Extraction Automatique d'Informations Pédagogiques Pertinentes à partir de Documents Textuels

Boutheina Smine1,2 Rim Faiz2 Jean-Pierre Desclés1

(1) LaLIC, Université Paris-Sorbonne, 28 rue Serpente, 75006 Paris, France.
Boutheina.Smine@etudiants.univ-paris4.fr , Jean-pierre.Descles@paris4.sorbonne.fr
(2) LaRODEC, IHEC de Carthage, 2016 Carthage Présidence, Tunisie. Rim.Faiz@ihec.rnu.tn

RÉSUMÉ. Plusieurs utilisateurs ont souvent besoin d'informations pédagogiques pour les intégrer dans leurs ressources pédagogiques, ou pour les utiliser dans un processus d'apprentissage. Une indexation de ces informations s'avère donc utile en vue d'une extraction des informations pédagogiques pertinentes en réponse à une requête utilisateur. La plupart des systèmes d'extraction d'informations pédagogiques existants proposent une indexation basée sur une annotation manuelle ou semi-automatique des informations pédagogiques, tâche qui n'est pas préférée par les utilisateurs. Dans cet article, nous proposons une approche d'indexation d'objets pédagogiques (Définition, Exemple, Exercice, etc.) basée sur une annotation sémantique par Exploration Contextuelle des documents. L'index généré servira à une extraction des objets pertinents répondant à une requête utilisateur sémantique. Nous procédons, ensuite, à un classement des objets extraits selon leur pertinence en utilisant l'algorithme Rocchio. Notre objectif est de mettre en valeur une indexation à partir de contextes sémantiques et non pas à partir de seuls termes linguistiques.

ABSTRACT. Different users need pedagogical information in order to use them in their resources or in a learning process. Indexing this information is therefore useful for extracting relevant pedagogical information in response to a user request. Several searching systems of pedagogical information propose manual or semi-automatic annotations to index documents, which is a complex task for users. In this article, we propose an approach to index pedagogical objects (Definition, Exercise, Example, etc.) based on automatic annotation of documents using Contextual Exploration. Then, we use the index to extract relevant pedagogical objects as response to the user's requests. We proceed to sort the extracted objects according to their relevance. Our objective is to reach the relevant objects using a contextual semantic analysis of the text.

MOTS-CLÉS : extraction d'informations, objets pédagogiques, carte sémantique, exploration contextuelle, algorithme Rocchio

KEYWORDS : Information retrieval, pedagogical objects, semantic map, Contextual Exploration, Rocchio algorithm

1 Introduction

Avec la croissance rapide de la quantité d'information disponible en ligne et dans les bases de données, les moteurs de recherche jouent un rôle important dans l'apprentissage en ligne, car ils peuvent soutenir l'apprenant dans sa recherche d'informations nécessaires à son apprentissage, à sa formation, etc. Toutefois, ces systèmes de recherche d'information sont basés sur l'indexation des termes sans tenir compte de la sémantique du contenu pédagogique (Dehors et al., 2005), (Buffa et al., 2005). Une meilleure alternative est de proposer une approche d'indexation basée sur l'annotation sémantique des informations pédagogiques qui sont attestées dans les documents. Par une telle indexation, les informations pédagogiques présentées par l'auteur d'un document sont capturées et le processus d'apprentissage ou d'enseignement pour l'élève ou l'enseignant respectivement est facilité.

Nous proposons, dans cet article, une approche d'indexation automatique d'informations pédagogiques à partir de documents. Notre travail consiste d'abord à annoter les segments textuels (objets) reflétant un contenu pédagogique (Définition, Exemple, Exercice, etc.). Ensuite, nous procédons à une indexation de ces objets annotés pour extraire ceux qui sont pertinents par rapport à une requête utilisateur. Enfin, nous procédons à un classement de ces objets en utilisant l'algorithme de classification Rocchio.

Dans la section 2, nous positionnons cette contribution par rapport aux travaux existants. Nous consacrons la section 3 à la définition de la notion d'objet pédagogique. Une description détaillée de notre approche d'indexation d'informations pédagogiques est le sujet de la quatrième section. Avant de conclure, nous illustrons les résultats des expérimentations de notre approche dans la cinquième section.

2 Etat des lieux autour de la recherche d'informations pédagogiques

Nous détaillons ici divers points de l'état de l'art liés à notre approche d'indexation d'objets pédagogiques, à savoir l'annotation, l'indexation, et l'extraction d'informations pédagogiques à partir de documents textuels.

L'annotation comme technique d'indexation est appliquée dans plusieurs systèmes comme le système QBLS (Dehors et al., 2005) qui est une partie de la plateforme TRIAL SOLUTION (Buffa et al., 2005). Dans cette dernière, les utilisateurs annotent les livres manuellement selon le rôle pédagogique de leur contenu, les sujets abordés dans leur contenu (mots clés) et leurs relations avec d'autres ressources (référence, prérequis, etc.). Le système QBLS a pour but de structurer le cours en se référant à une ontologie pédagogique constituée de fiches (définition, exemple, énoncé, procédure, solution, etc.) et de ressources pédagogiques abstraites (cours, thème, notion, question). Il existe aussi le système SYFAX (Smei et al., 2005) qui annote semi-automatiquement le document pédagogique selon plusieurs critères (type du document, point de vue de l'utilisateur sur le document, etc.).

