I.J. Modern Education and Computer Science, **2016, 6, 13-19**
Published Online June 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2016.06.02

Semantic Annotation of Pedagogic Documents

Benyahia Kadda
Evolutionary Engineering & Distributed Information Systems Laboratory
Djillali Liabes University of Sidi Bel-Abbes, Algeria
Email: benyahiaka@gmail.com

Lehireche Ahmed
Evolutionary Engineering & Distributed Information Systems Laboratory
Djillali Liabes University of Sidi Bel-Abbes, Algeria
Email: elhir@univ-sba.dz

***Abstract*—**To teach, teacher needs help for sharing these educational documents, and especially his knowledge. We present an approach to overcome the difficulty of sharing educational materials and facilitate access to content; we describe semantically these documents to make them accessible and available to different users. The main idea in our annotation approach is based on: (1) Identify key words in a document, to have a good presentation of the document, we extract the candidate words by applying a weighting process and another process using similarity measure, These keywords candidates are reconciled with ontology to determine the appropriate concepts. (2) As document reference generally other documents, we propagate the annotations of references for citing document.
(3) A process of validation will be applied each time an annotation is added in order to keep the coherence of the base of annotation.

After evaluation with several types of pedagogic documents, our approach achieved a good performance; this suggests that teachers can be greatly helped for the semantic annotation of their pedagogical documents.

***Index Terms*—**Semantic annotation, pedagogic document, metadata, information retrieval, ontology, validation.

I. Introduction

Teachers use different teaching documents such as programs, their courses or those of colleagues by subject area. The sharing of these documents on the web enriched the contents and thus to improve quality of the course ensured by the teachers. This document helps teachers to find knowledge close to their specialty and allowing them to enrich their pedagogic package.

On several years, the base of documents becomes increasingly bulky which make the localization a little slow especially a time of the teachers is limited.

Add a semantic layer to words of documents is one of the methods giving more semantics to the documents, and then the research becomes a meaningful, not just words. So a document must be described by a list of concepts linked by relations, it is the semantic annotation.

Yue and Francois [1] see that, a semantic annotation system of technical documents should have the following properties: (1) Be able to notice a concept and not just general types of instances as meaning of a term; (2) to provide accurate and reliable interpretation, taking into account the semantic models of this Treated field; (3) to have a good covering of the text, so that the cost textual fragments can be easily detected and connected.

When we know that systems of semantic annotation will be brought more often to help of sharing documents on the web, how do we admit that the recipe for an annotation system is nothing rathan than a miraculous act of faith, an act of mutual trust between the team that develops the annotation tools and users who takes delivery?

In our approach, we present a valid semantic annotation of pedagogic documents on the web. This approach aims to annotate a document by content and context, by content, we represents documents by keywords extracted with tow process (weighting and measuring similarity) which are connected to the ontology's concepts. By context, as documents reference generally other documents, we propagate the annotations of references to annotate the citing document. We then apply a validation module which make our annotations consistent. The rest of the article is structured as follows: we are discussing the state of art in the following section and present our contribution in Section 3. The experiments and their evaluation are presented in Section 4.

II. State of art

The annotation document is to associate information in these documents, to ensure a true and accurate description of their contents. All annotation tools that have proposed manipulate annotations, which were inspired from annotations that we are used to practice in the paper [2], for this reason, digital annotations, which we are interested for adapting this concept to electronic media. Technically, a semantic annotation consists of assigning metadata to an entity whose semantics are defined in ontology.

We present two types of annotations: annotation by content, using only the contents of the document and the annotation by context, using the relationships between documents.

Beginning with the annotation content, Semantic annotation takes into account the semantics of words, it uses ontologies to realize annotation, every concept of ontology is denoted by one or more terms, Desmontils [3] has index page with keywords attached to an ontology. Yan Bodain [4] proposed an annotation tool KATIA that can annotate a web page by selecting a text area and choosing the corresponding element of the ontology in the hyperbolic tree. Baziz [5] presented an annotation model that builds a semantic core for each document with the concepts and their proximity. Khelif[6] using a domain ontology biopuces to annotate documents, annotating medical texts based on a medical thesaurus is the subject of the work of Pouliquen et al[7]. Staab et al [8] underline the importance of using an ontology for creating semantic annotation, a comparison of the results of research systems, one based on freely generated annotations and the other annotations based on ontologies are made by soo et al[9] wherein they show the advantage of the use of the ontology. Saad [10] develope a automatic annotation tool that supports the semantic annotation of Arabic language Web documents and Maha et al[11] presents a lexical ontology for Arabic semantic relations.

