultats, le tableau 4.3 illustre les résultats obtenus.

	Similarité	Temps en milliseconde
Resnik	0.15	1461
Wu&Palmer	0.34	1399
Lin	0.21	1445
Jian Conrath	0.13	1492

TABLE 4.3 – Comparaison des différentes méthodes de calculs de similarité en termes de temps d'exécution

En se basant sur les résultats de cette comparaison, nous remarquons que la meilleure mesure est celle de wu&Palmer qui donne la plus grande similarité avec un temps d'exécution minimale. Par conséquent, nous choisissons d'utiliser la méthode de Wu&Palmer dans notre démarche proposée pour le calcul de la similarité entre deux documents.

2.6. La démarche proposée pour le calcul de similarité entre deux documents

Après l'indexation de chaque document nous pouvons calculer la similarité sémantique entre deux documents en se basant sur les approches précédemment citées. Traditionnellement, la similarité entre deux documents d et q est calculée selon le modèle d'espace vectoriel en tant que cosinus du produit interne entre leurs vecteurs de document $\overrightarrow{d} = (d_1, d_2, ..., d_n)$ et $\overrightarrow{q} = (q_1, q_2, ..., q_n)$:

$$sim(\overrightarrow{d}, \overrightarrow{q}) = \frac{\sum_{i=1}^{n} d_i \times q_i}{\sqrt{\sum_{i=1}^{n} d_i^2} \times \sqrt{\sum_{i=1}^{n} q_i^2}}, \tag{4.3}$$

où q_i et d_i sont les poids des mots dans les deux documents. Ce modèle est également connu sous le nom de *modèle sac des mots*.

Ce modèle présente quelques limites, par exemple l'absence de termes communs dans deux documents ne signifie pas nécessairement que les documents ne sont pas similaires. Des concepts semblables sémantiquement peuvent être exprimés avec des mots différents dans les documents, alors la comparaison directe par un vecteur basée sur les mots n'est pas efficace. Par exemple, le modèle vectoriel ne reconnaîtra pas les synonymes (par exemple "professeur","enseignant"). Ainsi, la caractéristique principale de notre approche est d'enrichir la représentation vectorielle par les synonymes.

Pour toutes ces raisons, notre approche vient pour surmonter ce problème en découvrant les termes sémantiquement similaires dans les documents et en utilisant la mesure de similarité de wu&palmer. La formule que nous allons utiliser pour le calcul de la similarité entre les documents était utilisée dans les travaux menés par Giannis Varelas (*Méthodes de similarité sémantique dans WordNet et leur application à la récupération d'informations sur le Web*)[208].

Notre approche pour calculer la similarité sémantique entre deux documents ($\overrightarrow{d}$ et $\overrightarrow{q}$), est défini par la formule suivante :

$$sim(\overrightarrow{d}, \overrightarrow{q}) = \frac{\sum_i \sum_j d_i \times q_j \times sim(c_i, c_j)}{\sum_i \sum_j d_i \times q_j}, \tag{4.4}$$

où :
i : représente l'indice du concept c_i du document d.
j : représente l'indice du concept c_j du document q.
d_i : est le poids du concept c_i dans le document d.
q_j : est celui du concept c_j dans le document q.
sim(c_i,c_j) : est la similarité sémantique entre les deux concepts c_i et c_j, calculée à partir de la mesure de wu&Palmer.
Sachant que :

$$sim(c_i, c_j) = \frac{2 \times profondeur(c)}{profondeur(c_i) + profondeur(c_j)}, \tag{4.5}$$

où :
- c est le concept le plus spécifique qui subsume c_i et c_j dans la hiérarchie Top-Level dans WordNet.
- Profondeur(c) : est le nombre d'arcs entre la racine Top-Level et le concept c dans word-Net.
- Profondeur(c_i) : est le nombre d'arcs entre la racine Top-Level et le concept c_i dans word-Net.
- Profondeur(c_j) : est le nombre d'arcs entre la racine Top-Level et le concept c_j dans word-Net.

2.7. La phase de la recommandation

La recommandation est la dernière étape de notre approche qui se base sur les résultats du calcul de la similarité sémantique. Dans notre travail de la réalisation d'un système de recommandation sémantique, le but est de retourner à partir d'une collection des documents (Corpus), les documents qui sont pertinents (similaires) par rapport au choix antérieur de l'apprenant.
Les documents retournés sont classés par ordre décroissant de la similarité sémantique, puis nous appliquons un seuil de recommandation à cette similarité. Si la similarité du document est supérieur ou égale à notre seuil fixé alors ce document sera retourné par le système sinon il n'est pas pertinent alors il sera rejeté par le système.

2.8. Expérience et évaluation de notre approche

2.8.1. Evaluation de notre système : temps d'exécution

Dans cette section nous présentons une comapraison des résultats pratiques de notre approche avec des approches déjà existées comme Wu & Palmer et Lin.
Le tableau 4.4 illustre les résultats de la similarité obtenus entre deux documents identiques et le temps d'exécution nécessaire pour le calcul.

	Similarité	Temps (msec)
Notre approche avec wu & palmer	0.76	2329
wu & palmer	0.34	1399
Notre approche avec Lin	0.56	2469
Lin	0.21	1445

TABLE 4.4 – Similarité et temps d'exécution de notre approche comparé avec Wu&Palmer et Lin

D'après le tableau 4.4 ci-dessus nous remarquons que notre approche avec Wu & palmer fournit de meilleurs résultats en termes du temps d'exécution qu'est inférieure à celui de l'approche de Lin. De plus, elle fournit une similarité plus grande que les autre approches

puisque nous calculons la similarité entre un document et lui-même.

Nous pouvons conclure d'après ces résultats pratiques que notre approche avec Wu & palmer est la plus efficace pour notre système de recommandation car elle permet de nous retourner rapidement les documents similaires sémantiquement à un autre document.

La figure 4.22 illustre les résultats obtenus en appliquant notre approche pour calculer la similarité sémantique entre un document et un corpus qui contient à chaque fois un nombre variable de documents. Nous calculons à chaque fois le temps nécessaire pour trouver la similarité entre un document et tous les documents du corpus.

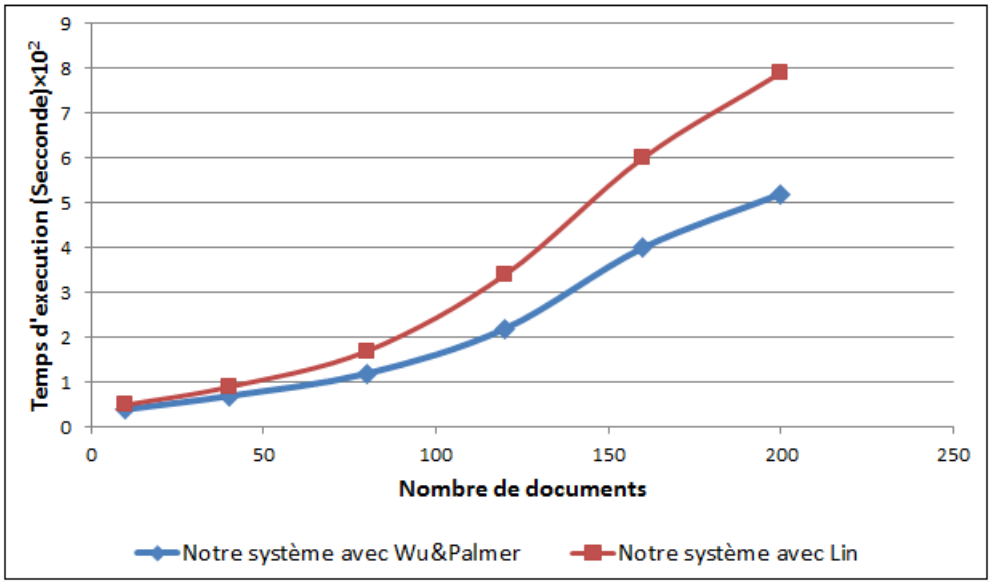

FIGURE 4.22 – Temps d'exécution en utilisant notre approche avec wu&Palmer et Lin

Nous remarquons que pour les deux cas de notre comparaison, soit avec Wu&Palmer ou avec Lin, si le nombre de documents augmente alors le temps d'exécution aussi augmente. Ainsi, avec l'approche de Wu & Palmer nous avons besoin d'un temps d'exécution moins important que celui nécessaire avec l'approche de Lin.

2.8.2. Evaluation de notre système : Rappel et Précision

Nous avons expliqué dans le troisième chapitre que la qualité d'un système doit être mesurée en comparant ses réponses avec les réponses idéales que l'utilisateur espère recevoir. Plus les réponses du système correspondent à celles que l'utilisateur espère, plus le système est performant.

Le principe de notre contribution est de recommander à partir d'une collection des documents textuels (Corpus) ; les documents qui sont similaires sémantiquement (pertinents) à un document choisi par l'apprenant sur notre plateforme e-learning.

L'idée est de calculer la similarité entre un document (document que l'apprenant a choisi) et chaque document du corpus $sim(doc_choisi, doci)$ après nous trions les résultats obtenus de façon décroissante afin d'appliquer un seuil de recommandation à ces résultats triés pour que notre système retourne seulement les documents pertinents. Dans cette section nous présentons les résultats de l'évaluation de notre système de recommandation sémantique comparé avec un système classique basé sur une méthode statistique qui est la fonction cosinus.
Pour réaliser une telle évaluation nous devons posséder les éléments suivants :

- Un ensemble de documents (corpus).
- Un ensemble de documents représentant les documents sur lesquels la recommandation doit être faite.
- La liste des documents pertinents pour chaque document choisi (c'est-à-dire nous devons connaitre les réponses idéales que le système doit retourner pour chaque document préféré).

Comme rappel de ce que nous avons vu dans le troisième chapitre, la comparaison des réponses d'un système avec les réponses idéales de l'utilisateur nous permet d'évaluer les deux métriques suivantes :

◇ **Précision** : La précision mesure la proportion de documents pertinents sélectionnés parmi tous les documents sélectionnés par le système.
Précision = nombre de documents pertinents retrouvés / nombre de documents pertinents.

◇ **Rappel** : Le rappel mesure la proportion de documents pertinents sélectionnés parmi tous les documents pertinents dans le corpus.
Rappel = nombre de documents pertinents retrouvés / nombre de documents retrouvés.

Un système qui aurait 100% pour la précision et pour le rappel signifie qu'il trouve tous les documents pertinents, et rien que les documents pertinents. Cela veut dire que les réponses du système à chaque requête sont constituées seulement des documents idéaux pour l'utilisateur. En pratique, cette situation n'arrive pas, souvent, nous pouvons obtenir un taux de précision et de rappel aux voisinages de 35%.

Le processus d'évaluation est comme suit :
Pour i = 1, 2, ... document-dans-la-base faire : évaluer la précision et le rappel pour les i premier documents dans la liste des réponses du système.
En conséquence, l'évaluation du système sera avec la courbe Précision/Rappel, qui représente l'évolution de la précision et du rappel c'est à dire pour chaque document retrouvé, nous calculons la précision et le rappel obtenus en considérant seulement le premier document comme réponse, puis les deux premiers, puis les trois premiers, etc.

Exemple illustratif :
Supposons que le nombre des documents idéals que le système doit retourner pour un document choisi par un apprenant est 40. La liste des recommandations du système est donnée

par le tableau 4.5 :

Liste de réponses	Pertinence
Doc1	oui
Doc2	non
Doc3	oui
Doc4	non
Doc5	oui
Doc6	non
Doc7	oui
Doc8	non
Doc9	oui
Doc10	non

TABLE 4.5 – Les réponses du système pour l'exemple illustratif

Nous considérons d'abord le premier document Doc1 comme la réponse du système. À ce point, Nous avons retrouvé un document pertinent parmi les 40 existants. Donc nous avons un taux de rappel de 0.025 avec un taux de précision égal 1. Le point de la courbe est (0.025, 1.0) et ainsi de suite jusqu'à qu'à la dernière réponse du système. Finalement, nous aurons le tableau 4.6 suivant :

Rappel	Précision
0.025	1
0.025	0.5
0.05	0.66
0.05	0.5
0.075	0.6
0.075	0.5
0.1	0.57
0.1	0.5
0.125	0.55
0.125	0.5

TABLE 4.6 – Taux du Rappel/Précision pour l'exemple illustratif

Un des problèmes pour tracer la courbe de Précision/Rappel avec le tableau non normalisé c'est comment faire pour fusionner les résultats de plusieurs tests pour un système ? La solution est de normaliser le tableau de chaque réponse par la règle de maximum :

- Nous fixons une valeur de rappel normalisée **r**,
- Nous sélectionnons les lignes du tableau non-normalisé qui ont une valeur du rappel $\geq$**r**,
- Nous mettons dans le tableau normalisé la valeur de la précision **max** parmi les valeurs des lignes sélectionnées.

En appliquant cette règle sur le tableau de précision/rappel précédent nous trouvons le tableau 4.7 normalisé suivant :

Rappel	Précision
0	1
0.1	0.57
0.2	0
0.3	0
0.4	0
0.5	0
0.6	0
0.7	0
0.8	0
0.9	0
1	0

TABLE 4.7 – Tableau normalisé de Rappel/Précision pour l'exemple illustratif

Pour évaluer notre système de recommandation sémantique par rapport au système basé sur la similarité cosinus, nous avons tracé la courbe du Rappel/Précision pour les deux systèmes dans le même graphe (la figure 4.23). L'évaluation est effectuée avec le même corpus de test.

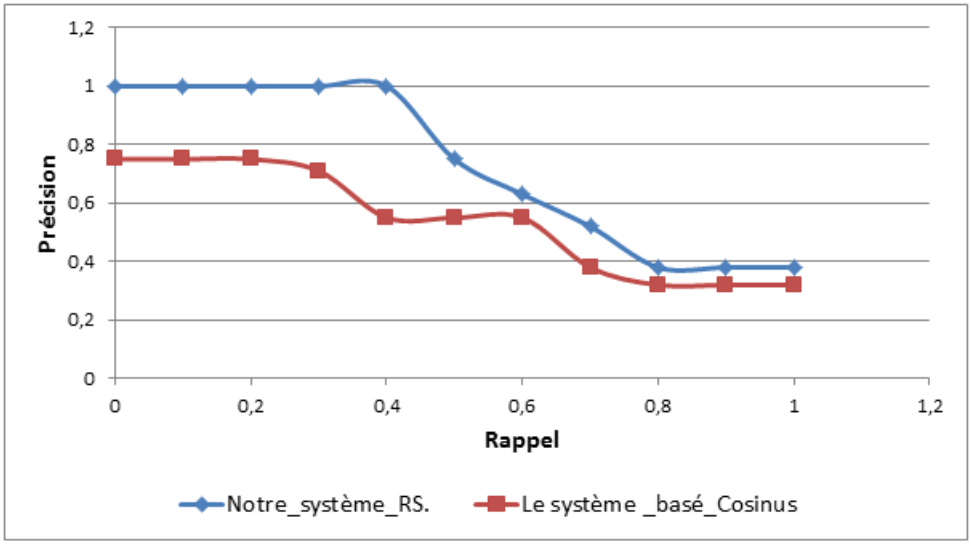

FIGURE 4.23 – Courbe rappel précision pour notre système de recommandation et le système de recommandation basé sur la similarité cosinus

Un système dont la courbe dépasse celle d'un autre est considéré plus performant. Pour conclure, d'après la courbe de la figure 4.23 nous déduisons que notre système de recommandation est plus performant que le système basé sur la similarité cosinus.