En vue d'indexer les documents, les annotations proposées par les différents systèmes cités ci-dessus sont stockées dans un entrepôt de connaissances pédagogiques. Par la suite, les réponses aux requêtes sont extraites à partir de cet entrepôt grâce à un moteur de recherche (*Corese* pour le système QBLS). Le système SYFAX applique un processus de raffinement de la requête basé sur une ontologie des types de documents pédagogiques et une autre ontologie des domaines des documents informatiques. Ceci permet d'extraire les documents pertinents par rapport à la requête.

Pour tous les systèmes présentés ci-dessus, une intervention humaine est requise afin d'enrichir les documents par des métadonnées. Cependant, la plupart des producteurs de ressources pédagogiques ne s'intéressent probablement pas au retour aux documents pour annoter leurs propres travaux. Notre travail se place dans cette perspective tout en procédant à l'automatisation du processus d'annotation.

D'autres travaux se sont intéressés à la recherche de ressources pédagogiques à partir du web (Thomson et al., 2003). Toutefois, le but de leur travail est limité à une extraction de métadonnées (Travaux Dirigés, Programme, Travaux Pratiques) relatives au document en entier en vue de les annoter et de les classifier. Toujours dans la même perspective, (Hassen et al., 2009) comparent l'efficacité des algorithmes Naïve Bayes et SVM dans la classification des ressources pédagogiques basée sur un ensemble de propriétés (catégorie du contenu, titre du cours, année, auteur, etc.).

A notre connaissance, ces travaux de recherche portant sur l'indexation de documents pédagogiques se sont intéressés à une indexation du document en l'annotant par un ensemble de métadonnées relatives à la totalité du document. D'autres approches basées sur des patrons linguistiques ont été appliquées dans plusieurs travaux pour extraire les définitions à partir de ressources pédagogiques afin de constituer un glossaire (Westerhout et al., 2008) ou encore pour répondre à divers types de questions (Greenwood et al., 2003). Cependant, les patrons sont appliqués la plupart du temps à extraire des objets pédagogiques de type "Définition" en raison de l'accessibilité des structures langagières relatives à ce type que ce soit sur le web (wikipédia, dictionnaires, etc.) ou dans d'autres sources comme les rapports, les manuels d'utilisation, etc.

Dans cet article, nous proposons une annotation automatique des informations pédagogiques avec des métadonnées sémantiques (Définition, Exemple, Exercice, etc.). Ce qui nous permettrait d'indexer ces informations en vue d'une extraction des informations pertinentes par rapport à une requête utilisateur.

3 Notion d'objets pédagogiques

Un utilisateur "extracteur" d'informations pédagogiques pertinentes est guidé dans sa lecture par certains passages, des segments textuels (phrases ou de paragraphes). L'hypothèse générale utilisée ici est d'essayer de reproduire ce que fait un humain, en particulier l'apprenant, en soulignant certains segments textuels reflétant un contenu pédagogique. Ces segments de type pédagogique, appelés objets pédagogiques, existent, généralement, dans les documents pédagogiques sous forme de définitions, exemples, exercices, plan, questions et réponses, etc. Un objet pédagogique peut être défini comme une entité numérique ou non (Flory, 2004) qui peut être utilisée ou citée dans un apprentissage. Dans notre cas, un objet pédagogique correspond à un segment textuel reflétant un contenu pédagogique.

Un apprenant pourrait être intéressé par une définition en formulant une requête, par exemple: trouver les documents qui contiennent "La définition du langage SQL". Un autre utilisateur recherche, en explorant de nombreux textes (encyclopédies spécialisées, manuels, articles), des exemples sur un concept (par exemple, «l'inflation» dans l'économie, «polysémique» en linguistique, ..) pour l'intégrer à ses ressources pédagogiques. Un autre utilisateur peut être intéressé, à la pratique des exercices sur un concept. L'objectif de ces types d'objets pédagogiques (Définition, Exemple, Exercice) est une annotation possible des segments textuels pédagogiques qui correspondent à une recherche guidée afin d'en extraire des objets pédagogiques à partir de textes.
Chaque type pédagogique, comme nous l'avons mentionné ci-dessus, est explicitement indiqué par les marqueurs linguistiques identifiables dans les textes. Notre hypothèse est que chaque type d'objet pédagogique laisse des traces discursives dans le document texte. Les types d'objets pédagogiques sont décrits comme suit:

- D'une part, une relation complexe entre les concepts dans une structure «carte sémantique» (Figure 1) et d'autre part un ensemble de classes et sous-classes d'unités linguistiques (indicateurs et indices).

- Un ensemble de règles communautaires où chaque règle concerne une classe d'indicateurs avec des indices différents.

La carte sémantique (Figure 1) est une organisation des types d'objets pédagogiques. Elle peut être conçue aussi comme une ontologie des types d'objets pédagogiques indépendamment des différents domaines d'application. En effet, les expressions de la carte sémantique pour un type d'objet sont les mêmes dans différents domaines comme l'informatique, mathématiques, gestion, ... car ces expressions sont utilisées par l'auteur pour exprimer une information pédagogique. Dans certains types de textes (textes narratifs, articles de presse,) les expressions pédagogiques ne sont pas présentes mais dans d'autres (support de cours, devoirs, travaux dirigés, ..), ces expressions organisent le texte et donnent des informations sur l'intention de l'auteur.

