The SRIDoP System Using Semantic Metadata for Web Database Processing

Boutheina Smine[1,3], Rim Faiz[2], Jean-Pierre Desclés[3]
[1]LARODEC
High Institute of Management of Tunis
Le Bardo
Tunisie
[2]LARODEC
IHEC de Carthage
2016 Carthage Présidence
Tunisie
[3]LaLIC
Paris Sorbonne University
28 Rue Serpente
Paris 75006
France
Boutheina.Smine@etudiants.univ-paris4.fr, Rim.Faiz@ihec.rnu.tn, Jean-Pierre.Descles@paris4.sorbonne.fr

***ABSTRACT** : Searching learning information from the web or from databases is a user's need to learn or to teach. In order to satisfy these user's needs, we proposed here a model which aims at automatically feeding texts with semantic metadata. These metadata would allow us to search and extract learning information from texts indexed in that way. This model is build up from two parts: the first part consists on a semantic annotation of learning objects according to their semantic categories (definition, example, exercise, etc.). The second part uses automatic semantic annotation which is generated by the first part to create a semantic inverted index which is able to find relevant learning objects for queries associated with semantic categories. To sort the results according to their relevance, we apply the Rocchio's classification technique on the learning objects. We have implemented a system called SRIDoP, on the basis of the proposed model and we have verified its effectiveness.*

Keywords: Semantic annotation, Learning information, Document Classification, Contextual Exploration

Received: 3 March 2011, Revised 14 April 2011, Accepted 21 April 2011

1. Introduction

With the rapid growth of the information amount available online and in databases, search engines play an important role within eLearning, since they can support the learner in looking for the needed information for his learning, training or teaching process. However, these information extraction systems are based on terms indexation without taking into account neither the semantics of learning information contents nor the context.

A better alternative is to realize an information retrieval system based on the semantic annotation of learning objects which are attested in the documents (*Definition, Exercise, Example, etc*.). By doing so, the learning objects presented by the author of a

certain document are captured and the learning or the teaching process for the student or the instructor respectively is facilitated. We propose in this paper a learning objects retrieval system based on a semantic annotation process with Contextual Exploration and on a learning objects indexation. To improve the results obtained with these two processes, a machine learning technique is applied to sort the results according to their relevance.

The rest of this paper is organized as follows: Section 2 deals with the presentation of related works on learning information processing. In section 3, we present the semantic learning categories for text mining. Our model for learning objects retrieval is detailed in section 4. Before concluding, we illustrate the evaluation results of the different parts of our model in the fifth section.

2. Related works

Several works provide infrastructure and services for learning information annotation, indexing, and retrieval from documents. Among these works, we can mention:

The QBLS system (Question Based Learning System) (Dehors & Faron-Zucker, 2006) which aims at firstly structuring the course referring to a learning ontology constituted of cards (definition, example, procedure, solution, etc) and secondly abstracting learning resources (course, topic, concept, and question). is a part of the TRIAL SOLUTION plateform (Dehors et al., 2006) where the users specify the learning role of the resources content (definition, theorem, explanation, etc), the "key words" and the "relations" with other resources.

We denote also the SYFAX system (Smei & Ben Hamadou) where the authors proposed several metadata relative to the document indicating the correspondence of the document with the user profile, the user point of view on the document, the documents type (TD, TP, etc),etc. The authors propose a refinement process of the request based on the ontology of educational material types and on the ontology of the computer science domain.

In order to index pedagogical documents, the systems presented above stored the generated annotations in knowledge databases from which response to user's queries are extracted.

For all the systems presented above, the problem of learning documents annotation is discussed from various sides: (1) the course is structured manually according to a pedagogical ontology in order to use it in an e-learning environment, (2) the course is semi-annotated by users to produce personalized course supports. In all cases, a human intervention is provided to enrich documents with metadata. Therefore, many producers of learning content are not interested in going back and annotating all their work.

There exist other works of (Hassen & Mihalcea, 2009) and (Thompson et al., 2003) which target the problem of finding educational resources on the web. The focus of their work was limited to metadata extraction relative to the whole document. A set of properties (Relevance, Content Categories, course title, instructor, year, etc.) was explored to annotate and classify the educational resources. Therefore, their methods don't enable to reach the contents of the documents in order to analyse their textual elements.