3. Conclusion

Dans ce chapitre, nous avons présenté la première partie de nos contributions. Premièrement, nous avons créé un système de génération de parcours d'apprentissage adapté au profil apprenant. L'idée est de transformer le problème du parcours adapté (liste d'activité pédagogique à suivre) ; en un problème d'optimisation. L'utilisation des algorithmes génétiques nous a permis d'automatiser la recherche de contenu adapté au profil de l'apprenant. Deuxièmement, nous avons abordé la recommandation des documents textuels aux apprenants, en se basant sur une approche hybride du calcul de similarité sémantique.

Par ailleurs, nos contributions ne se limitent pas à la recommandation sémantique des documents. Outre cela, dans le chapitre suivant nous allons soulever la problématique de l'accompagnement des apprenants (e-tutorat). C'est pourquoi, nous proposons un système multi-agents qui contribue au suivi du travail collaboratif des apprenants ainsi qu'une approche de la prédiction de leurs performances sur une plateforme e-learning.

Chapitre 5

Système multi-agents pour e-tutorat MASET et une approche de prédiction de la performance des apprenants

"La connaissance s'acquiert par l'expérience, tout le reste n'est que de l'information"

- Albert Einstein,

Introduction

L'exploration de données éducatives est une discipline émergente et prometteuse en sciences de l'éducation dont l'objet est l'extraction d'un savoir ou d'une connaissance à partir des données provenant des plateformes e-learning.
En effet, l'exploration de données utilise des algorithmes d'apprentissage automatique (Machine Learning) [209], pour doter les machines de programmes capables de percevoir leur environnement et pour construire des modèles prédictifs.
Par ailleurs, les systèmes multi-agents prennent aujourd'hui une place de plus en plus importante dans le domaine de e-learning. C'est une discipline qui s'intéresse aux comportements collectifs produits par les interactions entre agents. Ces agents, qui représentent des entités autonomes et flexibles, sont très utiles pour la modélisation, la conception et l'implémentation de systèmes intelligents, complexes, ouverts et dynamiques tels que les systèmes d'apprentissage en ligne.

Dans ce chapitre, nous allons présenter la deuxième partie de nos contributions. Tout d'abord, nous décrivons notre système MASET (Système Multi Agents pour E-Tutorat des apprenants engagés dans le Travail collaboratif en ligne) qui vise essentiellement à aider les tuteurs à surveiller le travail collaboratif des apprenants à travers leurs diverses interactions. En outre, nous proposons une étude comparative de trois algorithmes de datamining pour construire un modèle prédictif basé sur les arbres de décision afin de prédire le niveau des apprenants dans leur parcours d'apprentissage au cours d'une formation en ligne.

1. MASET : Notre système multi-agents pour le e-tutorat des apprenants

Les systèmes multi-agents s'avèrent une solution pertinente pour les plateformes e-learning parce qu'ils possèdent des caractéristiques qui permettent de les structurer d'une manière efficace et ouvrent de nouvelles perspectives d'assistance des apprenants. L'intérêt qu'ils suscitent est lié à leur capacité d'aborder les problèmes complexes d'une manière distribuée et de proposer des solutions modulaires et robustes. En effet, la puissance d'un système multi-agents ne provient pas du comportement individuel des agents mais des interactions entre ses agents. Le comportement collectif des agents se traduit à partir de leurs interactions : coopération, négociation et coordination.

Dans cette section, nous présentons notre système intitulé MASET (Système Multi Agents pour E-Tutorat des apprenants engagés dans le travail collaboratif en ligne) qui a pour objectif principal d'aider les tuteurs dans le suivi du travail collaboratif des apprenants et la surveillance de leurs interactions sur l'espace forum et via la communication sur l'espace chat. Les interactions des apprenants sur ces deux espaces sont évaluées par les IAAI(les Indicateurs d'Analyse Automatique des Interactions) et les résultats trouvés sont communiqués aux tuteurs. Notre système repose sur le middleware JADE (Java Agent DEvelopment Framework) et le LMS Moodle.

Ainsi, l'objectif essentiel de notre contribution est d'avoir un certain degré d'autonomie à l'aide de ces entités logicielles intelligentes (agents) qui sont des programmes auxquels nous pouvons déléguer des tâches spécifiques. Par ailleurs, nous devons souligner que ce système diffère d'un système traditionnel grâce à ces entités autonomes qui négocient entre elles. Ce système multi-agents est testé avec les données d'interactions issues d'une expérimentation menée avec les étudiants du master Ingénierie des Systèmes Informatiques de l'université Sultan Moulay Slimane de Béni Mellal.
Dans ce qui suit nous introduisons la notion d'agents et celle des systèmes multi-agents puis, l'apport de ces systèmes pour e-learning et nous terminions par une implémentation de notre système MASET.

1.1. La notion d'agent intelligent

Les agents intelligents sont des outils d'intelligence artificielle fréquemment utilisés en e-learning. Pour Ferber [210] : "*Un agent est une entité autonome, réelle ou abstraite (physique ou virtuelle), qui est capable d'agir sur elle-même et sur son environnement, qui, dans un univers multi-agent, peut communiquer avec d'autres agents, et dont le comportement est la conséquence de ses observations, de ses connaissances et des interactions avec les autres agents*". Pour Russell et Norwig [211], le concept d'agent en intelligence artificielle se rapporte à "*Tout ce qui peut être perçu comme percevant de son environnement à travers des capteurs et agissant sur l'environnement par les effecteurs*". Enfin, Patti Maes [212] décrit les agents comme "*Systèmes informatiques qui habitent un environnement dynamique complexe, capable d'agir de façon autonome dans cet environnement, et en le faisant réaliser*

un ensemble d'objectifs ou de tâches pour lesquelles ils sont conçus.".

D'après toutes ces définitions, nous définissons un agent comme une entité logicielle ayant la capacité de percevoir son environnement et d'y agir de manière plus ou moins autonome pour atteindre un objectif.

1.2. Les systèmes multi-agents (SMA)

Un système est dit à base d'agents si l'abstraction clé de sa modélisation est celle d'un agent. Le système peut être mono-agent ; comme les systèmes d'intelligence artificielle classique qui modélisent le comportement d'un seul agent intelligent et essaient de simuler dans une certaine mesure les capacités du raisonnement humain [213]. Le système peut également être à base de plusieurs agents interactifs, dans ce cas on parle de systèmes multi-agents.

Ces systèmes qui ont vu le jour avec l'avènement de l'intelligence artificielle distribuée (IAD) [214], s'intéresse aux comportements intelligents qui sont produits par la coopération d'agents dans un système distribué. Ainsi, les agents permettent de réduire la complexité et le temps de la résolution d'un problème par sa division en sous-problèmes, chaque sous problème est affecté à un agent intelligent indépendant dit "résolveur". Pour y parvenir ; la résolution du but commun est établie par l'organisation et la coordination des activités des agents [215]. Nous citons par la suite quelques plateformes utilisées pour le développement des systèmes multi agents :

MADKIT :(Multi-Agent Development Kit) [216] a été développé dès 1998 au LIRMM (Montpellier) par Olivier Gutchneck, est une plateforme des agent flexible, capable de supporter plusieurs modèles de communication simultanément.

JADE :(Java Agent DEvelopment) [217] est un framework de développement de systèmes multi-agents, open-source et basé sur le langage Java. Il offre en particulier un support avancé de la norme FIPA-ACL [218], ainsi que des outils de validation syntaxique des messages entre agents basés sur les ontologies.

GAMA :(Gis and Agent-based Modelling Architecture) [219] est une plateforme générique pour la modélisation et la simulation orientée agent. Le développement de cette plateforme est en parallèle avec l'implémentation de plusieurs modèles complexes qui appartiennent à différents domaines. La diversité des modèles développés nous aide à vérifier la généralité de la plateforme et l'assurer d'une part. Elle sert également à découvrir les fonctionnalités manquantes de la plateforme d'autre part.

JADEX : est une plateforme multi-agents développée en JAVA par l'université de Hambourg, compatible avec de nombreux standards et capable de développer des agents suivant le modèle BDI (Croyance Désire Intention) [220]. Le tableau 5.1 suivant présente une comparaison selon quelques caractéristiques des différentes plateformes citées au-dessus.

Caractéristiques	GAMA	JADEX	MADKIT	JADE
Etat	Logiciel libre	Logiciel libre	Logiciel libre	Logiciel libre
Documentation	Moyenne disponibilité	Moyenne disponibilité	Moyenne disponibilité	Haute disponibilité
Conformité FIPA	conforme	conforme	Non conforme	conforme
Langage de programmation	JAVA	JAVA, KQML , XML	GAML	JAVA, XML

TABLE 5.1 – Comparaison entre les différentes plateformes

En synthèse, nous pouvons conclure que l'ensemble des quatres plateformes étudiées dans cette section répond bien aux attentes de développeurs. Chacune fournit les outils adaptés pour permettre aux agents de communiquer que cela soit en interne, avec les agents qui sont sur le même nœud, ou en externe, avec les agents qui sont sur des nœuds différents.

1.2.1. La plateforme multi-agents JADE

La plate-forme JADE (Java Agent Development Framework) [217] est utilisée pour l'implémentation et la gestion des agents(figure 5.1). JADE est complètement développée en JAVA, suit les spécifications émises par l'organisme FIPA (Foundation for Intelligent Physical Agent) [221], collectif de chercheurs industriels et académiques ayant proposé de nombreux standards en lien avec les agents. JADE est un middleware qui facilite le développement des systèmes multi agents. Nous présentons dans ce qui suit une description des principales spécificités de JADE.
JADE contient :

- **Un Runtime Environment** : l'environnement où les agents peuvent vivre. Cet environnement d'exécution doit être activé pour pouvoir lancer les agents.
- **Une librairie de classes** : que les développeurs utilisent pour écrire leurs programmes correspondant aux agents.
- **Une suite d'outils graphiques** : qui facilitent la gestion et la supervision de la plateforme des agents.

Chaque instance du JADE est appelée *container* (conteneur), et peut contenir plusieurs agents. Un ensemble de conteneurs constituent une plateforme. Chaque plateforme doit contenir un conteneur spécial appelé *main-container* (conteneur principal) et tous les autres conteneurs s'enregistrent auprès de celui-ci dès leurs lancement.

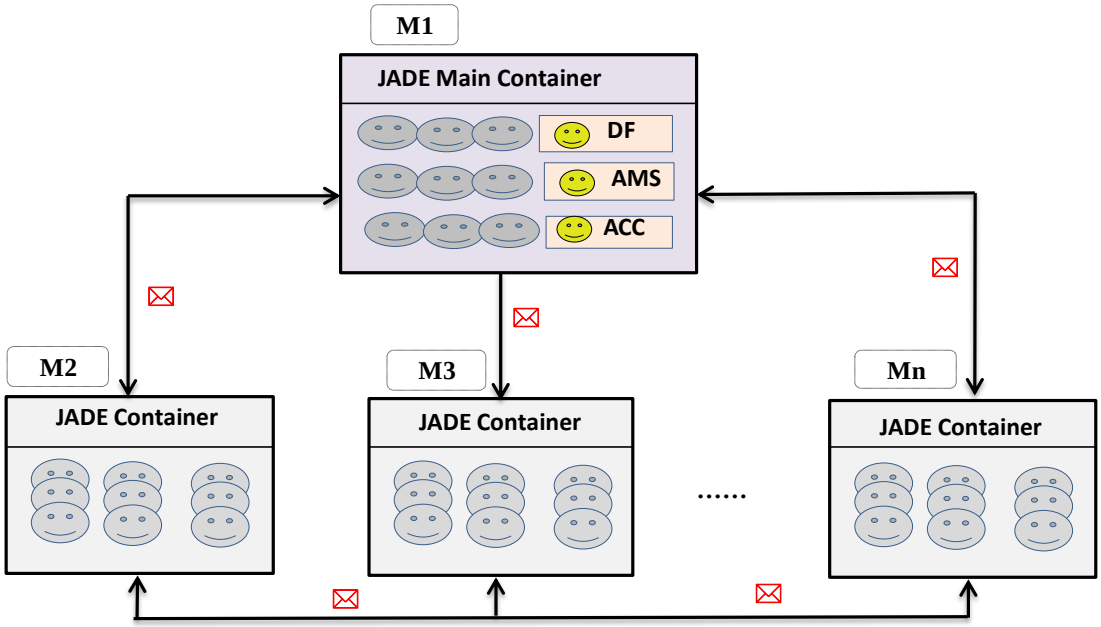

FIGURE 5.1 – Architecture de la plateforme Jade

Un main-container se distingue des autres (simples) conteneurs par le fait qu'il contient toujours des agents spéciaux appelés **DF**, **AMS**, **ACC** et **RMA** qui sont lancés automatiquement au lancement du conteneur principal.

- **AMS (Agent Management System)** : qui fournit le service de nommage (pour assurer par exemple que chaque agent possède un identifiant unique dans la plateforme) et qui représente l'autorité de la plateforme (par exemple il est possible de créer/arrêter des agents en envoyant des requêtes à l'AMS).
- **DF (Directory Facilitator)** : qui fournit un système de pages jaunes qui permet aux agents de retrouver les agents fournisseurs de services.
- **ACC (Agent Communication Channel)** : qui gère la communication entre les agents.
- **RMA (Remote Monitoring Agent)** : cet agent d'interface est exclusivement réservé à la gestion de l'application JADE en cours d'exécution. Il permet d'afficher un certain nombre d'interfaces interrogeables permettant de renseigner l'utilisateur humain sur le déroulement de la session JADE. Ainsi, en interrogeant le RMA, on peut notamment obtenir des informations graphiques provenant des deux agents précédemment cités.

Un autre agent spécial pour la surveillance des communications entre les autres agents est appelé (Sniffer Agent) : Cet agent est spécialisé dans la surveillance des flots de communications entre les différents agents. Par conséquent il est particulièrement utilisé pour s'assurer du bon déroulement d'un protocole de communication.

Les agents sous JADE sont menus d'une autonomie dans le sens de l'indépendance de fonctionnement. En pratique, cela se caractérise par leur capacité à effectuer un ensemble

de comportements qui leur sont propres, sans la nécessité d'une interaction externe. Ces comportements sont appelés *Behavior*. JADE propose également des classes de bases pour des types de comportements spécifiques (CyclicBehaviour, ParallelBehaviour, ReceiverBehaviour, SenderBehaviour, SequentialBehaviour...) [217].

Un langage de communication est nécessaire dans tous les SMA pour les interactions entre les différents agents. Ce langage est basé sur la théorie des actes de communication [217]. Un message échangé entre les agents sera directement liée à une action ou un acte très spécifique de communication. Dans ce langage, un certain nombre d'actes de communication appelé *performative* sont définis. Ces actes de communication sont utilisés pour marquer un message en fonction de son objectif global.