Le premier niveau de la carte sémantique (Figure 1) présente 6 types d'objets pédagogiques : (i) Cours, (ii) Plan, (iii) Exercice, (iv) Exemple, (v) Définition, (vi) Caractéristique. Par exemple, les règles du type d'objet "Définition" sont déclenchés par la présence de noms ou de verbes définitoires (par exemple: "*est défini*", et l'annotation sémantique est attribuée si des indices linguistiques, comme les prépositions (l'indice de l'exemple précédent est "*par*"), sont trouvés dans le contexte de l'indicateur.

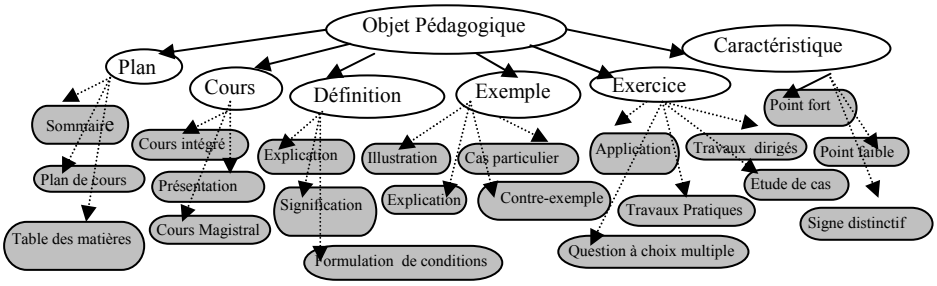

Figure 1 : Carte sémantique des types d'objets pédagogiques

4 Approche proposée pour la recherche d'informations pédagogiques à partir de documents

L'approche que nous proposons se décompose en deux principales parties: dans la première partie, nous procédons à une annotation sémantique des segments textuels représentant des objets pédagogiques (Smine et al., 2010). La deuxième partie exploite les annotations générées par la première partie pour créer un index qui est capable de localiser les segments textuels pertinents par rapport à des requêtes associées aux types pédagogiques (Définition, Exemple, Exercice, etc.). Pour classer les réponses selon leurs pertinences, nous appliquons l'algorithme de classification Rocchio sur les objets pédagogiques extraits.

4.1 Annotation des objets pédagogiques

4.1.1 Segmentation

La segmentation est la détermination des limites des unités linguistiques (unités comme proposition, phrase, paragraphe, etc.). La segmentation des textes en petites unités (phrases) reste encore une tâche difficile à réaliser, vu qu'un point suivi d'une majuscule ne peut pas déterminer le début ou la fin d'un segment. Il est nécessaire de prendre en compte tous les marqueurs typographiques. Il existe des travaux qui considèrent l'aspect multilangue dans leur segmentation comme le travail de (Mourad, 2002) qui propose de définir un segment textuel en se basant sur une étude systématique des marques de ponctuation. Nous avons effectué la segmentation de nos documents en intégrant les règles linguistiques développées par Mourad. Pour chaque document segmenté, le résultat obtenu est un fichier XML balisé par des balises <Section>, <Paragraphe>, <Phrase>.

4.1.2 Annotation des objets pédagogiques

Pour annoter les objets, nous explorons la technique d'Exploration Contextuelle 'EC' (Desclés, 1997). C'est une technique de traitement linguistique et sémantique du langage, qui fait appel à des marqueurs discursifs explicites (morphèmes, mot, expression, etc.) caractéristiques d'une intention pragmatique de l'auteur. 'EC' consiste à appliquer des règles dans un contexte déterminé par des indices. Elle a l'avantage d'être indépendante d'un domaine particulier, car les règles décrivant les structures linguistiques sont indépendantes d'un domaine particulier. C'est une méthode qui a été validée par les travaux de (Djioua et al., 2006) et (Elkhlifi et al., 2010). En plus, 'EC' ne nécessite pas une analyse morphosyntaxique du texte, ce qui réduit considérablement le temps d'exécution pendant l'implémentation de la méthode.

Par l'exploration contextuelle du contenu des documents, nous pouvons repérer et annoter les objets pédagogiques contenu dans ces documents, par exemple, « des exemples de requêtes SQL», « des exercices sur le langage UML », « les définitions d'une ou de plusieurs notions », etc. Ces objets sont exprimés par des structures langagières comme « …se définit par… », « est défini par… » pour le type *Définition* ou « Exercices sur… », « Travaux dirigés » pour le type *Exercice*. Ils sont explicitement indiqués par des indicateurs linguistiques identifiables dans les textes (verbes, noms, adjectifs). Ces indicateurs sont parfois polysémiques, ils ont besoin d'indices linguistiques pour clarifier l'indétermination. Les relations reliant les

indicateurs aux indices sont définis dans le cadre des règles. Une règle (IdR) se déclenche au moment de l'identification de l'un de ses indicateurs (Indicateur) ensuite elle essaye de localiser des indices linguistiques dans le contexte gauche (CL_1, CL_2) et/ou droite (CR_1, CR_2) de l'indicateur ce qui confirme ou non la valeur sémantique exprimée par l'indicateur. A chaque type d'objet pédagogique correspond un ensemble de règles. Des exemples de règles sont présentés dans le tableau suivant (Tableau 1).