In this paper, we propose a model which aims at automatically annotating learning objects according to their semantic categories (Definition, Example, Exercise, etc.) in order to index and extract learning objects as response to the user's query. Then, a machine learning technique is applied on the extracted objects to sort them according to their relevance.

3. Semantic learning information categories for text mining

Users looking for relevant learning information from documents are primarily students, learners and teachers. They proceed by guided readings which give preferential processing to certain textual segments having learning contents. The aim of this hypothesis is to reproduce "What makes naturally a learner who underlines certain segments relating to a particular learning category (Definition, Example, Exercise, etc.) which focus his attention". Indeed, such a learner could be interested by the definition as a learning category, by formulating a request such as: find textual segments which contain "Definition of an incremental process". Another user will search by exploring many texts (Course support, Assignments, etc.) examples on SQL language. Yet another user may be interested to practice exercises on a concept to integrate it to its resources. The aim of these information learning categories for text mining is at a focused reading and a possible annotation of the learning textual segments corresponding to a guided research in order to extract learning objects from texts. We considered this learning information as learning objects having several categories (see Figure 1) and which can be used or cited for learning, teaching, etc. Each learning object category is explicitly indicated by identifiable linguistic markers in the texts.

The learning information categories (Figure 1) are described as follows: (1) On the one hand, a complex relation between concepts inside a structured "semantic map" and on the other hand a set of classes and subclasses of linguistic units (indicators and indices). (2) A set of contextual exploration rules where each rule relates a class of indicators with different indices.

The semantic map is like an organization "in intention" of learning object categories, who's the classes of indices are extensional counterparts. The semantic map can be conceived also as ontology of learning object categories independently of different application domains. Indeed, the expressions of the semantic map for a learning object category are the same in different domains like Informatics, mathematics, management, … since these expressions are used by the author to express a learning information.

The first level of the semantic map makes it possible to release 6 learning object categories: (i) Course, (ii) Plan, (iii) Exercise, (iv) Example, (v) Definition, (vi) Characteristic. For instance, the Definition learning category rules are triggered by occurrence of definition nouns or verbs, and the semantic annotation is assigned if linguistic clues, like prepositions, are found in the indicator's context.