Ontopop, an annotation tool that is based on the combination of information extraction tools (IE) with knowledge representation tools of the Web service is presented by Amardeilh [12].

For the annotation by context using the relationships between documents, Kessler [13] proposed a method of bibliographic coupling based on the assumption that two articles that cite one or more common documents a significant relationship. The co-citation method based on the principle that two references of any dates, commonly cited together have a thematic parity [14], Starting from its principles, Lylia [15] used the technique of spreading independent annotation of content and uses a thematic grouping references built from an unsupervised fuzzy classification.

Boudebza.s [16] focuses on the annotation capitalization and reuse Within Communities of Practice of E-learning. Hakim et al proposed ontology of semantic annotation of learner for reuse in an pedagogic annotation tool. Vadd et al [17] discuses focus question based Inquiry based learning an active teaching learning procedure also student centered.

For semantic annotation systems nothing or almost hadn't been done about the problem of validation especially when it comes to validating bases of annotations that have defined by approximate way. We place ourselves in the context where we have a set of pedagogic documents, the domain ontology and seeking to create a valid annotation.

III. Description of the Approach

To overcome the problem of sharing of educational materials and facilitate access to content, we propose a method of annotation and validation. The task of our system is to take as input a web document and out putting the same content enriched by a valid annotations based on representations of knowledge more or less formal.

Our approach is composed of three modules illustrated in Fig 1.

Module of annotation by content: The module that extracts the candidates words in documents .the candidates word are combined with the concepts of ontology.

Module of annotation by context: the module will extract the references cited in the document and import its annotations as annotation of the document.

Validation module: the module that tests the consistency of annotations, it fires every time an annotation is created.

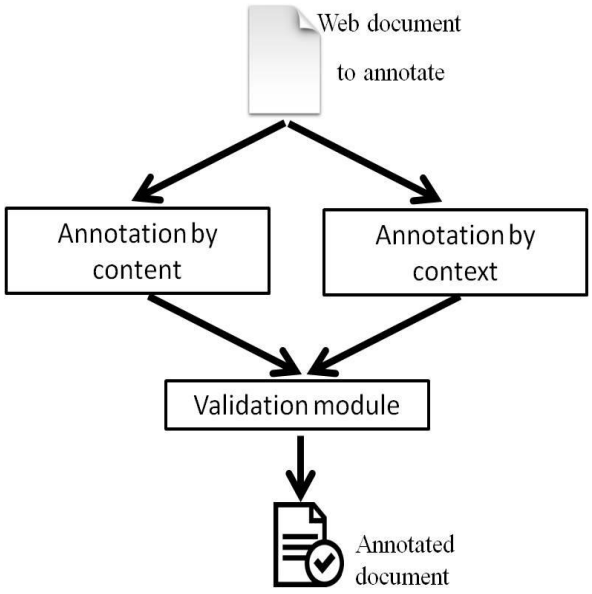

Fig.1. Block diagram of the proposed approach

We present in the following the operation of each module.

a) Module of annotation by content

The objective of this phase (fig 2) is to find the most important words in the document that will be associated with the concepts of the ontology.

1) Selection and cleaning of words

The main question is how to use these textual resources accessible to all, how to profit from these linguistic databases and how to extract the words which be used by such annotation system for represent the document. Linguistic treatment represent the document to be annotated a set of simple and important terms.

First starting with text segmentation step, however, when one makes the occurrence statistics, we see that the most frequent words are function words (or words tools, empty words), as "of", "an", "the", etc. that playing only a syntactic role and giving little sense to documents, so it would not be necessary to take them into consideration in the annotation stage and therefore eliminating these empty words is the second step.