1.2.2. Quelques solutions e-learning basées sur les agents

La technologie des agents a été appliquée dans divers applications pour l'éducation : les agents de recherche d'information, les agents de traitement des informations de l'élève, les agents de recueil ou de génération des Feedbacks, les agents pédagogiques, les agents de tutorat, etc. Les études empiriques effectuées pour évaluer l'efficacité des agents intelligents dans l'enseignement en ligne ont montrés que les agents peuvent améliorer le taux d'achèvement, la satisfaction des apprenants et la motivation. Dans ce qui suit nous présentons quelques solutions e-learning basées sur les agents :
F-SMILE Multi-Agent System [222] : Le système fonctionne dans un environnement multi-agents d'apprentissage intelligent qui peut fournir un tutorat adaptatif basé sur la modélisation de l'apprenant à travers les services web.
aLFanet Multi-Agent System [223] : Le système offre des services éducatifs adaptés aux besoins individuels et collaboratifs de l'apprenant.
MASCE [224] : C'est un système multi-agents pour l'apprentissage collaboratif à distance. Il complète l'apprentissage traditionnel en présentiel. En effet, MASCE permet à l'apprenant de revoir les documents de cours, de demander de l'aide et d'évaluer l'aide fournie afin de permettre au système d'avoir une liste des meilleurs aidants, d'interagir avec leurs tuteurs ou les autres apprenants à l'aide des outils de travail collaboratif.

Dans le même contexte, et dans le cadre de nos travaux de recherche nous avons développé deux contributions : Premièrement, nous avons proposé un prototype appelé MASET (Multi Agents System for E-Tutoring learners engaged in online collaborative work) [225].
Deuxièmement, nous avons proposé une approche multi-agent basée sur la plateforme JADE qui a pour objectif la conception, la modélisation et l'implémentation d'une application (*scolarité intelligente*) en utilisant les systèmes multi-agents [226].

Les systèmes d'apprentissage que nous avons présentés dans cette section sont basés sur l'utilisation des agents intelligents dans le domaine de e-learning. Ils permettent d'améliorer les processus d'adaptation des expériences d'apprentissage en profitant des avantages des SMA qui facilitent la modélisation, la décomposition des traitements et des interactions. Ils permettent aussi d'améliorer les requêtes de recherches des ressources numériques pour pouvoir aider l'apprenant dans sa formation.

1.3. Les indicateurs de productivité des apprenants calculés par les agents

Les besoins des tuteurs au suivi des activités d'apprentissage collaboratif des apprenants font l'objet de plusieurs systèmes d'analyse automatiques des interactions. En effet, l'intégration d'indicateurs cognitifs, sociaux et émotionnels sont misent en œuvre pour aider les tuteurs dans le suivi des apprenants et des groupes d'apprenants.
Le schéma suivant montre les indicateurs calculés par MASET pour assister le travail collaboratif des apprenants. L'agent tuteur recueille la valeur de ces indicateurs pour les communiquer à l'acteur tuteur :

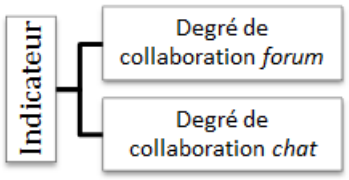

FIGURE 5.2 – Indicateurs calculés par MASET

Comme cité dans le chapitre de l'état de l'art à propos des IAAI, nous identifions deux types d'indicateurs pour notre système (figure 5.2).
Ces indicateurs peuvent être classés en deux catégories :

- Indicateur de la productivité des apprenants
- Indicateur à caractère social

Pour la première catégorie « Indicateur de la productivité des apprenants », nous avons choisi de mettre en œuvre l'indicateur « productivité » des apprenants dans le forum. Cet indicateur est calculé à partir des données collectées à partir des espaces forum basé sur le nombre de documents déposés par chaque apprenant.
Pour la deuxième catégorie « Indicateur à caractère social », qui se réfère aux modes ou à la qualité de la communication et de collaboration au sein d'un groupe. Cet indicateur concerne les actions et les contributions des différents membres, dans l'espace de travail du groupe, et dans ce contexte nous avons choisi de mettre en oeuvre « Indicateur de la collaboration des apprenants dans le chat » qui indique le degré de l'interactivité des apprenants au cours d'une activité d'apprentissage en calculant le nombre de messages échangés dans le chat par période.

1.4. Le comportement des agents

La communication dans les systèmes multi-agents est la base de la distribution des agents et de la résolution coopérative des problèmes. En communiquant, les agents sont capables

de coopérer, de négocier, de coordonner leurs actions ou de réaliser des tâches en commun. Afin de facilité la communication et l'interopérabilité entre les agents, plusieurs langages de communication ont été développés à titre d'exemple FIPA-ACL [218].

Un système multi-agent se caractérise généralement, comme précisé dans la section ci-dessus, par la capacité d'échange et de communication entre les agents qui y sont définis. Notre système est développé en se basant sur le framework JADE ; ces échanges utilisent le langage ACL. Dans ce qui suit, nous présentons quelques interactions implantés dans notre système MASET.
Comme expliqué ci-dessus, l'agent tuteur peut interagir avec d'autres agents qui sont responsables du calcul des différents indicateurs. La figure 5.3 montre le diagramme de séquence qui régit ce type d'interactions.

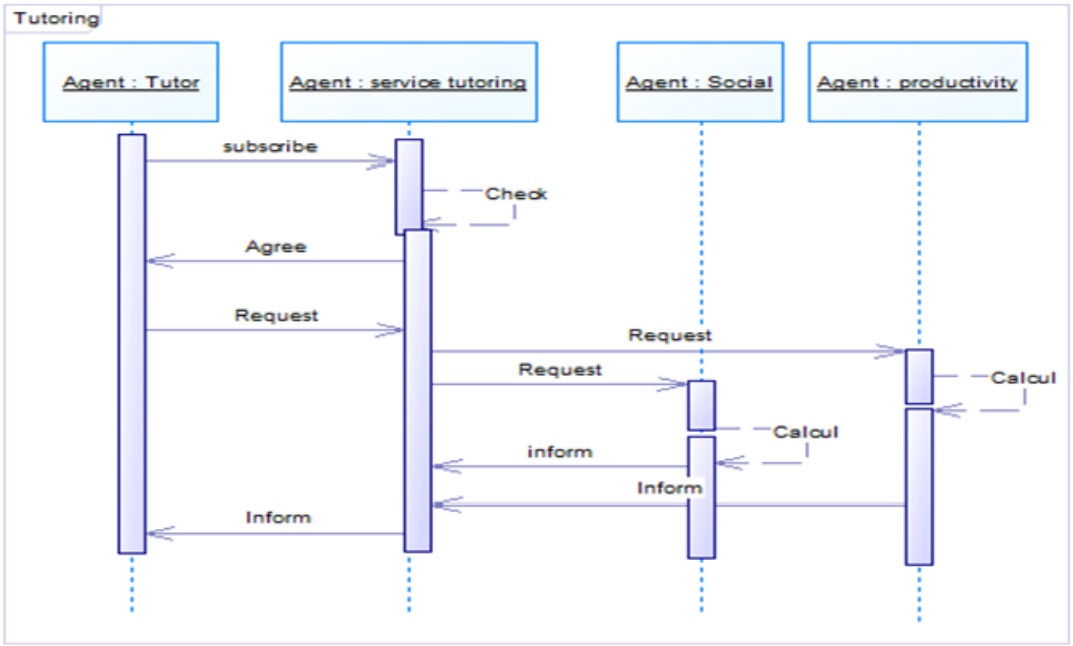

FIGURE 5.3 – Le diagramme de séquence de MASET

1.5. Implémentation du système MASET avec la technologie agent

MASET que nous proposons est un système qui a pour objectif principal l'analyse des interactions afin d'améliorer le tutorat en ligne. Les résultats fournis par les IAAI sont utilisés par le tuteur pour assister les apprenants. Par ailleurs, pour atteindre ces objectifs, nous avons opté pour une architecture distribuée où les agents communiquent et collaborent entre eux afin de réaliser les tâches qui leurs sont confiées (figure 5.4).

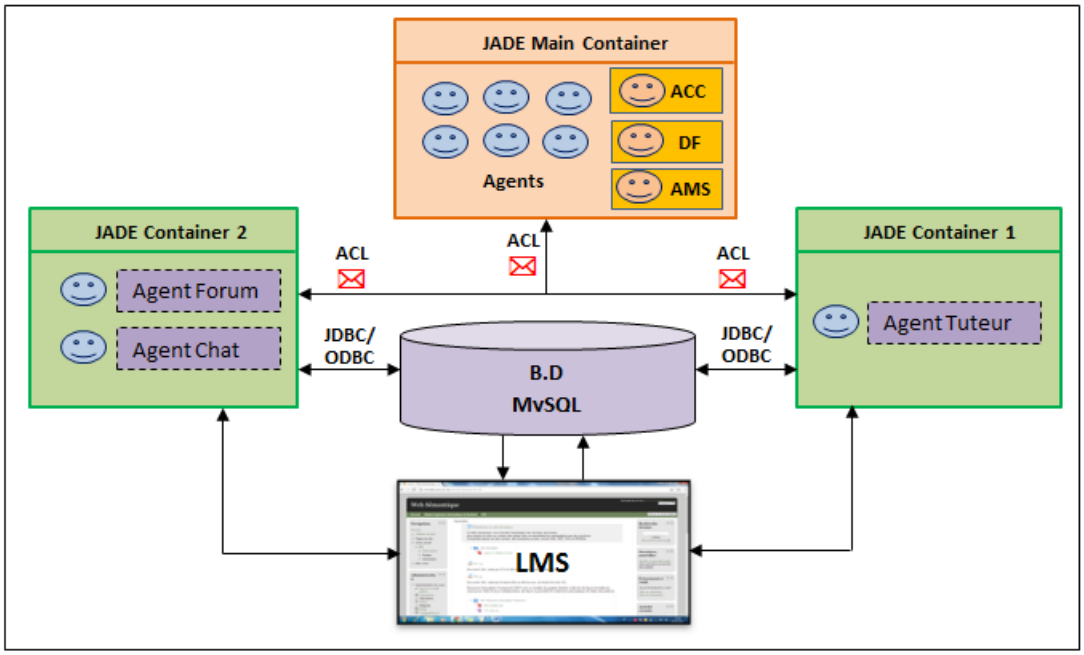

FIGURE 5.4 – Architecture du système MASET

Nous avons créé un système multi-agents distribué sur des appareils mobiles, qui supportent le système Androïde, et sur des ordinateurs desktops. Cette application permet aux tuteurs de surveiller les activités des apprenants dans les forums et les sessions du chat. Nous avons créé trois agents : *Agent_tuteur*, *Agent_forum* et *Agent_chat* ; chacun de ces agents fonctionne comme déjà expliqué dans la figure ci-dessus.

L'interface de la figure 5.5 permet à l'*Agent_tuteur* de se déployer en entrant l'adresse IP de la machine qui contient le conteneur principal.

FIGURE 5.5 – L'interface de connexion pour le tuteur

L'*Agent_forum* et *Agent_chat* sont déployés sur un autre container sur la machine qui contient le conteneur principal. C'est dans cette machine que tous les autres agents seront

déployés (figure 5.6).

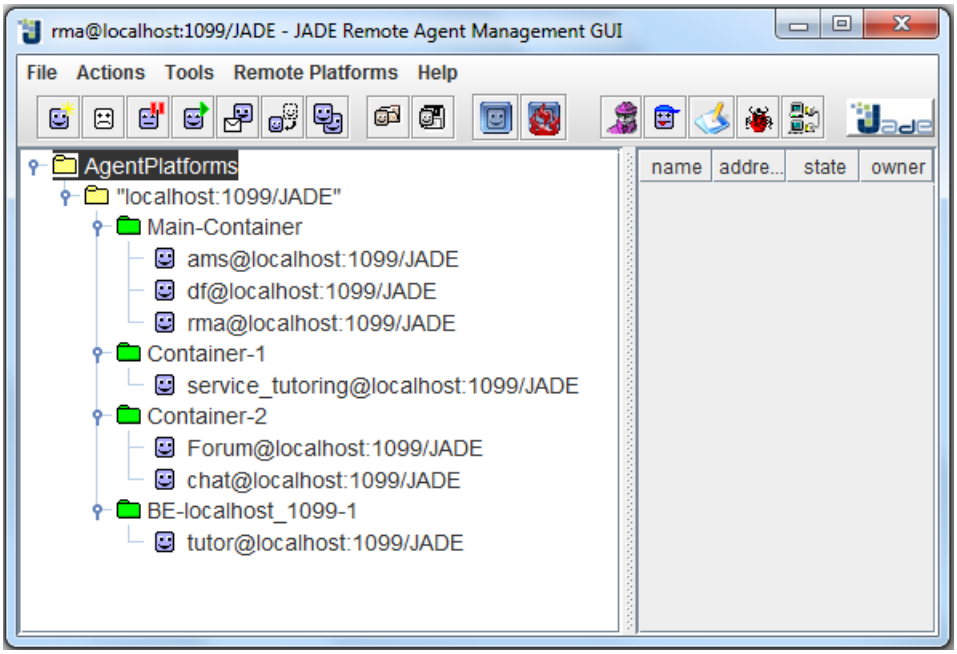

FIGURE 5.6 – Les agents JADE déployés lors de l'exécution de notre système MASET

La vérification de la circulation des messages entre les agents de notre système MASET est faite grâce à l'agent **sniffer** de JADE. Ce dernier permet la visualisation des messages échangés entre les différents groupes d'agents (figure 5.7).

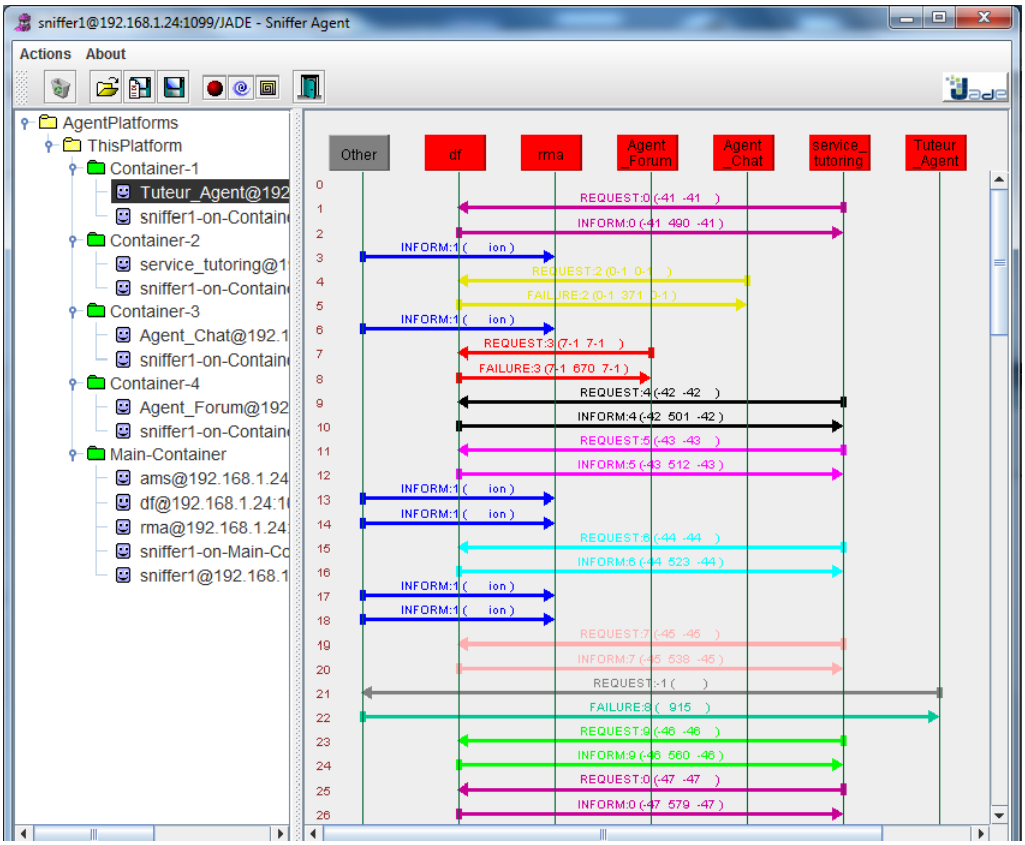

FIGURE 5.7 – La surveillance des différents agents de notre système MASET

La figure 5.8 permet d'afficher les valeurs des indicateurs calculés par l'*Agent_forum* et *Agent_chat*.