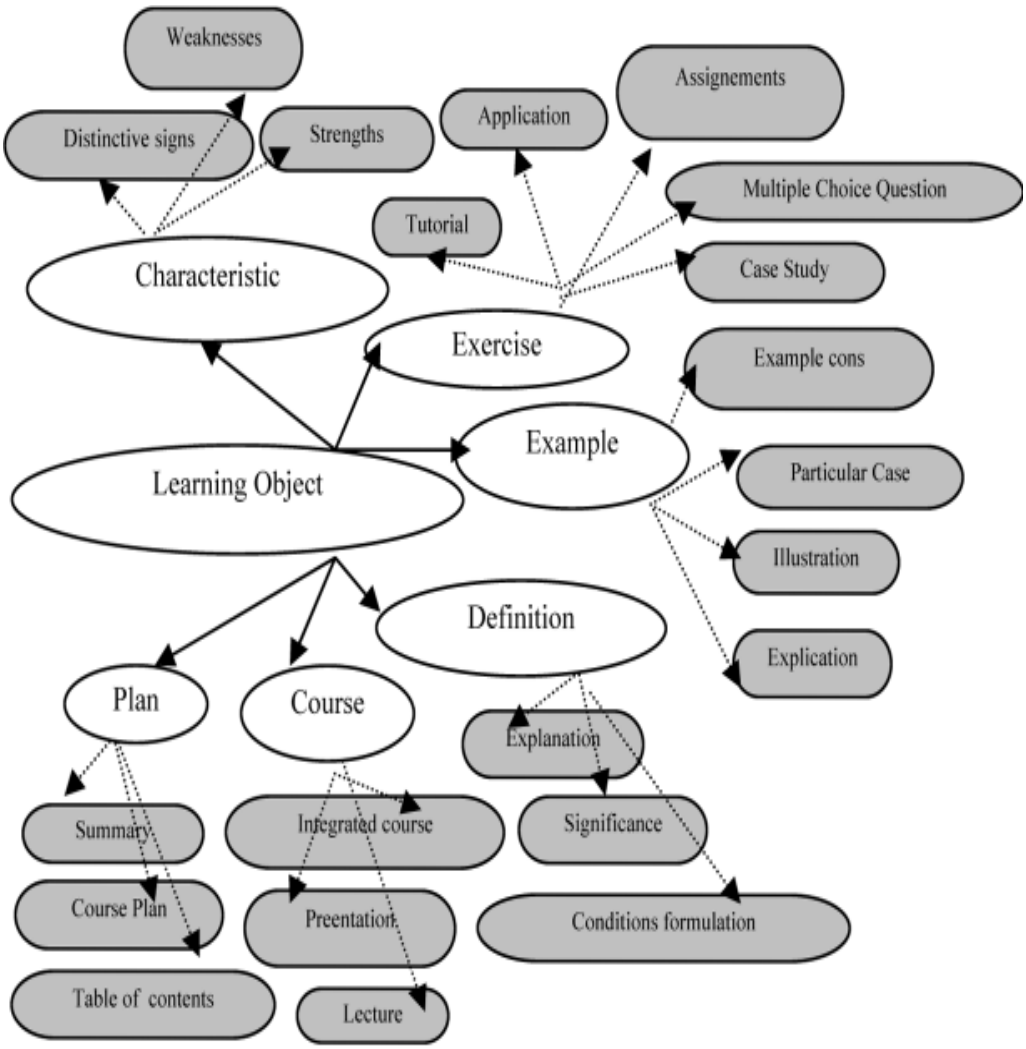

Figure 1. A learning objects semantic map

4. Our learning information retrieval model

Our model is built up from two parts: The first part consists of an automatic annotation of pedagogical texts according to learning

object categories (Smine et al., 2010, 2011). The second part uses automatic semantic annotation which is generated by the first part to create a semantic inverted index which is able to find relevant objects for queries associated with learning categories such as Definition, Exercise, Example, etc. Then, we propose to sort these objects according to their relevance using the Rocchio classification algorithm.

4.1 Learning Objects Annotation

4.1.1 Segmentation

Before applying the annotation task, the content of the considered document has to undergo a segmentation action which lies in determining the unit's borders (unit as sentence, paragraph, etc.). We have implemented our own segmentor based on the segmentation rules developed in (Mourad, 2002)] where defined a textual segment starting from a systematic study of the punctuation marks. Our plain text documents are then transformed into XML structured documents (titles, sentences, paragraphs, etc.).

4.1.2 Learning Objects Annotation Process

For the annotation process, we unfold the *Contextual Exploration technique* 'EC' (Desclés, 1997,2006) which call upon knowledge exclusively linguistic and present in the texts. This linguistic knowledge is structured in form of lists and is capitalized in a knowledge base. There are two kinds of lists: indicator lists on the one hand, contextual index (clue) lists on the other hand. Indicators are specific to a given information learning category (i.e.: to recognize a *Definition*, to locate an *Example*, etc.). Each indicator is seen as associating a set of heuristic rules of Contextual Exploration. The application of a rule called by an indicator, amounts seeking explicitly, in the indicator context, the linguistic indexes complementary to the indicator, in order to be able to solve the task. In addition, it doesn't need a morpho-syntactic analysis which reduces considerably the execution time of the method (Elkhlifi & Faiz, 2009), (Djioua et al., 2006), (Elkhlifi & Faiz, 2010).

We focus on the learning object categories (see the semantic map) to construct our contextual exploration rules. We go through each document in order to extract linguistic structures that define the learning object categories, i.e. the category "Definition" can be expressed by several structures : "...is defined as…", "The definition of ….is….", "To define …, we say that…". These linguistic structures are expressed by discursive markers (indicators and clues) which are represented in a list of verbs, prepositions, nouns, etc. Relations binding indicators and clues are defined within Contextual Exploration rules. The rule is triggered when one of its indicators is detected within the textual segments. These rules must identify an indicator (Ii) then locate linguistic clues to the left (CLi) and/or to the right (CRi) context of the indicator, which involves the confirmation or not of the semantic value carried by the indicator.