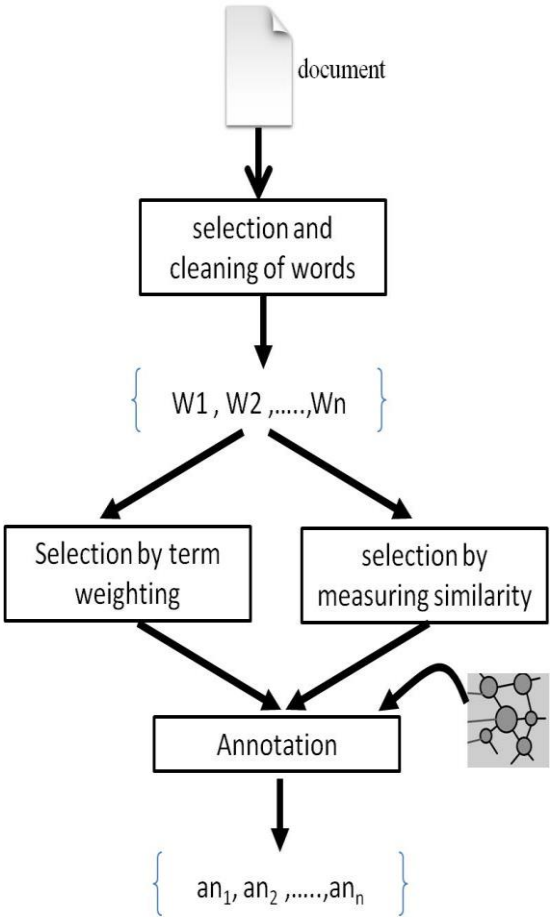

Fig.2. Annotation by content

2) Selection by term weighting

The goal now is to find the words that best represent the content of a document. Based on the principle of [18] "when an author writes a text, he repeats certain terms to develop an aspect of the subject", It is generally accepted a word that often appears in the text is an important concept. Thus, the first approach is to choose words representatives according to their frequency of occurrence. The easiest way is to set a threshold on the frequency: if a word occurrence frequency exceeds the threshold, then it is considered important to the document. But generally, the simple occurrence of word cannot indicate the topic, the meaning or purpose of a text.

The weighting process should provide an iconic representation, compact and informative of document content. It should provide an important indicator to discriminate the terms of each against the other. This important indicator (weight terms) is often measured from three settings: the term frequency, document frequency of term and standardized length document. Several methods are proposed in the literature to measure the term "significant". We interested in the local weighting which the principle is:

The local weighting measures the local representation of a term. It takes into account the local information of the term that depend only on the given document, and gives the importance of the term in this document. We used the logarithmic function that combines tfij (the occurrence frequency of the term ti in the document dj) with a logarithm, is given by:

$$\text{imp} = \propto + \log(\text{tf}_{ij}) \quad (1)$$

Where α is a constant.

This is proposed by [19], aims to mitigate the effects of wide differences between the frequencies of occurrence of words in the document. Thus, by choosing the words which have higher frequencies than the threshold defined by the user to get the words whose informativeness is highest

3) Selection by measuring similarity

In this step, we determined the weight of a word in web document; this semantic weight calcul is based on the similarity measure.

According to a literature study by [20] on the various existing similarity measures, the principle of taxonomic distance-based measures is to count the number of edges separating both ways in a taxonomy. A figure(Fig 3) [21] represents the relationship between any two directions C1 and C2 in a taxonomy compared to their common sense most specific C and relative to the root of the taxonomy

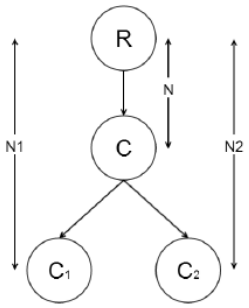

Fig.3. Distances in the ontology used for Wu&Palmer measure of similarity of concepts

Measuring Rada[22] is the first to use the distance between the nodes corresponding to the two sense on the links hyponymy and hyperonymy:

$$\text{SimRada}(c1, c2) = \text{d}(c1, c2) = \text{N1} + \text{N2} \quad (2)$$

The terms located deeper in the taxonomy are always closer than the most general terms, Wu and Palmer[21] proposes to take into account the distance between the common ancestors most specific and root for remedied.

$$\text{SimWup} = \frac{2*\text{N}}{\text{N1}+\text{N2}+2*\text{N}} \quad (3)$$

Leacock and Chodorow[23] are also based on the measurement of Rada, but rather to normalize the relative depth of taxonomy in relation to the senses, they choose a normalization compared to the total depth of taxonomy D and normalize with a logarithm:

$$\text{SimLCH} = -\log\left(\frac{N1+N2}{2*D}\right) \quad (4)$$

We use SimLCH measure with Wordnet , because of its simplicity offered to quantify the similarity of two concepts by semantic distance discovery path graph.

In this phase, a word is accepted if and only if it is strongly related to other words on this document. This decision depends on the selection of a user-defined threshold belongs to the interval [0,1].