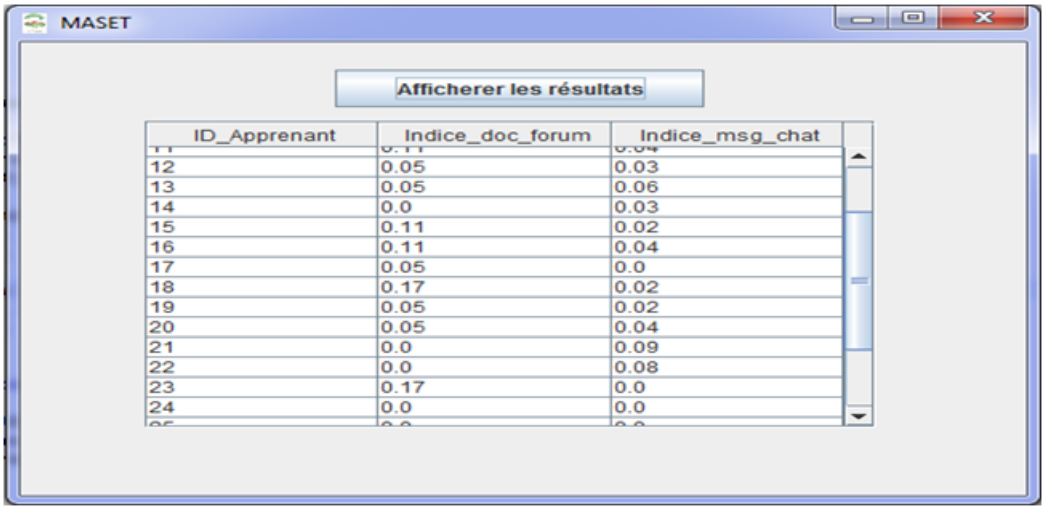

FIGURE 5.8 – Le résultat du calcul des indicateurs

1.6. Résultats et évaluation

Pour évaluer notre système, une expérience a été menée auprès de 24 étudiants en master Ingénierie Informatique et Système. Un certain nombre de cours et travaux pratiques qui font partis d'un module de la formation est misent en ligne. De même, la collaboration entre les étudiants est obligatoire lors de la réalisation de leurs activités.
L'analyse des interactions entre les étudiants en utilisant MASET est illustrée par les figures suivantes :
Premièrement, l'*Agent_forum* responsable du calcul de l'indicateur de la productivité des étudiants dans le forum (figure 5.9) :

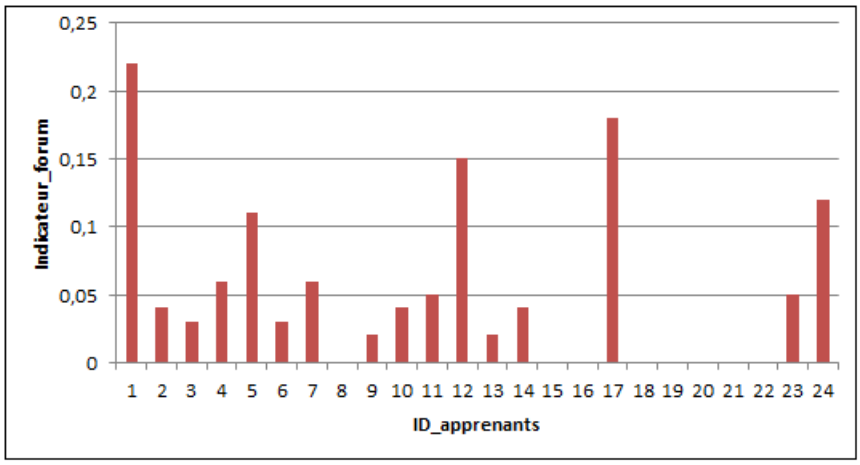

FIGURE 5.9 – La productivité des apprenants dans l'espace forum

Deuxièmement, l'*Agent_chat* responsable du calcul de degré d'interactivité des apprenants dans le chat (figure 5.10) :

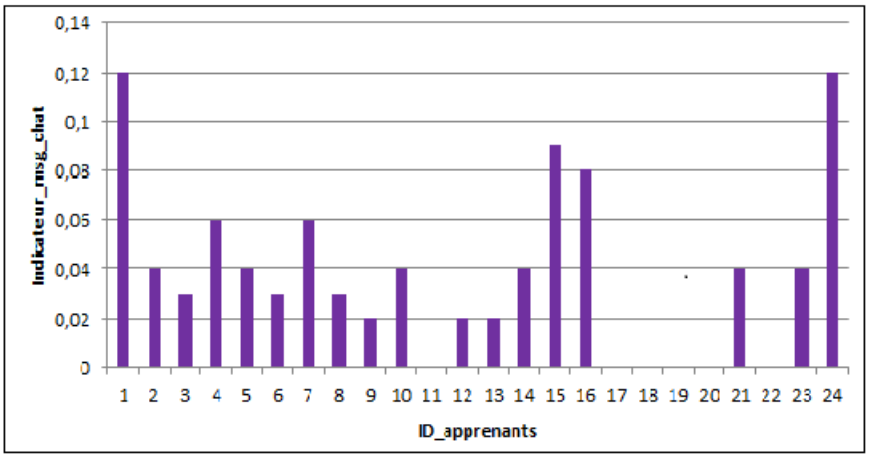

FIGURE 5.10 – Degré d'interactivité des apprenants dans l'espace chat

En conséquence, il faut souligner que l'objectif principal de ce système est d'aider les tuteurs dans le suivi du travail collaboratif des apprenants et de leurs diverses interactions.

Dans cette contribution nous avons introduit les indicateurs d'analyse automatique des interactions. Puis, nous avons également présenté les spécifications que nous avons suivies pour développer notre système qui a été mis en pratique avec les étudiants universitaires dans le LMS Moodle.

Comme nos recherches se concentrent sur comment aider un tuteur à surveiller les activités des apprenants sur les plateformes e-learning, nous proposons dans ce qui suit un modèle prédictif basé sur les arbres de décision pour prédire la performance des apprenants [227].

2. La prédiction de la performance des apprenants

La capacité de prédire la performance des apprenants sur une plate-forme e-learning est un facteur décisif dans les systèmes éducatifs actuels. En effet, l'apprentissage par les arbres de décision utilise des algorithmes sophistiqués et efficaces basés sur l'utilisation de modèles prédictifs. Ces derniers constituent un outil d'aide à la décision pour l'évaluation de la valeur d'une caractéristique d'une population en se basant sur l'observation des autres caractéristiques de la même population.

Comme on pouvait s'y attendre, nous proposons un modèle prédictif basé sur les arbres de décision pour prédire le niveau des apprenants dans leur parcours d'apprentissage durant une formation en ligne. Le choix de l'algorithme le plus efficace pour la construction de ce modèle prédictif est réalisé en se basant sur une étude comparative entre ces différents algorithmes d'arbres de décision [228].
Nous présentons dans ce qui suit le principe de l'apprentissage par arbres de décision, puis nous introduisons une description de notre modèle prédictif et les résultats obtenus :

2.1. Apprentissage automatique

L'apprentissage automatique est un champ d'étude de l'intelligence artificielle qui concerne la conception, l'analyse et l'implémentation de méthodes permettant à une machine (au sens large) d'apprendre et d'améliorer sa performance par expérience [209]. En effet, l'objectif principal de l'apprentissage automatique est de remplir des tâches difficiles ou impossibles à remplir par des moyens algorithmiques classiques.

L'apprentissage automatique se base sur un ensemble d'algorithmes issus de disciplines scientifiques diverses telles que les statistiques, la science informatique et l'intelligence artificielle, afin de construire des modèles à partir des données, cela consiste à trouver des structures selon des critères fixés au préalable, et d'en extraire un maximum de connaissances.
De plus, les algorithmes d'apprentissage automatique peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient : apprentissage supervisé [229], apprentissage non supervisé [230] et apprentissage par renforcement [231].

2.2. Apprentissage par arbres de décision

L'apprentissage par arbre de décision désigne une technique d'apprentissage supervisé basée sur l'utilisation d'un arbre de décision comme modèle prédictif [232]. Il est considéré comme un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les *feuilles* de l'arbre), et sont atteintes en fonction de décisions prises à chaque étape. L'arbre de décision est un outil utilisé dans des domaines variés tels que la sécurité, la fouille de données, la médecine, etc. Il présente plusieurs avantages : la simplicité de compréhension, d'interprétation et la performance sur de grands jeux de données. Il s'agit de plus d'une représentation calculable automatiquement par des algorithmes d'apprentissage supervisé.

L'idée générale des arbres de décision est de diviser récursivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage par des tests définis à l'aide des attributs jusqu'à ce que l'on obtienne des sous-ensembles d'exemples ne contenant (presque) que des exemples appartenant tous à une même classe.
Dans toutes les méthodes, nous trouvons les trois opérateurs suivants :

- *Décider si un nœud est terminal*, c'est-à-dire décider si un nœud doit être étiqueté comme une feuille.
- *Sélectionner un test à associer à un nœud*, soit aléatoirement, soit en utilisant des critères statistiques (si ce nœud n'est pas terminal).
- *Affecter une classe à un noeud* (si ce nœud est terminal).

Les méthodes vont différer par les choix effectués pour ces différents opérateurs, c'est-à-dire selon le choix d'un test (par exemple, utilisation du coût et de la fonction entropie) et le critère d'arrêt (quand arrêter la croissance de l'arbre c'est-à-dire quand décider si un nœud est terminal).

Nous pouvons alors définir un schéma général d'algorithme, sans spécifier comment seront définis les trois opérateurs décrits plus haut, comme suit :

Données : échantillon S
Initialiser l'arbre courant à l'arbre vide ; la racine est le nœud courant;
tant que *noeud courant n'est pas terminal* **faire**
  si *le noeud est terminal* **alors**
    Affecter ce nœud à une classe;
  sinon
    Sélectionner un test et créer autant de nouveaux nœuds fils qu'il y a de réponses possibles au test;
  fin
  Passer au noeud suivant non exploré s'il en existe;
  Jusqu'à obtenir un arbre de décision
fin

Algorithme 2 : Apprentissage générique par arbre de décision

Avec un tel algorithme, nous décidons qu'un nœud est terminal lorsque tous les exemples associés à ce nœud, ou du moins la plupart d'entre eux sont dans la même classe. De plus, un arbre de décision parfait est un arbre dont tous les exemples de l'ensemble d'apprentissage soient correctement classifiés. Lorsque plusieurs classes sont en concurrence, nous pouvons choisir la classe la plus représentée dans l'ensemble de l'échantillon, ou en choisir une au hasard. La sélection d'un test à associer à un nœud est plus délicate.

Tandis que, l'idéal serait de trouver un critère qui permet d'arrêter la croissance de l'arbre au bon moment. Malheureusement, dans l'état actuel des recherches, un tel critère n'a pu être trouvé. En outre, le risque d'arrêter trop tôt la croissance de l'arbre est plus important que de l'arrêter trop tard. Par conséquent, les méthodes utilisées pour la construction d'un arbre de décision peuvent contenir plusieurs anomalies liées au problème de surapprentissage (overfitting). Il s'agit de la déduction d'informations plus que supporte l'ensemble de données d'apprentissage. L'arbre peut être aussi d'une taille très importante qui peut épuiser les ressources de stockage. Pour surmonter ce problème, nous effectuons des opérations d'élagage [233] qui consistent à éliminer de l'arbre les branches les moins significatives. L'élagage peut être effectué avant ou après l'apprentissage, nous parlons souvent de pré et post-élagage [234].

2.2.1. Exemple illustratif

Nous allons considérer un exemple très simple pour introduire les algorithmes d'apprentissage par arbres de décision. Nous prenons en entrée un ensemble de données classées, et nous fournissons en sortie un arbre où chaque nœud final (feuille) représente une décision (une classe) et chaque nœud non final (intermédiaire) représente un test. Chaque feuille représente la décision d'appartenance à une classe de données vérifiant tous les tests du chemin menant de la racine à cette feuille. L'exemple du tableau 5.2 montre un ensemble de données avec quatre attributs : Ciel, Température, Humidité, Vent et l'attribut à prédire **Jouer**.

Jour	Ciel	Température	Humidité	Vent	Jouer
J1	soleil	chaud	élevé	faible	non
J2	soleil	chaud	élevé	fort	non
J3	couvert	chaud	élevé	faible	oui
J4	pluie	doux	élevé	faible	oui
J5	pluie	froid	normale	faible	oui
J6	pluie	froid	normale	fort	non
J7	couvert	froid	normale	faible	oui
J8	soleil	Doux	élevé	faible	non
J9	soleil	froid	normale	faible	oui
J10	pluie	doux	normale	fort	oui
J11	soleil	doux	normale	fort	oui
J12	couvert	doux	élevé	fort	oui
J13	couvert	chaud	normale	faible	oui
J14	pluie	doux	élevé	fort	non

TABLE 5.2 – Echantillon des données météorologiques

L'arbre appris à partir de cet ensemble de donnée est le suivant (figure 5.11) :

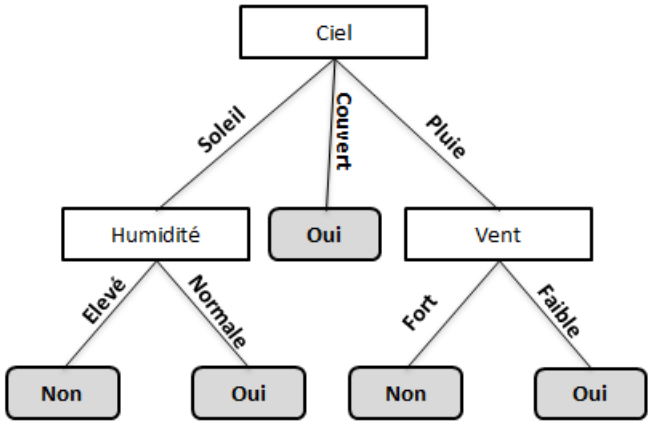

FIGURE 5.11 – Exemple d'arbre de décision

Nous pouvons remarquer que toutes les données ayant l'attribut Ciel="Soleil" et l'attribut Humidité="Normale" appartiennent à la classe ("*oui*"). En effet, toute nouvelle donnée peut être classée en testant ses valeurs d'attributs l'un après l'autre en commençant de la racine jusqu'à atteindre une feuille c'est-à-dire une décision.

Pour construire un tel arbre, plusieurs algorithmes existent : ID3 [235], CART [236], C4.5 [237], RF [238], etc. Nous commençons généralement par le choix du meilleur attribut puis nous affectons cet attribut à la racine. Pour chaque valeur de cette attribut, nous créons un nouveau noeud fils de la racine après nous classons les exemples dans les noeuds fils. Si tous les exemples d'un noeud fils sont homogènes, nous affectons leur classe au noeud, sinon nous

commençons à partir de ce noeud. L'algorithme continue d'une manière récursive jusqu'à obtenir des nœuds qui contiennent des données homogènes.