For each category of the semantic map, we defined the set of rules which covers all the possible linguistic form of learning object. We have developed about 180 rules. We start from a textual example to generalize all linguistic structures. This method permits to define incrementally a solid base of rules. Indeed, we give the permission to the user to manage the EC rule base (adding, updating, deleting rules) through the Access Database system. The Table 1 shows some examples of rules. In this table, IdR denotes the identifier of the rule; CL_1, CL2 denote the left clues and CR_1, CR_2 denote the right clues.

IdR	CL_1	CL_2	Indicator	CR_1	CR_2
RD1	is|are		defined	as	
RD2			is | are	a|an|the	
RC1	The| A		Characteristic| Characteristics	of	is|are
RE1	This is	an| the	example| examples	of	

Table 1. Examples of Contextual Exploration Rules

For example, the EC rule RD1 (see Table I) would follow these steps to annotate a textual segment as a *Definition*:

• Express the semantic of the "*Definition*" category by means of a relevant indicator, represented in this case by the verb "defined".

• To confirm the indicator's "definition semantic", we need first to identify in the sentence terms of the list CL1 (the verb is" or "are") in the left context

• Indicator needs another expression like the preposition "as" in the right context to allow the annotation of the sentence as a definition.

The whole rules relative to the various categories and their respective indicators and clues constitute the linguistic resources that we employed to annotate learning objects.

We take an extract from a pedagogical document (Figure 2)

Chapter 1

SQL Presentation

SQL is a complete language of Relational Database Management. It is not a conceptual language. It was conceived, in the 70th years, by IBM.

Figure 2. An extract from a learning document

When a rule of the learning category *Definition* is applied to the example above, it permits to annotate, as a definition, the sentence "SQL is a complete language of Relational Database Management". The Definition learning category is detected through the expression *"is"* which is an occurrence I_i belonging to the Definition indicator and the right clue CR1 "*a*".

For a rule of the learning category "*Course*", it is enough to find an occurrence Ii on the title level to annotate the document as a "*Course*". The nominal indicator of this rule is the word "*Course*" and other words like "*Chapter*", "*Course Notes*", etc. Beyond the title, the existence of a "Course" indicator does not imply an annotation of the document as a *Course*.

We noticed that the sentence "*It is not a conceptual language*" illustrates the case of negative clues (CRN, CLN). In fact, the presence of "*not*" prevent the annotation of the segment as a Definition although the presence of the indicator "*is*" and the clue "*a*".

With regard to the learning category "Exercise", the indicator can be verbal (*a*) or nominal (b), i.e.: (*a*) "*Formulate an SQL clause*" The indicator is the verb "*Formulate*", (b) "*Exercises on SQL Requests*" has as indicator the noun "*Exercises*"

We have introduced another parameter to the rule which is the emplacement of the term expressed by the user's query. This is due to the fact that the place of the term expressed in the query varies according to the rule applied to annotate the learning objects, i.e. for the category Definition, the term "*SQL Language*" can exist in the beginning of the sentence "*SQL Language is defined as the .....*", or in the middle of the sentence "*X has defined the SQL Language as ……*". We have designed this emplacement with a set of values, relatively to the indicator, and the clues of the rule (left or right of the indicator or the clues), i.e. ***LIND*** indicates that the emplacement is to the left of the indicator and ***RCL1*** indicates that the emplacement is to the right of the left clue.