IdR	CL_1	CL_2	Indicateur	CR_1	CR_2	Type|Sous-type de l'objet pédagogique
RD1	est|sont		défini|définie| définis	par		Définition|Explication
RD2			est | sont	le|la|un|une |des|les		Définition|Explication
RC1	La|Les Des|Une		caractéristique| caractéristiques	du|de|des	est| sont	Caractéristique|Signes distinctifs
RE1	Voici	un|l'|les |des	exemple| exemples	du|de|des		Exemple|Illustration

Tableau 1 : Des exemples de règles

Nous avons ajouté un composant à chaque règle qui représente l'emplacement du terme de la requête à rechercher dans le cadre du segment exprimant l'objet pédagogique. Le besoin d'ajouter ce composant est né de la variation de l'emplacement du terme à rechercher avec la variation des structures langagières exprimant les objets pédagogiques. Ceci permet d'identifier les segments textuels exprimant le type d'objet pédagogique ainsi que le concept demandés par l'utilisateur. Par exemple, pour le même type d'objet pédagogique "Définition" : le terme à rechercher "*Maintenance*" peut exister au début du segment "*La maintenance est définie comme l'ensemble des activités destinés à maintenir ou à rétablir un bien dans un état de sûreté de fonctionnement*" ou au milieu du segment pour le cas "*L'AFNOR a défini la maintenance comme étant l'ensemble des activités de remise en état de fonctionnement d'un système*". Sans la considération de ce paramètre, le système peut ne pas extraire l'objet demandé par l'utilisateur comme par exemple, pour le type *Cours*, la plupart de ses règles d'EC exigent un emplacement du terme de la requête au niveau du Titre du document. Au cas où le terme est recherché ailleurs du titre, le résultat de la recherche sera erroné.

De ce fait, l'emplacement du terme est un paramètre qui diffère d'une règle à une autre selon la structure langagière exprimée par cette dernière. Nous avons désigné cet emplacement par une étiquette, qui prendra une valeur parmi un ensemble fini de valeurs désignant l'emplacement du terme par rapport aux indicateurs et indices. Par exemple, GIND indique le terme se place à gauche de l'indicateur ou TITRE indique que l'emplacement du terme est au niveau du titre du document. En fait, dans plusieurs cas, le titre peut nous révéler des connaissances sur le contenu du document.

Pour chaque type d'objet de la carte sémantique (cf. Figure 1), nous avons défini un ensemble de règles qui couvrent toutes les formes linguistiques possibles de l'objet pédagogique. Nous avons commencé par un exemple textuel relatif à chaque type pour généraliser toutes les structures langagières. Cette méthode permet de définir de manière incrémentale une base solide de règles. Nous avons développé en totalité environ 200 règles. L'ensemble des règles développées, ainsi que la carte sémantique forment les ressources linguistiques utilisées dans notre approche.

Nous prenons un extrait de texte à partir d'un document pédagogique

Chapitre 1

Présentation de SQL

SQL est un langage complet de gestion de bases de données relationnelles. Il n'est pas un langage conceptuel. Il a été conçu, dans les années 70, par IBM. Il est devenu le langage standard des systèmes de gestion de base de données (SGBD).

Pour le type d'objet pédagogique "*Définition*", la règle RD2 (cf. Tableau 1), appliquée à l'exemple ci-dessus, permet d'annoter la phrase " *SQL est un langage complet de gestion de bases de données relationnelles*". Le type d'objet pédagogique est détecté grâce à l'expression "*est* " qui est une occurrence Ii de l'indicateur du type "*Définition*" et l'indice droit CR1 "*un*".

Pour le type "*Cours*", le repérage de l'occurrence Ii au niveau du titre est suffisant pour annoter le document comme un cours. L'indicateur nominal de l'objet pédagogique est le mot "Cours", et d'autres noms comme "Chapitre, "Notes de cours". A part le titre, l'existence de l'indicateur "Cours" n'implique pas l'annotation du document comme un cours.

Notons que la phrase "*Il n'est pas un langage conceptuel*" illustre le cas des indices négatifs. En effet, la présence de l'expression "n'….pas" annule l'annotation du segment comme Définition, malgré la présence de l'indicateur "*est*" et l'indice droit CR1 "*un*".

Afin d'annoter le segment " *Il a été conçu, dans les années 70, par IBM*" comme une "*Caractéristique*", nous détectons en premier lieu l'expression "*a été conçu*" ensuite nous cherchons, dans le contexte droit de l'indicateur, le CR1 "*par*". En cas où les deux éléments (Ii et CR1) sont présents, alors le système annote le segment comme une caractéristique.

Concernant le type d'objet "Exercice", l'indicateur peut être verbal (a) ou nominal (b), par exemple :

- (a) "*Formulez une clause SQL……*" a comme indicateur verbal "*Formulez*"
- (b) "*Exercices sur requêtes SQL*", son indicateur est le nom "*Exercices*"

4.2 Génération de l'index

Notre objectif, par l'annotation, est de générer un index sémantique contenant à la fois des objets pédagogiques annotés selon leur type, en utilisant la méthode d'annotation détaillée ci-dessus, et l'emplacement du terme de la requête spécifié par la règle appliquée pour annoter l'objet. Cet index servira à extraire les objets répondant à la requête utilisateur. Les métadonnées générées par les annotations des différents objets sont stockés dans une base de données. Pour chaque objet pédagogique annoté, les métadonnées suivantes sont introduites dans l'index : (1) L'objet pédagogique annoté (OBJECT), (2) Chemin du document analysé (PATH), (3) Type de l'objet annoté (TYPE), (4) Identifiant de la règle appliquée pour annoter le segment (IDRule) et (5) L'emplacement du terme de la requête (TERMEMP). La figure suivante (Figure 3) montre deux exemples d'objets annotés.