Données : échantillon D
Résultat : Arbre de decision
initialization;
si *tous les exemples de D sont de la même classe C* **alors**
  Retourner N comme une feuille étiquetée par C ;
fin
si *la liste des attributs est vide* **alors**
  Retourner N comme une feuille étiquetée de la classe de D
fin
Selectionner l'attribut A du meilleur Gain dans D ;
Etiqueter N par l'attribut selectionné ;
Liste d'attributs ← Liste d'attributs - A ;
pour *(chaque valeur ;Vi ; de A)* **faire**
  Soit Di le sous ensemble de D ayant la valeur de A = Vi ;
  Attacher à N le sous arbre généré par l'ensemble Di et la liste des attributs ;
fin

Algorithme 3 : Construction d'un arbre de décision

En réalité la conception d'un algorithme pour la construction d'arbre de décision n'est pas si simple, plusieurs problèmes doivent être résolus :

- Comment choisir l'attribut qui sépare le mieux l'ensemble de données ? Nous parlons souvent de la variable de segmentation.
- Comment choisir les critères de séparation d'un ensemble selon l'attribut choisi, et comment ces critères varient selon que l'attribut soit numérique ou symbolique ?
- Quel est le nombre optimal de critères qui minimise la taille de l'arbre et maximise la précision ?
- Quels sont les critères d'arrêt de ce partitionnement, sachant que souvent l'arbre et d'une taille élevée ?

Nous présentons ci-dessous quelques fondements mathématiques des principaux algorithmes d'apprentissage par arbres de décision comme CART et C4.5.

2.2.2. Choix de la variable de segmentation

Il s'agit de choisir parmi les attributs des données, celui qui les sépare le mieux du point de vue de leurs classes déjà connues. Pour choisir le meilleur attribut, on calcule pour chacun une valeur appelée *Gain* qui dépend des différentes valeurs prises par cet attribut. Cette mesure est basée sur les recherches en théorie d'informations menées par C.Shannon [239]. Par exemple, l'algorithme ID3 utilise le concept d'entropie introduite initialement par Shannon en 1948.

Soit un ensemble S d'exemples dont une proportion p_+ est positive et une proportion p_- est négative (bien entendu, $p_+ + p_- = 1$), l'entropie de E est donnée par la formule suivante :

$$E(S) = -p_- \log_2(p_-) - p_+ \log_2(p_+), \tag{5.1}$$

sachant que : $0 \leq E(S) \leq 1$.
Si $p_+ = 0$ ou $p_- = 0$ alors $E(S) = 0$.
Ainsi, si tous les exemples sont tous positifs ou tous négatifs, l'entropie de la population est nulle.
Si $p_+ = p_- = 0.5$ alors $E(S) = 1$.

Dans le cas où nous avons k classes :

$$E(S) = -\sum_{i=1}^{k} p_i \log_k(p_i) \tag{5.2}$$

L'entropie est représentée graphiquement dans la figure 5.12 en fonction de la probabilité. La quantité E a un certain nombre de propriétés intéressantes qui justifient son utilisation comme mesure raisonnable pour le choix de variable de segmentation de l'ensemble des exemples.

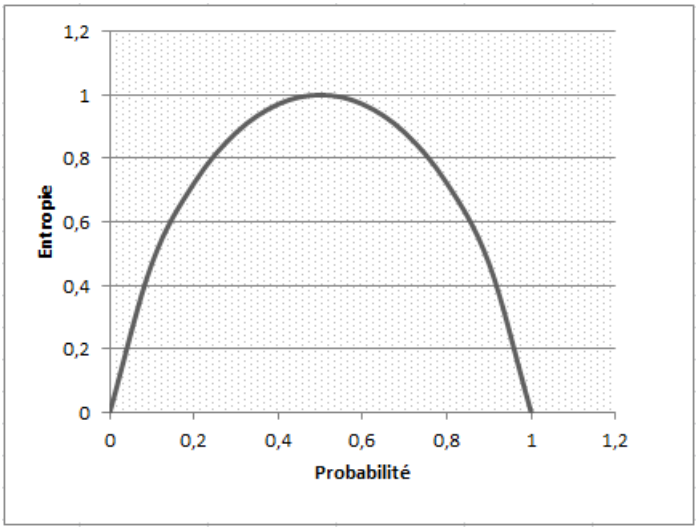

FIGURE 5.12 – L'allure de l'ontropie dans l'intervalle [0,1]

Le critère d'évaluation des partitions caractérise l'homogénéité des sous-ensembles obtenus par division de l'ensemble. Ces métriques sont appliquées à chaque sous-ensemble candidat et les résultats sont combinés pour produire une mesure de la qualité de la séparation. Ainsi, nous disposons d'une fonction permettant de mesurer le degré de séparation des classes pour tout échantillon, il s'agit de la fonction *Gain* définie par la formule suivante :

$$Gain(S, Q) = E(S) - \sum_{i=1}^{k} p_i E(S, Q_i), \tag{5.3}$$

où $E(S)$ est l'ontropie de l'ensemble S et $E(S, Q_i)$ est l'ontropie lié à l'attribut i.

Application numérique liée à notre exemple courant du tableau 5.2 :

$E(S) = -\frac{3}{8}log_2(\frac{3}{8}) - \frac{3}{8}log_2(\frac{3}{8}) = 0.954$

Le gain de l'attribut **Humidité** est calculé comme suit :
$G(S, Humidite) = 0.954 - \frac{7}{14} \times 0.985 - \frac{7}{14} \times 0.592$
$G(S, Humidite) = 0.151$

Le gain de l'attribut **Vent** est calculé comme suit :
$G(S, Vent) = 0.954 - \frac{8}{14} \times 0.811 - \frac{6}{14} \times 1$
$G(S, Vent) = 0.048$

Le gain de l'attribut **Température** est calculé comme suit :
$G(S, Temperature) = 0.954 - \frac{4}{14} \times 1 - \frac{6}{14} \times 0.92 - \frac{4}{14} \times 0.81$
$G(S, Temperature) = 0.042$

Le gain de l'attribut **Ciel** est calculé comme suit :
$G(S, Ciel) = 0.954 - \frac{5}{14} \times 0.971 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.971$
$G(S, Ciel) = 0.247$

La condition à vérifier pour déterminer la variable de segmentation est de prendre la variable du gain d'information maximum.
Le gain maximal est obtenue pour le choix de l'attribut **Ciel**. Nous remarquons que le choix du test **Température** est très mauvais, ce qui correspond bien à l'intuition.

Nous allons, dans le paragraphe suivant, présenter quelques algorithmes utilisés pour la constriction des arbres de décision en particuliers ID3, C4.5, CART.

2.3. Algorithmes de construction d'arbres de décision

2.3.1. L'algorithme ID3

L'algorithme ID3 [235] construit l'arbre de décision récursivement. À chaque étape il calcule parmi les attributs restants pour la branche en cours, celui qui maximisera le gain d'information. C'est-à-dire l'attribut qui permettra le plus facilement de classer les exemples à ce niveau de cette branche de l'arbre. Le calcul se fait à base de l'entropie de Shanon déjà présentée ci-dessus. L'algorithme suppose que tous les attributs sont catégoriels ; si des attributs sont numériques, ils doivent être discrétisés pour pouvoir l'appliquer. ID3 se base sur **l'algorithme 2** présenté précédemment dans ce chapitre.

2.3.2. L'algorithme CART

L'algorithme CART "Classification And Regression Trees" [236], construit un arbre de décision d'une manière analogue à l'algorithme ID3. Contrairement à ce dernier, l'arbre de

décision généré par CART est binaire (un nœud ne peut avoir que 2 fils) et le critère de segmentation est l'indice de Gini [240].
À un attribut binaire correspond un test binaire. À un attribut qualitatif ayant n modalités, nous pouvons associer autant de tests qu'il y a de partitions en deux classes, soit 2^{n-1} tests binaires possibles. Enfin, dans le cas d'attributs continus, il y a une infinité de tests envisageables. Dans ce cas, on découpe l'ensemble des valeurs possibles en segments, ce découpage peut être réalisé par un expert ou de façon automatique.

2.3.3. L'algorithme C4.5

L'algorithme C4.5 [237] est une amélioration de l'algorithme ID3, appelé aussi (J48), il prend en compte les attributs numériques ainsi que les valeurs manquantes. Cet algorithme s'appuie sur le gain défini dans (l'équation 5.3) et la fonction entropie (l'équation 5.2) combiné avec une fonction SplitInfo [241] pour évaluer les attributs à chaque itération.

Pour les attributs à valeur sur intervalle continu : l'algorithme permet de les gérer de la façon suivante : Si l'attribut C_i a un intervalle continu de valeurs. Nous examinons les valeurs de cet attribut dans les données d'apprentissage. Supposons que ces valeurs sont en ordre croissant, A_1, A_2, ..., A_m. Ensuite pour chacune de ces valeurs, on partitionne les enregistrements entre ceux qui ont des valeurs de C_i inférieures ou égales à A_j et celles qui ont des valeurs supérieures à à A_j. Pour chacune de ces partitions nous calculons le gain, ou le GainRatio et nous choisissons la partition qui maximise le gain.

Pour les attributs à valeurs manquantes : Dans de nombreux problèmes concrets, il existe certains attributs dont les valeurs ne sont pas renseignées. Par exemple, si nous disposons des données de patients, il est très probable que toutes les mesures ne soient pas disponibles car elles n'ont pas pu être faites pour tout les patients. Pour classifier un exemple possédant des valeurs manquantes à l'aide d'arbres de décision, nous procèdons comme dans le cas standard, lorsque nous rencontrons un test et que la valeur de l'attribut est manquante, nous considèrons la branche majoritaire. Pour la phase d'apprentissage, nous supposons que la valeur de cet attribut suit la distribution des valeurs connues.

2.4. Descriptif de notre modèle prédictif

Les apprenants se connectent à la plate-forme e-learning pour suivre une formation en réalisant des activités d'apprentissage. Notre modèle intègre les algorithmes d'arbres de décision pour évaluer la performance des apprenants et leur travail collaboratif en se basant sur les données récoltées à partir de la plate-forme e-learning sur lesquels les apprenants sont inscrits (figure 5.13).

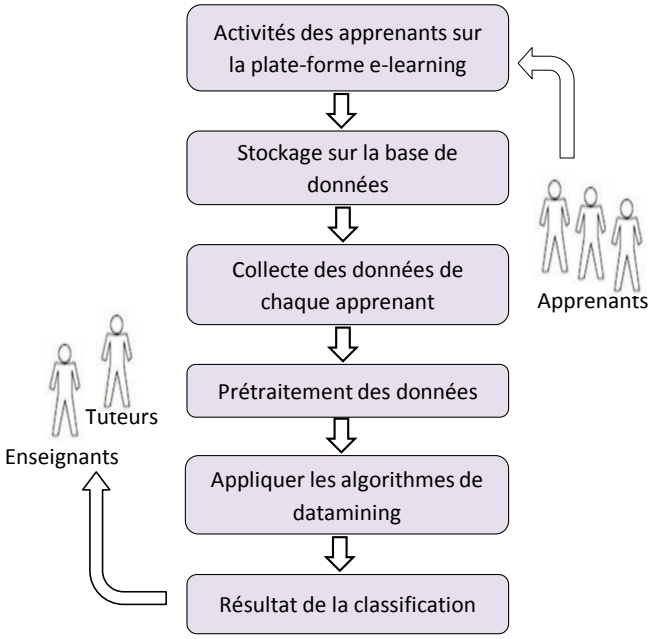

FIGURE 5.13 – Architecture de notre modèle prédictif

En fait, grâce à ce modèle, nous pouvons avoir une idée sur le niveau des apprenants et évaluer leur performance au cours d'une formation sur la plateforme e-learning.

Comme notre objectif est de prédire la performance des apprenants (*Decision* est l'attribut à prédire) en se basant sur leurs données produites lors d'une activité d'apprentissage ($Mark\_exam, quiz\_p, wiki...$), le tableau 5.3 présente les différents attributs utilisés dans notre étude de prédiction :

Attributs	Description	Valeurs possibles
T_wiki	Type de wiki réalisé	{A,B,C,D}
Nb_assignment	Nombre d'attributions effectuées.	Un entier
Quiz_p	Nombre de quiz passé	Un entier
Quiz_f	Nombre de quiz échoué	Un entier
msg_forum	Nombre de messages envoyés au forum	Un entier
msg_read	Nombre de messages lus sur le forum	{A,B,C}
Indication_tutor	Le grade donné par le tuteur aux apprenants	Un entier
T_time	Temps total passé pour réaliser les quiz	Un entier
Mark_exam	La note d'examen est supérieure à 10	Yes, No
Sex	Sexe de l'apprenant	{M,F}
Age	Âge de l'apprenant	Un entier
Specialty_L	Spécialité de l'apprenant	{chemistry, mathematics, physics, computer science}
Decision	Boolean	{ accept or reject}

TABLE 5.3 – La signification des attributs qui caractérisent les apprenants

De même, nous avons un échantillon des mesures avec une série d'attributs choisis pour leur liaison (la puissance discriminative) avec l'attribut que nous cherchons à prédire. Ainsi, l'arbre de décision est créé à partir de cet échantillon des mesures (la base d'apprentissage) et les résultats sont généralement exprimés comme une probabilité de satisfaction ou non de l'attribut prédit.

2.5. Expérience et évaluation

Notre base d'apprentissage se compose de 270 apprenants [242]. Chaque apprenant est décrit par 13 attributs, l'objectif étant de classer automatiquement un apprenant dans la classe qui correspond à son rendement en fonction de ses activités et de son travail collaboratif sur la plate-forme e-learning. Nous avons défini deux classes d'apprenants selon la valeur du treizième attribut (*Decision*).

Le modèle prédictif dérivé de notre base de connaissances par l'algorithme C4.5 est illustré sur la figure 5.14 :

```
Mark_exam = non
|   Specialty = D
|   |   Quiz_P <= 4: Refuser (2.0)
|   |   Quiz_P > 4: Accepter (10.0/1.0)
|   Specialty = A: Refuser (120.0/18.0)
|   Specialty = B
|   |   sexe = masculin: Accepter (21.0/6.0)
|   |   sexe = feminin: Refuser (10.0/1.0)
|   Specialty = C
|   |   msg_read = C: Accepter (9.0/1.0)
|   |   msg_read = A
|   |   |   sexe = masculin: Accepter (4.0/1.0)
|   |   |   sexe = feminin: Refuser (5.0)
|   |   msg_read = B: Accepter (0.0)
Mark_exam = oui
|   T_wiki = D
|   |   Quiz_P <= 7
|   |   |   Specialty = D: Accepter (2.0)
|   |   |   Specialty = A: Refuser (9.0/2.0)
|   |   |   Specialty = B: Accepter (6.0/1.0)
|   |   |   Specialty = C: Accepter (1.0)
|   |   Quiz_P > 7: Accepter (52.0/2.0)
|   T_wiki = C
|   |   age <= 35: Refuser (8.0/1.0)
|   |   age > 35: Accepter (3.0)
|   T_wiki = B: Refuser (4.0/1.0)
|   T_wiki = A: Refuser (4.0/1.0)

Number of Leaves  :     18

Size of the tree :      28
```

FIGURE 5.14 – Règles correspondants à l'algorithme C4.5

Comparaison de la précision et du temps d'exécution des algorithmes de construction d'arbres de décision ID3, C4.5 et CART (tableau 5.4) et (figure 5.15) :

	ID3	C4.5	CART
Temps d'exécution (Sec)	0,01	0,05	0,27
Taux de reconnaissance (%)	0,7519	0,8730	0,8641
Taux d'erreur (%)	0,2481	0,1370	0,1259

TABLE 5.4 – Comparaison de la précision et du temps d'exécution de ID3, C4.5 et CART

Le résultat de cette étude montre que le modèle basé sur C4.5 est le meilleur prédicteur en termes de temps d'exécution (0,05 seconde) et le taux de reconnaissance (87.30%).