4) Annotation

In this step using domain ontology's, we made a passage of the keywords candidate on ontology to define the corresponding concepts. With each passage of one term on ontology, a set of concepts will be presented to teachers to choose the concepts used to create association (word ,concept).

b) Module of annotation by context

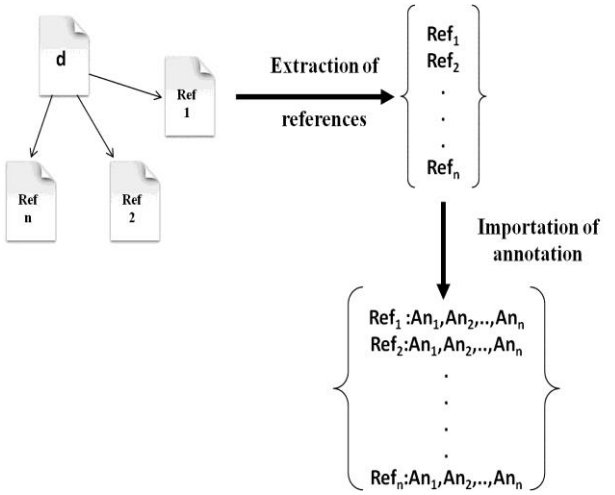

Fig.4. annotation by context

A pedagogic document generally refers to other documents, in this phase, we are interested in the part reference in the document, figure (fig4) presents annotation by context , for each document D, we extract the references ref1,ref2,.., refn which are themselves documents. The results of this phase a set of references.

For each reference of the previous phase, imports of them their annotations which are generally concepts defined in ontology used for annotation without needing content of the document.

c) The validation module

In this phase we are basing on the probability that two different teachers can choose the same concept (keywords) to describe a word is low, making it difficult to achieve consistency in the annotation. Figure (fig5) shows the steps of validation.

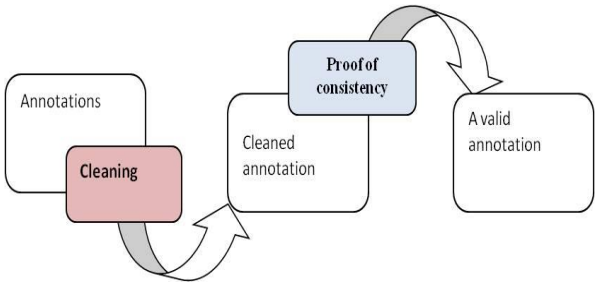

Fig.5. Validation of annotation

Specifications of inconsistency on:

-The redundancy of annotations; and thus the need for cleaning.

-Annotations in conflict that make all inconsistent annotations where the demonstrate consistency becomes a very useful step.

1) Cleaning

Eliminate redundant elements because two references may share common concepts.

2) Proof of consistency

Apply a validation mechanism to make all consistent annotations; it may happen that the resource will be annotated by concepts semantically disjoint and therefore the semantics must recalculate to keep one and eliminate others. We define a search service inconsistency, this service possible to focus the search for inconsistency whenever the annotation is created or propagated. The annotation must be revised when inconsistencies are detected.

The basic idea is to break down the basic annotations maximum subsets of annotations on the same resources, once this decomposition done, proof of consistency is reduced to simple tests of consistency for each subset, the detecting of conflict in annotations of the same resource make an inconsistent subset and the annotations of a subset will be deemed "correct" after he has proved consistency subset. The validation of all the annotations is to make consistent each subset

IV. Experimentation and Results

Our approach does not remain at the theoretical level; we have developed a tool that combines all steps of our approach. This tool is validated on a set of pedagogical document of computer science specialty.

We show, using a set of documents for the benefit of the approach we have proposed to annotate a pedagogic document and the validation of the annotation created.

For this, we use 160 pedagogic documents annotated by teachers .

Our approach is divided into two stages of testing:

1- Proof of interest in the use of two types of annotation (by content and by context)

2- Proof of the usefulness of our validation module for validating the created annotations.

The objective of this study is to evaluate the performance of our approach. To validate our approach of annotation of pedagogic documents, we have developed a tool using Java under the Eclipse environment.

Our test corpus, we collected a set of documents consisting of 53 course documents, directed works 40, 47 PowerPoint Presentation and 20 practical work. The average length of these documents in the corpus is 10 pages.

To evaluate the annotation process, the corpus was annotated by two experts for each spotted pedagogic document, they specify its type.

First, we tested the system without the application of the validation module; the results of the annotation process performed by our system are shown in Table 1.

We defined a quality index:

$$I_{qa} = \frac{Nac}{Na} \quad (5)$$

Na: number of created annotation.
Nac: number of annotation annotated correctly

Table 1. Results of the annotation without validation module

Nature of Document	Na	Nac	Iqa(%)
Course	3650	2800	77%
File of directed works	700	480	69%
File of practical work	120	80	67%
Presentation	790	580	73%

Secondly, we tested the system with the application of the validation module; the results of the annotation process performed by our system are shown in Table 2.