EDIT	OBJECT	PATH	TYPE	IDRULE	TERMEMP
	La production est une transformation de ressources appartenant à un système productif et conduisant à la création de biens ou de services.	C:\Documents and Settings\Boutheina SMINE\Mes documents\Evaluation\Gestion de Production.txt	Définition:Explication	RD2	GIND
	2) Exprimer en algèbre relationnel les requêtes suivantes et donner ses résultats : checkbld Nom des immeubles ayant plus de 10 étages. checkbld Qui habite le « Koudalou » ? checkbld Nom et Profession des personnes ayant emménagé avant 1994. checkbld Gérant des immeubles ayant un appartement de plus de 150 m². checkbld Dans quel immeuble habite un acteur ? checkbld Age et profession des occupants de l'immeuble géré par « Ross » ? checkbld Qui n'habite pas un appartement géré par « Ross » ?	C:\Documents and Settings\Boutheina SMINE\Mes documents\Evaluation\Base de donnée.txt	Exercice:Travaux Dirigés	RE5	TITRE

Figure 3 : **Deux exemples d'objets annotés et indexés**

Afin de pouvoir extraire les objets pédagogiques qui contiennent des termes de la requête, nous avons utilisé la base de synonymes WOLF (qui représente la partie traduite en Français du dictionnaire WordNet) permettant d'enrichir la requête en prenant en compte tous les termes équivalents au terme de la requête. Ce

dernier est remplacé par la liste de ses synonymes. Ceci permet d'étendre le champ de la recherche. La requête est ainsi composée des termes à rechercher (par exemple "Langage SQL") et du type d'objets pédagogiques requis par l'utilisateur (par exemple : Exercice).

Grâce à un moteur de recherche (implémenté sous la plateforme *Lucene*), le système se connecte à l'index généré et retient les documents contenant des objets pédagogiques de même type que celui énoncé dans la requête (Exercice). Ensuite, le moteur procède à une recherche des termes de la requête (Langage SQL ainsi que ses synonymes) à partir des objets annotés et indexés. Cette recherche s'effectue dans l'emplacement désigné par la règle avec laquelle est annoté l'objet pédagogique. Par exemple, si l'emplacement du terme spécifié par la règle est DIND, le terme de la requête est recherché à droite de l'indicateur de la règle appliquée (Dans ce cas règle de type Exercice). Dans le cas où la requête est composée du type pédagogique "Exercice" et le terme «Langage SQL", le moteur de recherche procède comme suit:

- il extrait tous les objets pédagogiques trouvés dans l'index associé à l'annotation «Exercice»

- Pour chaque objet extrait, il recherche le terme "langage SQL» et ses synonymes dans l'emplacement spécifié par la règle d'annotation.

- Sélection, à partir des objets pédagogiques extraits, les objets comportant une occurrence du terme «langage SQL» ou ses synonymes dans le bon emplacement.

- Afficher toutes les informations présentes dans l'index associé à chaque objet pédagogique sélectionné.

4.3 Classement des objets pédagogiques

Après l'extraction des objets pédagogiques répondant à la requête utilisateur, une autre étape suit pour classer les réponses dans un ordre croissant selon leur similarité avec la requête. Pour classer ces objets, nous avons utilisé l'algorithme de Rocchio (Rocchio, 1971), adapté à la classification des textes (Ittner et al., 1995). L'utilisateur choisit un concept pour le correspondre au terme de sa requête, parmi une liste de 15 concepts appartenant à différents domaines (domaine de l'informatique, économie, génie mécanique, biologie, etc.). Ce sont des concepts auxquels appartient l'ensemble des documents du corpus d'annotation et d'indexation. Le concept choisi représente la classe C_{user} par rapport à laquelle les objets seront classés selon leur pertinence. Rappelons que nous considérons un objet pédagogique comme un segment textuel ayant différentes tailles (Phrase, paragraphe, document, etc.) selon le type de l'objet.

Nous représentons les données (les objets d'apprentissage et de test) par des vecteurs de poids numériques. Le vecteur de poids pour le *m* ième objet pédagogique est $V^m = (p_1^m, p_2^m, \ldots\ldots, p_l^m)$, où *l* est le nombre de termes index utilisés. Nous utilisons comme termes des mots singuliers et composés. Nous adoptons la mesure de poids TF-IDF (Salton, 1991) et nous définissons le poids p_k^m comme suit :

$$p_k^m = \frac{f_k^m \log(N/n_k)}{\sum_{j=1}^{l} f_j^m \log(N/n_j)}$$

Avec *N* est le nombre total d'objets, n_k est le nombre d'objets dans lesquels le terme index *k* apparaît, et f_k^m est :

$$f_k^m = \begin{cases} 0 & q = 0 \\ \log(q) + 1 & Sinon \end{cases}$$

Avec *q* est le nombre d'occurrences du terme index *k* dans l'objet *m*. Dans l'algorithme de Rocchio, un prototype est produit pour chaque classe *C*. Ce prototype est représenté par un vecteur singulier $\vec{c}_j$ de même dimension que le vecteur de poids original $v^1, \ldots\ldots, v^N$. Pour chaque classe *C*, the *k ième* terme dans son prototype est défini comme

$$\vec{c}_j = \frac{\alpha}{|C_j|} \sum_{m \in C_j} p_k^m - \frac{\beta}{|N - C_j|} \sum_{m \notin C_j} p_k^m$$

Avec *Cj* est l'ensemble de documents appartenant à la classe *C*. Les paramètres α et β contrôlent la contribution des exemples positifs et négatifs par rapport au vecteur prototype. Nous utilisons les valeurs standards α =4 et β = 16 (Buckley et al., 1994).