4.2 Indexing Annotated Objects

The aim of this step is to build up a multiple index composed of learning objects (sentences, paragraphs …), semantic annotations (Definition, Example, Plan …). Each learning object is associated with several important pieces of information such as:

• a semantic annotation (Definition, Exercise, Plan, …) according to the semantic categories used in the annotation process

• document URI (Uniform Resource Identifier) for the identification of the document path

• the full-text content of the learning object for a relevant answer to users

• the emplacement of the term enounced in the user's query

We have implemented a learning information retrieval system, called SRIDoP (**S**ystème de **R**echerche d'**I**nformations à partir de Documents Pédagogiques), on the basis of the proposed model. SRIDoP uses a query language which is based at the same

time on both linguistic terms (constitutive of textual segments) and semantic learning categories (definition, example, exercise …). Let us see some queries for the "Exercise" category. The answer to the query "Exercises on SQL Language?", in French "Exercices sur le langage SQL?" gives a set of learning objects (textual segments) grouped through a document URI (the annotated document by the annotation process). Each learning object presents a semantic annotation ("Exercise" annotation for this example).
The search engine proceeds as follows:

• The query, in French, has two important functions: a learning object category ("Exercise") and the term "SQL Language".

• SRIDoP extracts all learning objects fund in the index associated with the annotation "Exercise"

• For each object extracted, SRIDoP searches the term "SQL Language" and its synonyms in the emplacement enounced in the index. For the term synonyms, we used a component of the synonyms dictionary WOLF (a French version of WordNet) to replace the query term by its synonyms. For example, if the term emplacement is RIND, the system looks for the term "SQL Language" in the right of the indicator.

• Selection from these learning objects, all objects within an occurrence of the term "SQL Language" or its synonyms in the well emplacement.

• Display all present information in the index related to each learning object selected.

4.3 Sorting the learning objects

We propose to sort the objects displayed by our system according to their relevance. So, we implemented a version of Rocchio algorithm (Rocchio, 1971), as adapted to text categorization by (Ittner et al., 1995). Our choice is justified by the fact that a learning object can be classified to more than one class. I.e. An object concerning the SQL Language can also concern the Data Base System and so on. We note that the Vector Salton Machine technique can satisfy this assumption by applying the Rocchio's algorithm.

First, the user has to correspond the terms of his request to a topic from a set of fifteen topics of different fields. The topic chosen represent the class C_{user} against which the objects will be sorted. We note that we consider a learning object as a textual segment having different sizes (sentence, paragraph, document, and so on).

The application of the Rocchio classifier can be divided into three steps: pre-processing, learning and sorting. The pre-processing includes objects formatting and terms extraction. We use single and compound words as terms.

The learning objects are extracted from the learning corpus collected within the annotation and indexation steps. In the learning step, we presented these objects as vectors of numeric weights. The weight vector for the mth object is $V^m = (p_1^m, p_2^m, \ldots\ldots, p_l^m)$, where l is the number of indexing terms used. We adopted the TF-IDF weighting (Salton,1991) and define the weight p_k^m to be :

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

Here, N is the number of objects, n_k is the number of objects in which the term index k appears, and f_k^m is:

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

Where q is the number of occurrences of the indexing term K in object *m*. We produced a prototype for each class *C*. This prototype is represented as a single vector $\vec{c}_j$ of the same dimension as the original weight vectors $v^1, \ldots, v^N$. For class C, the K'th term in its prototype is defined to be :

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

Where C_j is the set of all objects in class *C*. The parameters α and β control the relative contribution of the positive and negative examples to the prototypes vector, we use the standard values $\alpha = 4$ et $\beta = 16$ (Buckley et al., 1994).

When the learning step is achieved, we launched the sorting step and we measured the similarity between the objects given as response to the user's query and the class chosen by the user C_{user}. Each object is first converted into weight vector $\vec{o}$ using TF-IDF

weighting, and then compared against the class prototype $\vec{c}_{user}$ using the cosines measure:

$$\cos(\vec{c}_{user}, \vec{O}) = \frac{\vec{c}_{user} * \vec{O}}{|\vec{c}_{user}|\vec{O}|}$$

Objects having a similarity cosine measure lower then a threshold θ are selected and then sorted ascending against their similarity measure with the prototype $\vec{c}_{user}$. The θ value varies according to the learning objects category. i.e. the object content annotated as a *Course* contains more significant terms than an object of category *Exercise* ($\theta_{Course} < \theta_{Exercise}$). We take into account only positive values of the similarity measures.