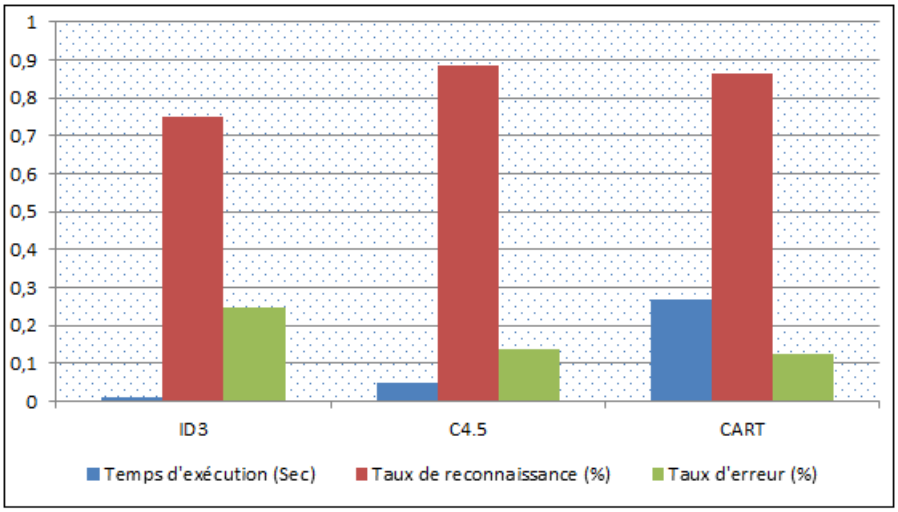

FIGURE 5.15 – Comparaison de la précision et du temps d'exécution des algorithmes ID3, C4.5 et CART

Pour conclure, les résultats de cette prédiction facilitent grandement la tâche des tuteurs et des enseignants. Elle présente l'apprentissage automatique par des arbres de décision qui peuvent être utilisés dans des contextes pratiques pour prédire le rendement scolaire des apprenants.
Cette étude contribue à une meilleure répartition des apprenants pour identifier ceux qui nécessitent une attention particulière de la part des tuteurs afin d'assurer un accompagnement efficace.

Conclusion

Dans ce chapitre, nous avons présenté une vision globale de l'intégration des agents intelligents et des arbres de décision dans le domaine de e-learning.

Premièrement, nous avons présenté notre système multi-agents MASET dont l'objectif principal est d'aider les tuteurs dans le suivi du travail collaboratif des apprenants et la surveillance de leurs interactions. Deuxièmement, nous avons décrit un modèle prédictif de la performance des apprenants en utilisant les arbres de décision. Ainsi, nous avons expliqué à travers ce modèle que la prédiction se fait à travers la construction d'un arbre dont chaque nœud correspond à une décision.

Conclusion et perspectives

Conclusion générale

Le développement des systèmes e-learning adaptatifs représente un enjeu majeur sur le niveau pédagogique et technologique. A ce sens, il est impraticable de fournir le même document éducatif par le même style à tous les apprenants qui sont de nature hétérogène. Donc, l'adaptation de contenu selon les besoins et les préférences des apprenants à tout moment est indispensable pour pouvoir les motiver et les guider et ainsi éviter l'échec ou l'abandon de ces apprenants.

Plusieurs questions ont été posées au début de cette thèse, nous en rappelons ici quelques-unes : Qu'est-ce que l'adaptabilité des contenus ? Comment adapter les contenus à un apprenant ? Comment recommander un document texte à un apprenant en se basant sur son contenu ? Ces questions ont constitué le fil conducteur de l'ensemble de nos recherches.

Les travaux présentés dans cette thèse se situent dans le cadre des problèmes d'adaptation de la formation au profil de l'apprenant. Nous avons utilisé les algorithmes génétiques pour pouvoir générer automatiquement, à partir de différents cours, un parcours de formation adapté au profil de l'apprenant et à l'objectif pédagogique visé. Dans le cadre de cette adaptation, nous nous sommes ainsi concentrés sur la recommandation sémantique des ressources pédagogiques aux apprenants. Nous avons décrit les systèmes de recommandation, qui sont considérés comme un sous-ensemble des systèmes adaptatifs, proposant une solution au problème de surcharge d'information par proposition de recommandations d'items. Le but de notre système de la recommandation sémantique était de retourner les documents qui sont susceptibles d'intéresser l'apprenant. Dans cette partie, nous avons présenté une nouvelle approche pour le calcul de la similarité sémantique entre les documents texte d'un corpus pour une recommandation sémantique basée sur le contenu.

Par ailleurs, nous avons proposé un système MASET qui vise essentiellement à aider les tuteurs à surveiller le travail collaboratif des apprenants à travers leurs diverses interactions. Puis, nous avons présenté une étude comparative de trois algorithmes de datamining (ID3, CART et C4.5) pour construire un modèle prédictif basé sur les arbres de décision afin de prédire le niveau des apprenants dans leur parcours d'apprentissage au cours d'une formation en ligne. Ce modèle contribue à une meilleure répartition des apprenants pour identifier ceux qui nécessitent une attention particulière de la part des tuteurs afin d'assurer un accompagnement efficace.

Certes, les contributions proposées sont loin d'être parfais. La partie suivante présentera un ensemble de perspectives qui vont permettre l'amélioration de nos contributions en enrichissant leurs fonctionnalités.

Perspectives

Le travail réalisé dans le cadre de cette thèse nous ouvre plusieurs perspectives de recherche intéressantes que nous comptons développer :

- Dans une première orientation, il s'agit d'appliquer notre approche dans plusieurs domaines d'études et d'étendre nos expérimentations à des applications sur des groupes d'apprenants.
- Nous comptons aussi comparer les résultats d'utilisation d'autres algorithmes dans la génération du parcours adapté.
- Dans une deuxième direction, nous comptons mettre en œuvre une ontologie pour l'évaluation des apprenants dans un environnement d'apprentissage en ligne.
- Dans une troisième direction, nous comptons analyser le feedback des apprenants pour générer le parcours adapté.
- Dans une quatrième direction, nous comptons exploiter les données massives extraites des réseaux sociaux pour contribuer à l'apprentissage adaptatif.

Bibliographie

[1] Louise Lafortune and Colette Daudelin. *Accompagnement Socioconstructiviste : Pour S'Approprier une Réforme en Éducation*, volume 3. PUQ, 2001.

[2] Cédric Jacquiot. *Modélisation logique et générique des systèmes d'hypermédias adaptatifs*. PhD thesis, Université Paris-Sud XI, 2006.

[3] Serge Garlatti and Yannick Prié. Adaptation et personnalisation dans le web sémantique. *Revue I3*, page 19, 2004.

[4] Patrick Mendelsohn and Pierre Dillenbourg. Le développement de l'enseignement intelligemment assisté par ordinateur. In *Symposium Intelligence Naturelle et Intelligence Artificielle*, 1991.

[5] Pierre Tchounikine. Platon-1 : quelques dimensions pour l'analyse des travaux de recherche en conception d'eiah. 2004.

[6] Tom Murray. Authoring intelligent tutoring systems : An analysis of the state of the art. *International Journal of Artificial Intelligence in Education (IJAIED)*, 10 :98–129, 1999.

[7] Gilbert Paquette, Jacqueline Bourdeau, France Henri, Josianne Basque, Michel Leonard, and Marcelo Maina. Construction d'une base de connaissances et d'une banque de ressources pour le domaine du téléapprentissage. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation (STICEF)*, 10, 2003.

[8] Emmanuel G Blanchard. *Motivation et culture en e-Learning*. Université de Montréal, 2007.

[9] Hugh L Burns and Charles G Capps. Foundations of intelligent tutoring systems : An introduction. *Foundations of intelligent tutoring systems*, pages 1–19, 1988.

[10] Henry M Halli. Curriculum and instruction in automated tutors. *Foundations of intelligent tutoring systems*, page 79, 2013.

[11] Gordon I McCalla and Jim E Greer. Granularity-based reasoning and belief revision in student models. In *Student modelling : The key to individualized knowledge-based instruction*, pages 39–62. Springer, 1994.

[12] Kurt VanLehn. Student modeling. *Foundations of intelligent tutoring systems*, 55 :78, 1988.

[13] James R Miller. The role of human-computer interaction in intelligent tutoring systems. *Foundations of intelligent tutoring systems*, pages 143–189, 1988.

[14] Alberto Amaral and Vincent Lynn Meek. *The higher education managerial revolution ?*, volume 3. Springer Science & Business Media, 2003.

[15] M Gallagher. E-learning in australia : universities and the new distance education. *Centre for Educational Research and Innovation OECD*, 2001.

[16] Annie Joyce Vullamparthi, Himadri S Khargharia, BS Bindhumadhava, and Nelaturu Sarat Chandra Babu. A smart tutoring aid for the autistic-educational aid for learners on the autism spectrum. In *Technology for Education (T4E), 2011 IEEE International Conference on*, pages 43–50. IEEE, 2011.

[17] Samir Bourekkache, Okba Kazar, Laïd Kahloul, Faiez Gargouri, and Benharkat Aïcha-Nabila. Un environnement sémantique à base d'agents pour la formation à distance (e-learning). In *10ième édition de la conférence sur Avancés des Systèmes Décisionnels-ASD 2016,*, 2016.

[18] Christian Depover and Louise Marchand. *E-learning et formation des adultes en contexte professionnel*. De Boeck Supérieur, 2002.

[19] Pierre Mœglin. A la recherche de l'industrialisation du tutorat à distance. *Distances et savoirs*, 3(2) :251–265, 2005.

[20] Emmanuel Houzé and Régis Meissonier. Performance du e-learning : de l'amélioration des résultats de l'apprenant à la prise en compte des enjeux institutionnels. *Systèmes d'Information et Management*, 10(4) :87, 2005.

[21] Pierre Dillenbourg, Charline Poirier, and Laure Carles. Communautés virtuelles d'apprentissage : e-jargon ou nouveau paradigme. *A. Taurisson et A. Sentini. Pédagogies. Net. Montréal, Presses*, pages 11–47, 2003.

[22] Jean-Philippe Pernin. A propos d'objets pédagogiques. In *Actes du colloque" Entre technique et pédagogie : la création de contenus multimédia pour l'enseignement et la formation*, pages 33–45, 2004.

[23] Fayrouz Soualah-Alila. *CAMLearn : Une Architecture de Système de Recommandation Sémantique Sensible au Contexte. Application au Domaine du M-Learning.* PhD thesis, Université de Bourgogne, 2015.

[24] Jean-Philippe Pernin. Objets pédagogiques : unités d'apprentissage, activités ou ressources. *Revue" Sciences et Techniques Educatives", Hors série*, pages 179–210, 2003.

[25] Jean-Philippe Pernin and Anne Lejeune. Dispositifs d'apprentissage instrumentés par les technologies : vers une ingénierie centrée sur les scénarios. In *Technologies de l'Information et de la Connaissance dans l'Enseignement Supérieur et de l'Industrie*, pages 407–414. Université de Technologie de Compiègne, 2004.

[26] Erik Duval and Wayne Hodgins. Learning objects revisited. *Online education using learning objects (open and flexible learning)*, pages 71–82, 2004.

[27] Katherine Nelson. Constraints on word learning ? *Cognitive Development*, 3(3) :221–246, 1988.

[28] Filip Neven and Erik Duval. Reusable learning objects : a survey of lom-based repositories. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 291–294. ACM, 2002.

[29] Christine Michel and Soufiane Rouissi. E-learning : normes et specifications. étude des spécifications lom et ims-qti caractérisant des documents numériques interéchangeables et réutilisables pour l'acquisition et l'évaluation des connaissances. *la*

Revue Document Numérique numéro spécial sur les nouvelles facettes du document électronique dans l'éducation, 2003.

[30] Edward R Jones. Implications of scorm^{TM} and emerging e-learning standards on engineering education. In *Proceedings of the 2002 ASEE Gulf-Southwest Annual Conference*, pages 20–22, 2002.

[31] Daniel Burgos, Michel Arnaud, Patrick Neuhauser, and Rob Koper. Ims learning design : la flexibilité pédagogique au service des besoins de la e-formation. 2005.

[32] Andreas Holzinger, Thomas Kleinberger, and Paul Muller. Multimedia learning systems based on ieee learning object metadata (lom). 2001.

[33] Dominique Guin and Luc Trouche. Un assistant méthodologique pour étayer le travail documentaire des professeurs : le cédérom sfodem 2008. *Repères IREM*, 72 :5–24, 2008.

[34] Nicolas Delestre, Nicolas Malandain, and Boulares Ouchenne. Une ontologie owl pour le cdm-fr. In *Conférence des Technologies de l'Information et de la Communication pour l'Enseignement*, 2014.

[35] Jacques Perriault. Le numérique : une question politique. *Hermès, La Revue*, (1) :183–189, 2004.

[36] Marie-Hélène Abel. Utilisation de normes et standards dans le projet memorae. *Distances et savoirs*, 2(4) :487–511, 2004.

[37] Jean Zahnd, François Hurter, Pierre-Olivier Vallat, and Haute Ecole Pédagogique BEJUNE. Intégration des tice dans l'apprentissage, une approche influencée par les objets pédagogiques. *Actes de la Recherche*, page 39, 2005.

[38] Anne Lejeune. Ims learning design. *Distances et savoirs*, 2(4) :409–450, 2004.

[39] Gilbert Paquette. *L'ingénierie pédagogique : pour construire l'apprentissage en réseau*. Puq, 2002.

[40] IMS GLC. Ims learning tools interoperability information model base document v0. 9.4. *IMS GLC*, 2007.

[41] Josep Blat, Toni Navarrete, Ayman Moghnieh, and Helena Batlle. A qti management system for service oriented architecture. 2006.

[42] Alberto Abelló Gamazo, María Elena Rodríguez González, Antoni Urpí Tubella, Xavier Burgués Illa, María José Casany Guerrero, Carme Martín Escofet, Quer Bosor, and Maria Carme. Learn-sql : automatic assesment of sql based on ims qti specification. In *8th IEEE International Conference on Advanced Learning Technologies*, pages 592–593, 2008.

[43] Xingchen Song. Teaching and learning experince with learning management systems : An adapted is success model in lms context. In *IT in Medicine and Education (ITME), 2011 International Symposium on*, volume 2, pages 148–152. IEEE, 2011.

[44] Moo-Chee Lee and Yun-Kung Chung. Using object-orientation to conceptualize an adaptive learning content management system modeling. In *Advanced Computer Control (ICACC), 2010 2nd International Conference on*, volume 3, pages 56–60. IEEE, 2010.

[45] Martin Dougiamas and Peter Taylor. Moodle : Using learning communities to create an open source course management system. 2003.

[46] Lesta A Burgess. Webct as an e-learning tool : A study of technology students' perceptions. 2003.