Une fois l'apprentissage achevé, nous classons les nouveaux objets fournis comme réponses à la requête utilisateur. Ce classement se fait selon leur pertinence par rapport à la classe C_{user} choisie par l'utilisateur. Les objets à classer sont tout d'abord convertis en vecteurs de poids, et puis comparés aux vecteurs de poids prototypes des différentes classes en utilisant la mesure de similarité cosinus.

La mesure de similarité entre l'objet de vecteur $\vec{O}$ et la classe C_{user} de vecteur $\vec{C}_{user}$ est définie comme :

$$\cos(\vec{C}_{user}, \vec{O}) = \frac{\vec{C}_{user} * \vec{O}}{\left|\vec{C}_{userl}\right|\left|\vec{O}\right|}$$

Les objets ayant une valeur de similarité avec la classe C_{user} supérieure à un seuil θ sont sélectionnés, ensuite classés dans un ordre croissant selon la valeur de leurs similarités par rapport à la classe C_{user}. La valeur du seuil θ varie selon le type d'objet pédagogique. Par exemple, un objet annoté par le type "*Cours*" contient plus de termes significatifs qu'un objet annoté par le type "*Exercice*" ($\theta_{Course} < \theta_{Exercice}$). Nous ne prenons en compte que les valeurs positives de la mesure de similarité. Les objets sélectionnés sont alors affichés pour constituer la fiche pédagogique demandée par l'utilisateur. Une fiche pédagogique rassemble les objets pédagogiques de type celui exprimée par l'utilisateur dans sa requête et correspondant au même concept que celui recherché par l'utilisateur. Cette fiche permet une accessibilité aux objets directement sans avoir accès au document en entier.

5 Expérimentations et Résultats

L'objectif de cette étape est d'évaluer les performances des différents modules. Un des indicateurs importants est donc le nombre des réponses pertinentes par rapport au nombre de documents indexés. Pour valider notre approche d'indexation d'objets pédagogiques, nous avons développé le système SRIDoP (**S**ystème de **R**echerche d'**I**nformations à partir de **Do**cuments **P**édagogiques) en utilisant le langage Java sous l'environnement Eclipse et le système de gestion de base de données Oracle. SRIDoP comporte les trois modules suivants : Annotation et indexation des objets pédagogiques selon leurs types, Appariement entre la requête utilisateur et les objets pédagogiques indexés, et Classement des objets pédagogiques.

Notre corpus d'apprentissage ainsi que celui du test est le même pour toutes les étapes d'annotation, d'indexation et de classification. Pour le corpus d'apprentissage, nous avons collecté un ensemble de documents couvrant 15 concepts (ceux utilisés dans la génération de fiches pédagogiques). En fait, pour chacun de ces concepts, une requête a était formulée et exécutée sur le moteur de recherche Google. Les 20 premiers résultats sont collectés. Notons que le sens de quelques termes peut être ambigu, par exemple "Base" ou "Enregistrement". Pour désambiguïser la requête, nous ajoutons le terme "Données". Pour faire disparaitre l'ambiguïté, nous misons sur le type pédagogique des documents retournés en réponse. Les documents collectés sont constitués de 60 supports de cours, 65 Travaux Dirigés, 83 Présentation PowwerPoint, 30 Travaux Pratiques, et quelques documents de différentes natures (articles de Presse, articles scientifiques, etc.). La longueur moyenne de ces documents constituant le corpus d'apprentissage est 23 pages.

Notre corpus de test est composé de 1000 documents, principalement de nature pédagogique : des Supports de cours, des Travaux Dirigés, des présentations PowerPoint, des Travaux Pratiques, des manuels d'utilisation, et d'autres documents de différentes natures. La longueur moyenne des documents est 53.6 pages. Les documents ont différents formats (DOC, PDF, HTML, PPT, etc.).

5.1 Première étape : Annotation des objets pédagogiques

Pour évaluer le processus d'annotation, le corpus de test a été annoté par deux experts : pour chaque objet pédagogique repéré, ils précisent son type. Les résultats du processus d'annotation effectué par notre système SRIDoP sont illustrés dans le Tableau 2.

Type de l'objet pédagogique	NOA	NOAC	NOMAC	Précision (%)	Rappel (%)	F-Mesure (%)
Plan	88	85	98	96,59	86,73	91,40
Cours	72	60	85	83,33	70,59	76,43
Définition	228	140	266	61,40	52,63	56,68
Caractéristique	139	124	156	89,21	79,49	84,07
Exemple	357	349	376	97,76	92,82	95,23
Exercice	760	705	776	92,76	90,85	91,80

Tableau 2 : Les résultats de l'étape Annotation

$$\text{Précision} = \frac{NOAC}{NOA} \qquad Rappel = \frac{NOAC}{NOMAC} \qquad F - Mesure = 2 * \frac{\text{Précision} * Rappel}{\text{Précision} + Rappel}$$

Avec : **NOA** : Nombre d'objets annotés, **NOAC** : Nombre d'objets annotés correctement, **NOMAC**: Nombre d'objets annotés par les experts.