5. Experimentation and results

We have implemented the SRIDoP system using the language Java and the Platform Lucene to annotate, index and sort the learning objects. To constitute the learning corpus for all the steps, we collect a data set covering the fifteen topics used in the step of "Creation of learning card-index"(i.e.Local Networks, Job-shop Scheduling, Programming language, Database, Maintenance, and so on.). Starting with each of these topics, a query is constructed and run against the Google search engine, and the top 20 ranked search results are collected. Note that the meaning of some terms can be ambiguous, e.g.., "Base" or "Record" and thus we explicitly disambiguate the query by adding the word "data". By performing this explicit disambiguation, we can focus on the learning property of the documents returned by the search, rather than on the differences that could arise from ambiguities of meaning.

The set of documents collected is constituted by 60 supports of course, 65 Assignments, 85 PowerPoint Presentations, 30 Syllabus and pages of different natures (scientific articles, web sites pages, etc.). The average length of these documents is about 23 pages.

Our testing corpus is composed of 1000 documents in French, mainly of learning nature: Support of Courses, Assignments, PowerPoint presentations, Syllabus, and documents of different nature. These documents are files in different formats (DOC, PDF, PPT, HTML, TXT, etc.) and have an average length of 53.6 pages.

5.1 First step: Learning objects annotation

To evaluate this step, our testing corpus was annotated by two experts: for each learning object spotted, they affect to it a category. The results of the SRIDoP annotation process are illustrated in the table below where NOA: Total number of annotated objects, NOAC: Number of objects annotated correctly, NOMAC: Number of objects annotated by the experts:

Learning Object Category	NOA	NOAC	NOMAC	Precision (%)	Recall (%)	F-score (%)
Plan	88	85	98	96,59	86,73	91,40
Course	72	60	85	83,33	70,59	76,43
Definition	228	140	266	61,40	52,63	56,68
Characteristic	139	124	156	89,21	79,49	84,07
Example	357	349	376	97,76	92,82	95,23
Exercise	760	705	776	92,76	90,85	91,80

Table 2. Experimentation results of the Annotation step

$$Precision = \frac{NOAC}{NOMAC}$$

$$Recall = \frac{NOAC}{NOMAC}$$

$$\text{F - score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

According to the experimentations presented above, the annotation results are promising. Indeed, the precision of the annotation exceeds the 85% for most learning categories (Example, Exercise, Plan, etc). But, concerning the "Definition" category, the corresponding precision is average. This derives owing to the fact that certain rules can annotate at the same time objects reflecting or not a

"definition". Such the case of a "Definition" category rule which has as an indicator the occurrence "*is is|are* " and as clue *"a|an|the*". These indicators and clues may exist within a textual segment of a defining nature or not. During the experimental phase, we could also note that the effectiveness of the annotation is closely related to the document segmentation effectiveness.

5.2 Second step: Indexing annotated objects

To test this module, we formulated 25 queries for each learning category. These queries deal with the fifteen topics of the learning and testing corpus. For each learning category, we illustrated the number of the returned results and the number of the relevant results given the whole set of the entered queries. The results are presented in the table below (Table III), where **NR**: Total number of results, **NRP**: Number of relevant results, **NRRU**: Number of relevant objects existing in the index.

$$Precision = \frac{NRP}{NR}$$

$$Recall = \frac{NRP}{NRRU}$$

$$\text{F - score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Learning Object Category of the Query	NR	NRP	NRRU	Precision (%)	Recall (%)	F-score (%)
Plan	72	66	77	91,67	85,71	88,59
Course	43	35	54	81,40	64,81	72,16
Definition	156	112	193	71,79	58,03	64,18
Characteristic	94	86	112	91,49	76,79	83,50
Example	213	198	230	92,96	86,09	89,39
Exercise	517	465	520	89,94	89,42	89,68

Table 3. Experimentation results of the Learning Objects indexing

At the end of these experiments, we conclude that the results of the indexing step depend on the annotation results. The searching process quality improves with the annotation process one. This latter depends on the segmentation process quality as we have mentioned in the above.