[47] Dossou Anani Koffi DOGBE-SEMANOU, Anne Durand, M Leproust, and H Vanderstichel. Etude comparative de plates-formes de formation à distance. *le cadre du Projet@ 2L Octobre*, 2007.

[48] open source. Claroline, une plate-forme d'enseignement et d'apprentissage pour stimuler le développement pédagogique des enseignants et la qualité des enseignements : premières approches. consulté le 17/06/2016.

[49] Open source. Ganesha, une plateforme e-learning avec migration au m-learning. consulté le 17/06/2016.

[50] Open source. Dokeos, une plateforme e-learning. consulté le 17/06/2016.

[51] Jean-François Bourdet. Tutorat en ligne et création d'un espace formatif. *TICE et Didactique des Langues Étrangères et Maternelles : la problématique des aides à l'apprentissage*, page 203, 2008.

[52] Sinclair Goodlad and Beverley Hirst. *Peer Tutoring. A Guide to Learning by Teaching.* ERIC, 1989.

[53] Danielle Paquette. Le rôle des tuteurs et des tutrices : une diversité à appréhender. *Distances*, 5(1) :7–35, 2001.

[54] René Chalon. *Réalité mixte et travail collaboratif : IRVO, un modèle de l'interaction homme-machine.* PhD thesis, Ecole Centrale de Lyon, 2004.

[55] Marc Walckiers and Thomas De Praetere. L'apprentissage collaboratif en ligne, huit avantages qui en font un must. *Distances et savoirs*, 2(1) :53–75, 2004.

[56] Robert E Slavin. *Cooperative learning : Student teams. What research says to the teacher.* ERIC, 1982.

[57] François Mangenot. Tâches et coopération dans deux dispositifs universitaires de formation à distance. *Apprentissage des langues et systèmes d'information et de communication*, 6(1) :109–125, 2003.

[58] Angélique Dimitracopoulou. Analyse automatique des interactions pour le soutien à l'auto-régulation des participants dans des activités médiées.

[59] Toshio Mochizuki, Hiroshi Kato, Kazaru Yaegashi, Tomoko Nagata, Toshihisa Nishimori, Shin-ichi Hisamatsu, Satoru Fujitani, Jun Nakahara, and Mariko Suzuki. Promotion of self-assessment for learners in online discussion using the visualization software. In *Proceedings of th 2005 conference on Computer support for collaborative learning : learning 2005 : the next 10 years !*, pages 440–449. International Society of the Learning Sciences, 2005.

[60] Tharrenos Bratitsis and Angelique Dimitracopoulou. Indicators for measuring quality in asynchronous discussion forums. In *International Conference on Cognition and Exploratory Learning in Digital Era (CELDA2006), IADIS (International Association for Development of the Information Society.* Citeseer, 2006.

[61] Chris Teplovs and Marlene Scardamalia. Visualizations for knowledge building assessment. In *AgileViz workshop, CSCL*, 2007.

[62] Denise Barbeau et al. *Analyse de déterminants et d'indicateurs de la motivation scolaire d'élèves du collégial*. Collège de Bois-de-Boulogne ; Programme d'aide à la recherche sur l'enseignement et l'apprentissage,, 1994.

[63] Marco Aurélio Gerosa, Mariano Gomes Pimentel, Hugo Fuks, and Carlos JP Lucena. No need to read messages right now : helping mediators to steer educational forums using statistical and visual information. In *Proceedings of th 2005 conference on Computer support for collaborative learning : learning 2005 : the next 10 years !*, pages 160–169. International Society of the Learning Sciences, 2005.

[64] Pablo Reyes Cabrera. *Structural awareness in mediated conversations for collaborative learning environments*. PhD thesis, Le Mans, 2005.

[65] Laurie P Dringus and Timothy Ellis. Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education*, 45(1) :141–160, 2005.

[66] Tharrenos Bratitsis and Angelique Dimitrakopoulou. Data recording and usage interaction analysis in asynchronous discussions : The dias system. In *12th International Conference on Artificial Intelligence in Education AIED*, 2005.

[67] Alejandra Martinez, P De la Fuente, and Y Dimitriadis. Towards an xml-based representation of collaborative action. In *Designing for change in networked learning environments*, pages 379–383. Springer, 2003.

[68] Hichang Cho, Michael Stefanone, and Geri Gay. Social information sharing in a cscl community. In *Proceedings of the Conference on Computer Support for Collaborative Learning : Foundations for a CSCL Community*, pages 43–50. International Society of the Learning Sciences, 2002.

[69] Jack Mezirow et al. Learning to think like an adult. *Learning as transformation : Critical perspectives on a theory in progress*, pages 3–33, 2000.

[70] Peter Reimann. How to support groups in learning : More than problem solving. *Artificial Intelligence in Education (AIED 2003). Supplementary Proceedings*, pages 3–16, 2003.

[71] Valerie Shute and Brendon Towle. Adaptive e-learning. *Educational Psychologist*, 38(2) :105–114, 2003.

[72] Roger Nkambou, Riichiro Mizoguchi, and Jacqueline Bourdeau. *Advances in intelligent tutoring systems*, volume 308. Springer Science & Business Media, 2010.

[73] Nicolas Delestre. *Metadyne, un hypermédia adaptatif dynamique pour l'enseignement*. PhD thesis, Université de Rouen, 2000.

[74] Aziz Dahbi, Abdelghafour Berraissoul, et al. Conception d'un système hypermédia d'enseignement adaptatif centré sur les styles d'apprentissage : modèle et expérience. *Revue internationale des technologies en pédagogie universitaire/International Journal of Technologies in Higher Education*, 6(1) :55–71, 2009.

[75] Nicolas Delestre, Jean-Pierre Pécuchet, and Catherine Gréboval. L'architecture d'un hypermédia adaptatif dynamique pour l'enseignement. In *Nouvelles Technologies pour l'Information et le Communication dans les Formations d'Ingénieurs-NTICF'98*, pages p383–390, 1998.

[76] Aymeric Paccou, Guillaume Chiavassa, Jacques Liandrat, and Kai Schneider. A penalization method applied to the wave equation. *Comptes Rendus Mécanique*, 333(1) :79–85, 2005.

[77] Olivier Ridoux and Thierry Viéville. À propos de dualités en sciences et technologies de l'information et de la communication. *Interstices*, 2005.

[78] Robert M Bollet and Santiago Fallon. Personalizing e-learning. *Educational Media International*, 39(1) :39–45, 2002.

[79] Chih-Ming Chen, Hahn-Ming Lee, and Ya-Hui Chen. Personalized e-learning system using item response theory. *Computers & Education*, 44(3) :237–255, 2005.

[80] Marlène Villanova-Oliver. *Adaptabilité dans les systèmes d'information sur le Web : modélisation et mise en oeuvre de l'accès progressif.* PhD thesis, Institut National Polytechnique de Grenoble-INPG, 2002.

[81] Paul De Bra, Peter Brusilovsky, and Geert-Jan Houben. Adaptive hypermedia : from systems to framework. *ACM Computing Surveys (CSUR)*, 31(4es) :12, 1999.

[82] Amal Battou. Approche granulaire des objets pédagogiques en vue de l'adaptabilité dans le cadre des environnements informatiques pour l'apprentissage humain. 2012.

[83] Azough Samia. E-learning adaptatif : Gestion intélligente des ressources pédagogiques et adaptation de la formation au profil de l'apprenant. 2014.

[84] Thomas Bäck, DB Fogel, and Z Michalewicz. Handbook of evolutionary computation. *Release*, 97(1) :B1, 1997.

[85] Marie Duflo. *Algorithmes stochastiques.* 1996.

[86] Ingo Rechenberg. Evolution strategy and human decision making. *Human decision making and manual control*, pages 349–359, 1878.

[87] David B Fogel. *Artificial intelligence through simulated evolution.* Wiley-IEEE Press, 2009.

[88] John Holland. Les algorithmes génétiques. *Revue Pour La Science*, 179 :44–51, 1992.

[89] David E Goldberg. *Genetic algorithms.* Pearson Education India, 2006.

[90] John H Holland. Genetic algorithms. *Scientific american*, 267(1) :66–72, 1992.

[91] Gabriel A Dover and Charles Darwin. *Dear Mr. Darwin : Letters on the evolution of life and human nature.* Univ of California Press, 2000.

[92] Christopher R Houck, Jeff Joines, and Michael G Kay. A genetic algorithm for function optimization : a matlab implementation. *NCSU-IE TR*, 95(09), 1995.

[93] Jean-Marc Alliot and Nicolas Durand. Algorithmes génétiques. *Centre d'Etudes de la Navigation Aérienne*, 2005.

[94] Florence Amardeilh. *Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle.* PhD thesis, Université de Nanterre-Paris X, 2007.

[95] Francis Pisani and Dominique Piotet. *Comment le web change le monde : l'alchimie des multitudes.* Pearson Education France, 2008.

[96] Sareh Aghaei, Mohammad Ali Nematbakhsh, and Hadi Khosravi Farsani. Evolution of the world wide web : From web 1.0 to web 4.0. *International Journal of Web & Semantic Technology*, 3(1) :1, 2012.

[97] Tim O'reilly. What is web 2.0 : Design patterns and business models for the next generation of software. *Communications & strategies*, (1) :17, 2007.

[98] Veronica Barassi and Emiliano Treré. Does web 3.0 come after web 2.0 ? deconstructing theoretical assumptions through practice. *New media & society*, 14(8) :1269–1285, 2012.

[99] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. *The Semantic Web, Scientific American*, 2001.

[100] Emmanuelle Bermes and Gautier Poupeau. Les technologies du web appliquées aux données structurées. In *Séminaire IST Inria : le document numérique à l'heure du web de données*, pages 41–84. ADBS, 2012.

[101] Tim Bray, Jean Paoli, C Michael Sperberg-McQueen, Eve Maler, and François Yergeau. Extensible markup language (xml). *World Wide Web Consortium Recommendation REC-xml-19980210. http ://www. w3. org/TR/1998/REC-xml-19980210*, 16 :16, 1998.

[102] Sharon Adler, Alex Milowski, Jeremy Richman, Steve Zilles, et al. Extensible stylesheet language (xsl)-version 1.0. 2001.

[103] XML Schema and XML Schema. W3c recommendation. *World Wide Web Consortium (W3C)*, 2001.

[104] Dan Brickley and Ramanathan V Guha. Resource description framework (rdf) schema specification 1.0 : W3c candidate recommendation 27 march 2000. 2000.

[105] Fabien Gandon, Reto Krummenacher, Sung-Kook Han, and Ioan Toma. The resource description framework and its schema. *Handbook of Semantic Web Technologies*, 2011.

[106] Dan Brickley and Ramanathan V Guha. Resource description framework (rdf) schema specification 1.0 : W3c candidate recommendation 27 march 2000. 2000.

[107] Vladimir Tomberg and Mart Laanpere. Rdfa versus microformats : exploring the potential for semantic interoperability of mash-up personal learning environments. In *CEUR Workshop Proceedings*, volume 506, pages 102–109, 2009.

[108] Sean Bechhofer. Owl : Web ontology language. In *Encyclopedia of Database Systems*, pages 2008–2009. Springer, 2009.

[109] Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10) :2004, 2004.

[110] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet : A practical owl-dl reasoner. *Web Semantics : science, services and agents on the World Wide Web*, 5(2) :51–53, 2007.

[111] Grigoris Antoniou and Frank Van Harmelen. Web ontology language : Owl. In *Handbook on ontologies*, pages 67–92. Springer, 2004.

[112] François Goasdoué, Ioana Manolescu, and Alexandra Roatis. Répondre aux requêtes par reformulation dans les bases de données rdf. In *RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle)*, pages 978–2, 2012.

[113] Michèle Vialatte. *Description et applications du moteur d'inférence SNARK*. PhD thesis, 1985.

[114] Volker Haarslev and Ralf Möller. Racer : A core inference engine for the semantic web. In *EON*, volume 87, 2003.

[115] Youyong Zou, Tim Finin, and Harry Chen. F-owl : An inference engine for semantic web. In *International Workshop on Formal Approaches to Agent-Based Systems*, pages 238–248. Springer, 2004.

[116] Christine Golbreich. Combining rule and ontology reasoners for the semantic web. In *International Workshop on Rules and Rule Markup Languages for the Semantic Web*, pages 6–22. Springer, 2004.

[117] T Huang, W Li, and C Yang. Comparison of ontology reasoners : Racer, pellet, fact++. In *AGU Fall Meeting Abstracts*, volume 1, page 1068, 2008.

[118] Neha Dalwadi, Bhaumik Nagar, and Ashwin Makwana. Semantic web and comparative analysis of inference engines. *Int. J. of Computer Science and Information Technologies*, 3(3) :3843–3847, 2012.

[119] Georgios Meditskos and Nick Bassiliades. A rule-based object-oriented owl reasoner. *IEEE Transactions on Knowledge and Data Engineering*, 20(3) :397–410, 2008.

[120] Michael Kifer. Rule interchange format : The framework. In *International Conference on Web Reasoning and Rule Systems*, pages 1–11. Springer, 2008.

[121] Ian Horrocks, Peter F Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, Mike Dean, et al. Swrl : A semantic web rule language combining owl and ruleml. *W3C Member submission*, 21 :79, 2004.

[122] Philippe Fournier-Viger. *Un modèle de représentation des connaissances à trois niveaux de sémantique pour les sytèmes tutoriels intelligents*. Université de Sherbrooke., 2005.

[123] Franz Baader. *The description logic handbook : Theory, implementation and applications*. Cambridge university press, 2003.

[124] Paulo Pinheiro Da Silva, Deborah L McGuinness, and Richard Fikes. A proof markup language for semantic web services. *Information Systems*, 31(4) :381–395, 2006.

[125] Ian Horrocks et al. Daml+oil : A description logic for the semantic web. *IEEE Data Eng. Bull.*, 25(1) :4–9, 2002.

[126] Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. Trust management for the semantic web. In *International semantic Web conference*, pages 351–368. Springer, 2003.

[127] O Dubois. La signature numérique. *Spectra biologie*, 16(91) :35–36, 1997.

[128] Jean Charlet, Philippe Laublet, and Chantal Reynaud. *Le web sémantique*. Cépaduès-Ed., 2003.

[129] Jean Charlet, Bruno Bachimont, and Raphaël Troncy. Ontologies pour le web sémantique. *Revue Information, Interaction, Intelligence I3*, 2004.

[130] Frédéric Fürst. *Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationalisation*. PhD thesis, Nantes, 2004.

[131] Thomas R Gruber et al. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2) :199–220, 1993.

[132] Nicola Guarino. Formal ontology and information systems. In *Proceedings of FOIS*, volume 98, pages 81–97, 1998.

[133] Gerardus Adrianus Cornelis Maria van Heijst. *The role of ontologies in knowledge engineering*. Universiteit van Amsterdam, Faculteit der Psychologie, 1995.

[134] Mouhamadou Saliou Diallo, Moussa Lo, Cheikh Talibouya Diop, and Fatou Kamara Sangaré. Etat de l'art sur ontologies et extraction de connaissances.

[135] Nabila Chergui, Razika Driouche, and Zizette Boufaida. Un processus de mapping basé contexte pour l'intégration des ontologies. In *Actes du Cinquième Colloque sur l'Optimisation et les Systèmes d'Information COSI'2008*, page 263.

[136] Thomas R Gruber. *Ontolingua : A mechanism to support portable ontologies*, volume 27. Citeseer, 1992.