Nous remarquons que la précision de l'annotation dépasse les 85% pour la plupart des types d'objets (*Exemple, Exercice, Plan, etc.*). Notons que, pour le type « Définition », cette précision est moyenne. Ceci dérive du fait que certaines règles peuvent annoter à la fois des énoncés définitoires ou non. Tel le cas de la règle « R2 » ayant comme indicateur l'occurrence «**est un**». Cet indicateur peut identifier un segment de nature définitoire (exemple : « *UML **est un** langage de modélisation conceptuelle orienté objet* ») ou un autre segment de nature non définitoire (exemple : « *Le facteur temps **est un** des plus importants dans la réalisation d'un projet* »). Pendant la phase d'expérimentation, nous avons pu constater aussi que la qualité de l'annotation est étroitement liée à la qualité de la segmentation du document.

5.2 Deuxième étape : Indexation des objets pédagogiques

A travers une interface de recherche d'informations, l'utilisateur saisit les termes à rechercher, et choisit le type (et sous-type) de l'objet pédagogique relatif au terme à rechercher. Les réponses aux requêtes sont affichées sous forme de liens permettant d'accéder à l'objet pédagogique répondant au besoin de l'utilisateur. Pour tester ce module de recherche d'objets pédagogiques, nous avons formulé les mêmes 25 requêtes pour chacun des types d'objets pédagogiques. Ces requêtes appartiennent aux différents domaines du corpus. Pour chaque type d'objet, nous avons illustré le nombre de réponses ramenées et le nombre de réponses jugées pertinentes compte tenu de l'ensemble des requêtes formulées. Les résultats sont résumés dans le tableau suivant (Tableau 3).

Type de l'objet pédagogique exprimé par la requête	NR	NRP	NRRU	Précision (%)	Rappel (%)	F-Mesure (%)
Plan	72	66	77	91,67	85,71	88,59
Cours	43	35	54	81,40	64,81	72,16
Définition	156	112	193	71,79	58,03	64,18
Caracteristique	94	86	112	91,49	76,79	83,50
Exemple	213	198	230	92,96	86,09	89,39
Exercice	517	465	520	89,94	89,42	89,68

Tableau 3 : Les résultats de l'étape d'appariement Documents-Requête

$$\text{Précision} = \frac{NRP}{NR} \qquad Rappel = \frac{NRP}{NRRU} \qquad F - Mesure = 2 * \frac{\text{Précision} * Rappel}{\text{Précision} + Rappel}$$

Avec : **NR** : Nombre d'objets (réponses) retournés à l'utilisateur, **NRP** : Nombre d'objets (réponses) pertinents retournées à l'utilisateur, **NRRU**: Nombre d'objets pertinents.

A l'issue de ces expérimentations, nous remarquons que les résultats de l'indexation d'informations pédagogiques sont étroitement liés aux résultats de l'annotation (cf. Figure 4). La valeur de "F-Mesure" de l'extraction évolue avec la valeur de "F-Mesure" de l'annotation. Ceci s'explique par le fait, que l'extraction est effectuée à partir d'objets pédagogiques annotés et indexés. La qualité de la recherche s'améliore en améliorant celle de l'annotation. Cette dernière est elle-même dépendante de la qualité de segmentation comme nous l'avons déjà mentionné.

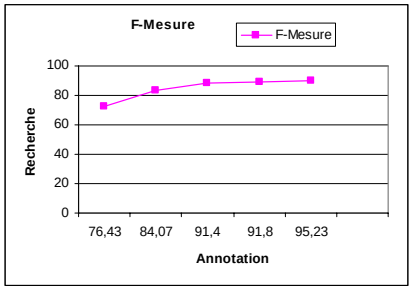

Figure 4 : Evolution des résultats de la recherche par rapport à celles de l'annotation

5.3 Troisième étape : Classement des objets pertinents

Après une extraction des objets pédagogiques, nous classons ces objets selon leur similarité avec la classe C_{user}. Suite à plusieurs expérimentations, nous avons fixé la valeur du seuil θ :

- 0.1 pour les types "Cours" et "Définition",
- 0.3 pour les types "Plan" et "Exemple",
- 0.45 pour les types "Caractéristique" et "Exercice".

Notons que d'un côté, diminuer la valeur de θ réduit l'ensemble des objets pertinents retournés à l'utilisateur. D'un autre côté, augmenter la valeur de θ amène à une sélection des objets non pertinents.

Nous avons assigné chaque objet à l'une de ces trois catégories : **A** (objets classés comme pertinents), **B** (objets classés correctement comme pertinents), **C** (objets pertinents). Les valeurs de précision, de rappel et de F-Mesure sont calculées pour chaque type d'objet pédagogique comme suit :

$$\text{Pr}\acute{e}cision = \frac{B}{A} \qquad Rappel = \frac{B}{C} \qquad F - Mesure = 2 * \frac{\text{Pr}\acute{e}cision * Rappel}{\text{Pr}\acute{e}cision + Rappel}$$

Nous illustrons ces valeurs relatives à chacun des types d'objets dans la Figure 5.