5.3 Third step: Sorting Learning objects

Following the extraction of learning objects, we sorted these objects according to their similarity with the class C_{user}. With reference to many experiments, we have fixed the threshold value è at : (i) 0.1 for the *Course* and *Definition* categories, (ii) 0.3 for the *Plan* and the *Example* categories, (iii) 0.45 for the *Characteristic* and *Exercise* categories.

On one side, decreasing the è value reduces the set of relevant objects, on the other side, increasing it leads to the selection of irrelevant objects.

We assigned each object into one of the following categories: **A** (objects sorted as relevant), **B** (objects sorted correctly as relevant), **C** (relevant objects). The precision and Recall and F-score for each learning category are calculated as:

$$Precision = \frac{B}{A}$$

$$Recall = \frac{B}{C}$$

$$\text{F - score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

We obtained an average of Precision=86%, of Recall=75%, and of F-score function= 80,12% for all the studied learning categories. Through our experiments, we conclude that the sorting step results depend not strictly on the annotation and indexation ones. There are other parameters which influence the classification results as the training corpus, the choice of the indexing terms, etc. We illustrate these results in Figure 3.

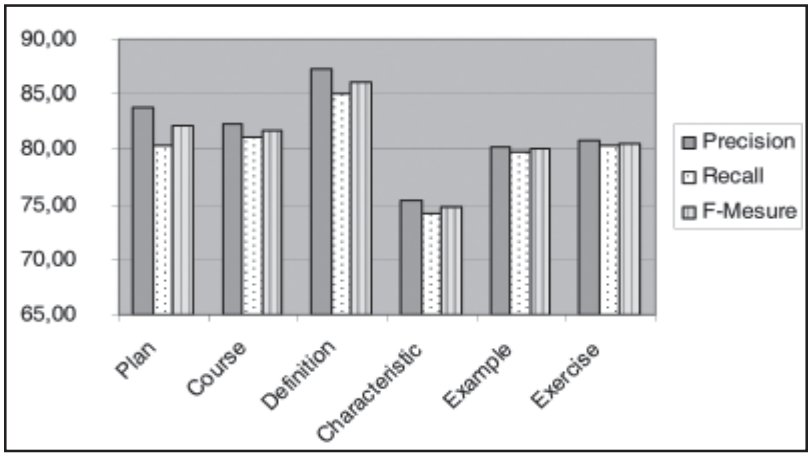

Figure 3. Precision, Recall et F-Mesure of the sorting objects step

The figure above shows, for each object type (represented on the x-axis), its precision value represented in blue, its recall value dotted and F-Measure value shown in stripes. We find that the precision values are between 75% and 87%, those of the recall are between 74% and 85%. Note that the sorting step results does not depend strictly on those of the annotation step but on other parameters such as the training corpus, the choice of index terms, etc.

6. Conclusion and Future Works

In this paper, we proposed a model for learning objects retrieval from documents. To develop it, we proceed by a semantic annotation of learning objects, then an indexation of these objects to find relevant learning objects for queries associated with semantic categories. Through the evaluation results, we observe the originality of a learning object indexation based on a semantic annotation relatively to a key-words searching system. This work comes within the context of learning objects processing and retrieval. Actually, it constitutes a considerable target in many application domains as the e-learning domain, training courses domain, data management systems, etc. One of the future works that we propose is to extend the semantic map of the pedagogical objects categories by other categories as Method, Author, etc. We also look forward to fuse the annotation and classification results using a score function to perform the accuracy SRIDoP system.

References

[1] Buckley, C., Salton, G., Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment, *In:* Proc. International ACM SIGIR Conference, pp. 292-300.

[2] Ittner, D.J., Lewis, D.D., Ahn, D. D. (1995). Text categorization of low quality images, *In: Proc. SDAIR-95, Las Vegas*, p. 301-315.

[3] Dehors, S., Faron-Zucker, C. (2006). QBLS: A Semantic Web Based Learning System. *In: Proc. of the World Conference on Educational Multimedia, Hypermedia & Telecommunications*, ED-MEDIA, Orlando.