[137] Ricardo De Almeida Falbo, Ana Candida Cruz Natali, Paula Gomes Mian, Gleidson Bertollo, and Fabiano Borges Ruy. Ode : Ontology-based software development environment. In *IX Congreso Argentino de Ciencias de la Computación*, 2003.

[138] Natalya F Noy, Monica Crubézy, Ray W Fergerson, Holger Knublauch, Samson W Tu, Jennifer Vendetti, Mark A Musen, et al. Protege-2000 : an open-source ontology-development and knowledge-acquisition environment. In *AMIA Annu Symp Proc*, volume 953, page 953, 2003.

[139] Valéry Psyché, Olavo Mendes, and Jacqueline Bourdeau. Apport de l'ingénierie ontologique aux environnements de formation à distance contribution of ontological engineering to distance learning environments.

[140] Faiçal Azouaou. *Modèles et outils d'annotations pour une mémoire personnelle de l'enseignant*. PhD thesis, Université Joseph-Fourier-Grenoble I, 2006.

[141] Sylvain Dehors, Catherine Faron-Zucker, Jean Paul Stromboni, and Alain Giboin. Des annotations sémantiques pour apprendre : l'expérimentation qbls. *Actes de la Journée thématique WebLearn, plate-forme AFIA*, 31, 2005.

[142] Ricardo R Amorim, Manuel Lama, Eduardo Sánchez, Adolfo Riera, and Xosé A Vila. A learning design ontology based on the ims specification. *Educational Technology & Society*, 9(1) :38–57, 2006.

[143] Suzanne Kabel, Robert De Hoog, Bob Wielinga, and Anjo Anjewierden. Indexing learning objects : Vocabularies and empirical investigation of consistency. *Journal of Educational Multimedia and Hypermedia*, 13(4) :405, 2004.

[144] Arezki Hammache and Rachid Ahmed-Ouamer. Système d'inférence pour une indexation de documents basée sur une ontologie de domaine. In *INFORSID*, pages 895–910, 2006.

[145] Badr Hssina, Belaid Bouikhalene, and Abdelkrim Merbouha. An ontology to assess the performances of learners in an e-learning platform based on semantic web technology : Moodle case study. In *Europe and MENA Cooperation Advances in Information and Communication Technologies*, pages 103–112. Springer, 2017.

[146] Brian McBride. Jena : A semantic web toolkit. *IEEE Internet computing*, 6(6) :55, 2002.

[147] Robin Burke. Hybrid recommender systems : Survey and experiments. *User modeling and user-adapted interaction*, 12(4) :331–370, 2002.

[148] Henri Isaac, Eric Campoy, and Michel Kalika. Surcharge informationnelle, urgence et tic. l'effet temporel des technologies de l'information. *Management & Avenir*, (3) :149–168, 2007.

[149] Chumki Basu, Haym Hirsh, William Cohen, et al. Recommendation as classification : Using social and content-based information in recommendation. In *Aaai/iaai*, pages 714–720, 1998.

[150] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[151] Amir Albadvi and Mohammad Shahbazi. A hybrid recommendation technique based on product category attributes. *Expert Systems with Applications*, 36(9) :11480–11488, 2009.

[152] An-Te Nguyen, Nathalie Denos, Catherine Berrut, and Bich-Thuy Dong Thi. Modèle de formation multiple de communautés dans un système de recommandation hybride. INFORSID, 2006.

[153] Mark Claypool, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes, and Matthew Sartin. Combining content-based and collaborative filters in an online newspaper. In *Proceedings of ACM SIGIR workshop on recommender systems*, volume 60. Citeseer, 1999.

[154] Bamshad Mobasher, Xin Jin, and Yanzan Zhou. Semantically enhanced collaborative filtering on the web. In *Web Mining : From Web to Semantic Web*, pages 57–76. Springer, 2004.

[155] Justin Basilico and Thomas Hofmann. Unifying collaborative and content-based filtering. In *Proceedings of the twenty-first international conference on Machine learning*, page 9. ACM, 2004.

[156] Zi-Ke Zhang, Chuang Liu, Yi-Cheng Zhang, and Tao Zhou. Solving the cold-start problem in recommender systems with social tags. *EPL (Europhysics Letters)*, 92(2) :28002, 2010.

[157] Marko Balabanović and Yoav Shoham. Fab : content-based, collaborative recommendation. *Communications of the ACM*, 40(3) :66–72, 1997.

[158] Sonia Ben Ticha. *Recommandation personnalisée hybride*. PhD thesis, Université de Lorraine, 2015.

[159] Hendrik Drachsler, Hans GK Hummel, and Rob Koper. Personal recommender systems for learners in lifelong learning networks : the requirements, techniques and model. *International Journal of Learning Technology*, 3(4) :404–423, 2008.

[160] Osmar R Zaíane. Building a recommender agent for e-learning systems. In *Computers in Education, 2002. Proceedings. International Conference on*, pages 55–59. IEEE, 2002.

[161] JESUS Bobadilla, Francisco Serradilla, Antonio Hernando, et al. Collaborative filtering adapted to recommender systems of e-learning. *Knowledge-Based Systems*, 22(4) :261–265, 2009.

[162] Katrien Verbert, Nikos Manouselis, Xavier Ochoa, Martin Wolpers, Hendrik Drachsler, Ivana Bosnic, and Erik Duval. Context-aware recommender systems for learning : a survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4) :318–335, 2012.

[163] Badr Hssina, Belaid Bouikhalene, and Abdelkrim Merbouha. A hybrid approach of semantic similarity calculation for a content-based recommendation of text documents on an e-learning platform. *International Arab Conference on Information Technology*, (88), 2016.

[164] Muriel Amar. *Les fondements théoriques de l'indexation : une approche linguistique*. PhD thesis, Lyon 2, 1997.

[165] Farah Harrathi. Extraction de concepts et de relations entre concepts à partir des documents multilingues : approche statistique et ontologique. *month*, 2009.

[166] Ludovic Tanguy. *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle informatique de la sémantique interprétative*. PhD thesis, Université de Rennes 1, 1997.

[167] Christian Jacquemin, Béatrice Daille, Jean Royauté, and Xavier Polanco. In vitro evaluation of a program for machine-aided indexing. *Information processing & management*, 38(6) :765–792, 2002.

[168] Mohamed Mohsen Gammoudi. *Méthode de Décomposition Rectangulaire d'une Relation Binaire : Une base formelle et uniforme pour la génération automatique des thesaurus et la recherche documentaire*. PhD thesis, 1993.

[169] Mike Scott and Christopher Tribble. *Textual patterns : Key words and corpus analysis in language education*, volume 22. John Benjamins Publishing, 2006.

[170] Ghassan Mourad. *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique de citations : réalisation des applications informatiques : SegATex et CitaRE*. PhD thesis, Paris 4, 2001.

[171] Benoît Lemaire. Limites de la lemmatisation pour l'extraction de significations. In *9e Journées internationales d'Analyse Statistique des Données Textuelles*, pages 725–732, 2008.

[172] Julie B Lovins. *Development of a stemming algorithm*. MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.

[173] Julie B Lovins. *Development of a stemming algorithm*. MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.

[174] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3) :130–137, 1980.

[175] Mustapha Baziz. *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. PhD thesis, Toulouse 3, 2005.

[176] Gerard Salton, Edward A Fox, and Harry Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11) :1022–1036, 1983.

[177] Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3) :129–146, 1976.

[178] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5) :513–523, 1988.

[179] Juan Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 2003.

[180] Pranas Zunde and Margaret E Dexter. Indexing consistency and quality. *American Documentation*, 20(3) :259–267, 1969.

[181] Peter HA Sneath, Robert R Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

[182] Yufeng Jing and W Bruce Croft. An association thesaurus for information retrieval. In *Intelligent Multimedia Information Retrieval Systems and Management-Volume 1*, pages 146–160. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 1994.

[183] Z Bellia, N Vincent, S Kirchner, and G Stamon. Assignation automatique de solutions à des classes de plaintes liées aux ambiances intérieures polluées. 8èmes journées d'extraction et de gestion des connaissances (egc 2008). *Sophia-Antipolis*, 2008.

[184] Alain Polguère. Remarques sur les réseaux sémantiques sens-texte. *A. Clas (éd.), Le mot, les mots, les bons mots, Presses de l'Université de Montréal*, 1992.

[185] George A Miller. Wordnet : a lexical database for english. *Communications of the ACM*, 38(11) :39–41, 1995.

[186] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.

[187] M Jacobzone and G Carbonnel. Coefficient de jaccard et coefficient de correlation. application aux ostracodes miocènes. *Paléoécologie des ostracodes*, pages 167–177, 1971.

[188] Henk Wolda. Similarity indices, sample size and diversity. *Oecologia*, 50(3) :296–302, 1981.

[189] Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1) :59–66, 1988.

[190] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, 2006.

[191] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64, 2000.

[192] V Bavi, T Beirne, N Bone, J Mohr, and B Neal. Comparison of document similarity metrics. *Computer Science Department, Western Washington University, Information Retrieval, Winter*, 2010.

[193] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[194] Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1) :188–230, 2004.

[195] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611, 2007.

[196] Joseph John Rocchio. Relevance feedback in information retrieval. 1971.

[197] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM, 1986.

[198] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[199] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, 19(1) :17–30, 1989.

[200] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[201] Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet : An electronic lexical database*, 305 :305–332, 1998.

[202] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[203] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304. Citeseer, 1998.

[204] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[205] Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet : An electronic lexical database*, 49(2) :265–283, 1998.

[206] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1) :5–53, 2004.

[207] Stéphane Chaudiron and Madjid Ihadjadene. Quelle place pour l'usager dans l'évaluation des sri ? In *Recherches récentes en sciences de l'information : Convergences et dynamiques*, pages 211–231. ADBS Editions, 2002.

[208] Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16. ACM, 2005.

[209] David E Goldberg and H John. Holland. genetic algorithms and machine learning. *Machine learning*, 3(2-3) :95–99, 1988.

[210] Jacques Ferber. *Objets et agents : une étude des structures de représentation et de communications en Intelligence Artificielle*. PhD thesis, Paris 6, 1989.

[211] Stuart Jonathan Russell, Peter Norvig, John F Canny, Jitendra M Malik, and Douglas D Edwards. *Artificial intelligence : a modern approach*, volume 2. Prentice hall Upper Saddle River, 2003.

[212] Pattie Maes. Artificial life meets entertainment : lifelike autonomous agents. *Communications of the ACM*, 38(11) :108–114, 1995.

[213] Jacques Ferber. Les systèmes multi-agents : un aperçu général. *Techniques et sciences informatiques*, 16(8), 1997.

[214] Sofiane Labidi and Wided Lejouad. *De l'intelligence artificielle distribuée aux systèmes multi-agents*. PhD thesis, INRIA, 1993.

[215] Brahim Chaib-Draa, Imed Jarras, and Bernard Moulin. Systèmes multi-agents : principes généraux et applications. *Principes et architectures des systèmes multi-agents*, 2001.

[216] Vladimir Gorodetski, Oleg Karsayev, Igor Kotenko, and Alexey Khabalov. Software development kit for multi-agent systems design and implementation. In *International Workshop of Central and Eastern Europe on Multi-Agent Systems*, pages 121–130. Springer, 2001.

[217] Fabio Bellifemine, Federico Bergenti, Giovanni Caire, and Agostino Poggi. Jade—a java agent development framework. In *Multi-Agent Programming*, pages 125–147. Springer, 2005.

[218] TCC FIPA. Fipa acl communicative act library specification component, foundation for intelligent physical agents, 2002.

[219] Duc An VO and Alexis DROGOUL. *Implantation des protocoles de communication FIPA dans la plate-forme GAMA*. PhD thesis, 2008.

[220] Lars Braubach, Winfried Lamersdorf, and Alexander Pokahr. Jadex : Implementing a bdi-infrastructure for jade agents. 2003.

[221] Fabio Bellifemine, Agostino Poggi, and Giovanni Rimassa. Jade–a fipa-compliant agent framework. In *Proceedings of PAAM*, volume 99, page 33. London, 1999.

[222] Maria Virvou and Katerina Kabassi. F-smile : An intelligent multi-agent learning environment. In *Proceedings of 2002 IEEE International Conference on Advanced Learning Technologies-ICALT*. Citeseer, 2002.

[223] Dimitar Nedev and Veselina Nedeva. Aspects of multi-agent system application in e-learning. In *International Scientific Conference Computer Science*, pages 18–19, 2008.

[224] Éric Alamartine, Damien Thibaudin, Nicolas Maillard, Catherine Sauron, Manolie Mehdi, Christian Broyet, and Christophe Mariat. Télémédecine : une expérience infructueuse de télé-expertise en néphrologie. *La Presse Médicale*, 39(5) :e112–e116, 2010.

[225] Badr Hssina, Belaid Bouikhalene, and Abdelkrim Merbouha. Towards an e-learning platform multi-agent based on the e-tutoring for collaborative work. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 9(4) :978–982, 2015.

[226] Fatiha Aityacine, Badr Hssina, and Belaid Bouikhalene. Jade multi-agent middleware applied to contribute to certificate management of students. *Indonesian Journal of Electrical Engineering and Computer Science*, 16(1) :176–181, 2015.

[227] Badr Hssina, Abdelkrimand Merbouha, and Belaid Bouikhalene. Predicting learners' performance in an e-learning platform based on decision tree analysis. In *International Arab Conference on Information Technology*. ACIT, 2016.

[228] Badr HSSINA, Abdelkarim Merbouha, Hanane Ezzikouri, and Mohammed Erritali. A comparative study of decision tree id3 and c4. 5. *Int. J. Adv. Comput. Sci. Appl*, 4(2), 2014.

[229] Xiaojin Zhu. Semi-supervised learning literature survey. 2005.

[230] Zhengdong Lu and Todd K Leen. Semi-supervised learning with penalized probabilistic clustering. In *Advances in neural information processing systems*, pages 849–856, 2004.

[231] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

[232] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. 1990.

[233] John Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine learning*, 4(2) :227–243, 1989.

[234] Ricco Rakotomalala. Arbres de décision. *Revue Modulad*, 33 :163–187, 2005.

[235] Devashsish Thakur, Nisarga Markandaiah, and D Sharan Raj. Re optimization of id3 and c4. 5 decision tree. In *Computer and Communication Technology (ICCCT), 2010 International Conference on*, pages 448–450. IEEE, 2010.

[236] Berthold Lausen, Willi Sauerbrei, and Martin Schumacher. Classification and regression trees (cart) used for the exploration of prognostic factors measured on different scales. 1994.

[237] Kemal Polat and Salih Güneş. A novel hybrid intelligent method based on c4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, 36(2) :1587–1592, 2009.

[238] Victor Francisco Rodriguez-Galiano, Bardan Ghimire, John Rogan, Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67 :93–104, 2012.

[239] Christiane Ferreira Lemos Lima, Francisco Marcos de Assis, and Cleonilson Protásio de Souza. Decision tree based on shannon, renyi and tsallis entropies for intrusion tolerant systems. In *Internet Monitoring and Protection (ICIMP), 2010 Fifth International Conference on*, pages 117–122. IEEE, 2010.

[240] Usama M Fayyad and Keki B Irani. The attribute selection problem in decision tree generation. In *AAAI*, pages 104–110, 1992.

[241] Johan Baltié, SCIA Specialisation, and Responsable M Adjaoute. Datamining : Id3 et c4. 5. *Epita SCIA*, 2002.

[242] Surjeet Kumar Yadav and Saurabh Pal. Data mining : A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv :1203.3832*, 2012.