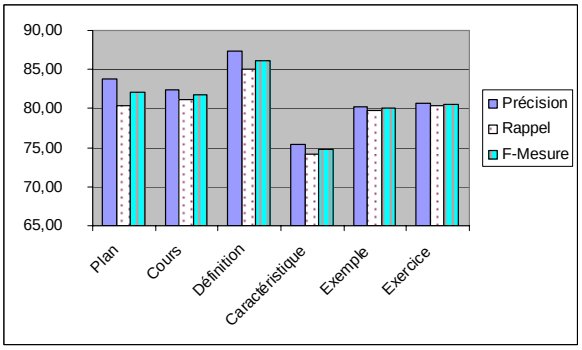

Figure 5 : Précision, Rappel et F-Mesure de l'étape de classement des objets

La figure ci-dessus présente, pour chaque type d'objet (représenté sur l'axe des abscisses), sa valeur de précision représentée en bleu, sa valeur de rappel en pointillé et sa valeur de F-Mesure représentée en rayures. Nous constatons que les valeurs de précision sont comprises entre 75% et 87% et que celles du rappel entre 74% et 85%. Notons que l'étape de classement ne dépend pas strictement de celles de l'annotation et d'appariement mais plutôt d'autres paramètres comme le corpus d'apprentissage, le choix des termes index, etc.

6 Conclusion et Perspectives

Dans cet article, nous avons proposé une approche d'indexation d'objets pédagogiques basée sur une annotation sémantique du texte par exploration contextuelle en vue d'une extraction des objets pédagogiques pertinents. Actuellement, notre travail présente un intérêt important dans plusieurs domaines d'application comme l'apprentissage en ligne, l'enseignement à distance (e-learning), l'éducation, etc. Pour évaluer notre approche, nous avons développé le système SRIDoP qui comprend les modules d'annotation, d'indexation, et de classement des objets selon leur pertinence. Nous remarquons, à travers les résultats d'évaluation, que notre approche permet d'avoir accès aux connaissances qui sont exprimées dans les textes selon un type donné, et de ramener des énoncés qu'un système de recherche d'informations classique ne parvient à capter par son approche d'indexation par mots clés.

L'un des travaux futurs que nous envisageons est l'extension de la carte sémantique des types d'objets pédagogiques par d'autres types comme Méthode, Auteur, Date, etc. Nous pensons aussi à la proposition d'une fonction score qui fusionne les résultats des deux modules d'annotation et de classement en vue de sélectionner les résultats pertinents.

Bibliographie

BUCKLEY C., SALTON G., ALLAN J. (1994). The effect of adding relevance information in a relevance feedback environment. Actes de *International ACM SIGIR Conference*, 292-300.

BUFFA M., DEHORS S., FARON-ZUCKER C., SANDER P. (2005). Vers une approche Web Sémantique dans la conception d'un système d'apprentissage. Revue du projet TRIAL SOLUTION, *AFIA*.

DEHORS S., FARON-ZUCKER C., STROMBONI J.P., GIBOIN A. (2005). Des annotations Sémantiques pour apprendre : l'Expérimentation QBLS. *WebLearn*.

DESCLES J.P. (1997). Système d'exploration Contextuelle. *Co-texte et calcul du sens*, Caen, 215-232.

DJIOUA B., FLORES, J.G, BLAIS A., DESCLES J.P., GUIBERT G., JACKIEWIEZ A., LE PRIOL F., NAIT BAHA L., SAUZAY B. (2006) Excom: an automatic annotation engine for semantic information. Dans *Proc. FLAIRS*, AAAI Press, Florida, 285-290.

ELKHLIFI A., FAIZ R. (2009). Automatic Annotation Approach of Events in News Articles. *International Journal of Computing & Information Sciences*, 19-28.

Elkhlifi A., FAIZ, R. (2010). French-Written Event Extraction Based on Contextual Exploration. Dans *Proc. FLAIRS*, AAAI Press, Florida.

FLORY L. (2004). Les caractéristiques d'une ressource pédagogique et les besoins d'indexation qui en résultent. *Journée d'étude sur l'Indexation des ressources pédagogiques numériques*, Ennsib, Villeurbanne.

GREENWOOD M.A., SAGGION H. (2004). A Pattern Based Approach to Answering Factoid, List and Definition Questions. Dans *Proc. RIAO 2004*, Avignon, France.

HASSAN S., MIHALCEA R. (2009). Learning to identify educational materials. Dans *Proc. RANLP*, Bulgaria.

ITTNER D.J., LEWIS D.D., AHN D. D. (1995). Text categorization of low quality images. Actes de *SDAIR*, Las Vegas, US, 301-315.

MOURAD G. (2002). La segmentation de textes par Explration Contextuelle automatiques, présentation du module SegATex. Dans *Inscription Spatiale du Langage : structure et processus ISLsp*, Toulouse.

ROCCHIO J. (1971). Relevance feedback information retrieval. *In Gerard Salton editor, The Smart retrieval system experiments in automatic document processing*, Prentice-Hall, Englewood Cliffs, NJ, 313-323.

SALTON G. (1991). Developments in automatic text retrieval. *Science*, 253 (5023), 974-980.

SMEI H., BEN HAMADOU A. (2005). Un système à base de métadonnées pour la création d'un cache communautaire-Cas de la communauté pédagogique. Dans *Proc. IEBC*, Hammamet, Tunisie.

SMINE B., FAIZ R., DESCLES J.P. (2010). Analyse de documents pédagogiques en vue de leur annotation. *Revue des Nouvelles Technologies de l'Information (RNTI)*, E-19, Ed. Cépaduès, 429-434.

THOMPSON C., SMARR J., NGUYEN H., MANNING C. (2003). Finding educational resources on the web : Exploiting automatic extraction of metadata. *Proc. ECML, Workshop on Adaptive Text Extraction and Mining*.

WESTERHOUT E., MONACHESI P. (2008). Creating glossaries using pattern-based and machine learning techniques. Dans *Proceedings of Map of Language Resources, Technologies and Evaluation*.