[4] Dehors, S., Faron-Zucker, C., Kuntz, R. (2006). Reusing Learning Resources based on Semantic Web Technologies. *In: Proc. of the International Conference on Advanced Learning Technologies*, Kerkrade.

[5] Desclés, J.P. (1997). Systèmes d'exploration contextuelle. *In C. Guimier (ed.) Cotexte et calcul du sens*, Presses Universitaires de Caen.

[6] Desclés, J.P. (2006). Contextual Exploration Processing for Discourse Automatic Annotations of Texts. *In: Proc. of The Florida Artificial Intelligence Research Society (FLAIRS)*, invited speaker, Florida, p. 281-284, AAAI Press.

[7] Desclés, J.P., Djioua, B. (2007). La recherche d'informations par accès aux contenus sémantiques : vers une nouvelle classe de systèmes de recherche d'informations et de moteurs de recherche (Aspects linguistiques et stratégiques). *Revue Roumaine de Linguistique*, Tome LII, N° 1-2, Janvier-Juin.

[8] Dister, A. (1997). Problématique des fins de phrase en traitement automatique du français. *In*: *À qui appartient la punctuation ? Actes du colloque international et interdisciplinaire de Liège,* Bruxelles.

[9] Djioua, B., Desclés, J-P., Mourad, G. (2007). Annotation et indexation des flux RSS par des relations discursives de citation et de rencontre : le système FluxExcom. *Analyse de texte par ordinateur, multilinguisme et applications*, 75e congrès de l'ACFAS, Canada.

[10] Djioua, B., Garcia-Flores, J., Blais, A., Desclés, J.P., Guibert, G., Jackiewe, A., Le Priol, F., Nait-Baha, L., Sauzay, B. (2006). EXCOM : an automatic annotation engine for semantic information. *In:* Proc. of The Florida Artificial Intelligence Research Society (FLAIRS), Florida, p. 285-290, AAAI Press.

[11] Elkhlifi, A., Faiz, R. (2009). Automatic Annotation Approach of Events in News Articles. *International Journal of Computing & Information Sciences*, p. 19-28.

[12] Elkhlifi, A., Faiz, R. (2010). French-Written Event Extraction Based on Contextual Exploration. *In:* Proc. of The Florida Artificial Intelligence Research Society (FLAIRS), AAAI Press, Florida.

[13] Hassan, S., Mihalcea, R. (2009). Learning to identify educational materials. *In:* Proc. of the Conference on Recent Advances in Natural Language Processing (RANLP), Bulgaria.

[14]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, 1971, p. 313-323.

[15] Mourad, G. (2002). La segmentation de textes par Exploration Contextuelle automatique, présentation du module SegATex. *Inscription Spatiale du Langage : structure et processus ISL sp.*, Toulouse.

[16] Salton, G (1991). Developments in automatic text retrieval. *Science,* 253 (5023). 974-980.

[17] Smei, H., Ben Hamadou, A. (2005). Un système à base de métadonnées pour la création d'un cache communautaire-Cas de la communauté pédagogique. *In:* Proc. of the International E-Business Conference,Tunisia.

[18] Smine, B., Faiz, R., Desclés, J.P. (2010). Analyse de documents pédagogiques en vue de leur annotation. *Journal of new information technologies (RNTI)*, E-19, Ed. Cépaduès, p. 429-434.

[19] Thompson, C., Smarr, J., Nguyen, H., Manning, C. (2003). Finding educational resources on the web : Exploiting automatic extraction of metadata. *In:* Proc. of the ECML Workshop on Adaptive Text Extraction and Mining.

[20] Elkhlifi, A., Faiz, R. (2008). Approche d'annotation automatique des événements dans les articles de presse.EGC'2008. 37-42.

[21] Faiz, R. (2006). Identifying Relevant Sentences in News Articles for Event Information Extraction*. Int. J. Comput.Proc. Oriental Lang.,* 1-19.

[22] Elkhlifi, A., Faiz, R. (2007). Machine Learning Approach for the Automatic Annotation of the Events. *In*: Proc. of The Florida Artificial Intelligence Research Society (FLAIRS), AAAI Press, Florida.