Table 2. Results of the annotation with validation module

Nature Document	Na	Nac	Iqa(%)
Course	2850	2800	98%
File of directed works	500	480	96%
File of practical works	87	80	92%
Presentation	600	580	97%

After these experiments, we note that the use of Using both types of annotations (by content and context) strengthens the semantics of documents since our tool merges the annotations created by the extraction of significant words of the document and those inherited by annotations propagation step of the resources cited in the annotated document, which increases the number of created annotations.

We note that the semantic annotation of pedagogic documents are closely related to the integration of the validation phase in the semantic annotation (Fig 6). The quality index "Iqa" of annotation by integrating the module of validation is higher than that of the annotation which does not use the validation module (Table1, Table2). This is explained by the fact that the number of created annotation is reduced (3650 without validation module and 2820 with validation module for course) which increases the "Iqa", because the validation module cleans and eliminates inconsistent and redundant annotations. Adding new annotation can make annotation base incoherent and therefore the revision is necessary in order to make it consistent either by proof of consistency or elimination of inconsistency especially after the fusion step of the two sets of annotations (by content and by context) where the redundancy may be exist.

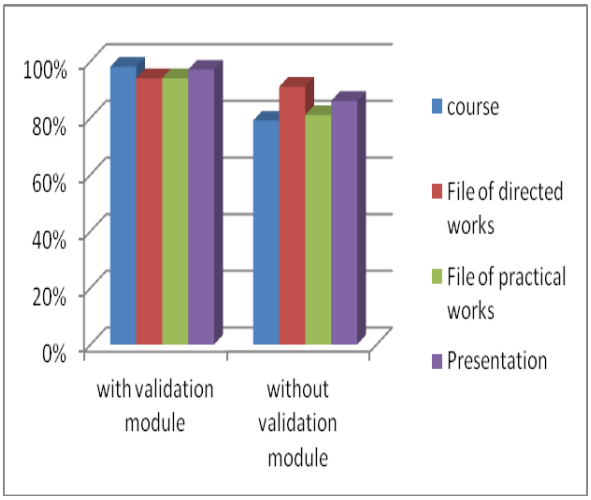

Fig.6. The quality index

We also note that the number of correct annotations is important (2800 for courses) and this is justified by the use of two types of keyword selection that represents the document in module of annotation by content, the selection by weighting and selection by similarity measure which give to annotators more of candidate words in the annotation phase.

The latter is itself dependent on the annotated document type which is an important factor affecting the quality of the annotation, the course document has reached 98%, by cons for directed works and practical work has not exceeded 92%. This is justified by the course content is rich and cited several documents which increases the number of annotations.

The index of quality of annotation exceeds 92% for most types of pedagogic documents, which explains the effect of the validation module on the one hand and the use of the hybrid method of annotation (by context and by content) which enrich the semantics of documents on the other hand.

V. Conclusion

This paper proposes an annotation approach to annotate pedagogical documents in order to improve searching effectiveness for teachers who share knowledge resources. A first step consists in discovering relevant words in a given document (keywords), based on a process to calculate words weight by weighting and by similarity measure. The keywords that result from those operations are presented to user who is responsible for choosing the right terms and also extend the resulting set

with the support of ontology (also referred as semantic networks). In the second step the annotations made in the references of a given document are also imported in a sort of back propagation. A final step for validation purposes removes redundancy and solves inconsistencies (concepts semantically disjointed). Some preliminary tests were made using a tool that implements the referred approach to evaluate the interest by using two types of annotation - by content (keyword extraction of document corpus) and by context (importation of annotations from references) - and also to determine the usefulness of the validation step.

We see, through the evaluation results, our approach allows to annotate and validate annotations which offers a act of mutual trust between the annotators and developers of search engines who takes delivery, especially when we know that the annotations will be brought more and more often to help search engines with decisions making for better answering the requests of teachers.

REFERENCES

[1] Ma, Y., Lévy, F., & Nazarenko, A. (2013). Annotation sémantique pour des domaines spécialisés et des ontologies riches}. In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles} (pp. 464-478). Association pour le Traitement Automatique des Langues}.

[2] S. Bringay, S. Barry, (Septembre 2003) « Information du dossier patient Tâches TV2 : Veille technologique : - Annotations » Projet HTSC, 1er rapport de veille « Document numérique médicaux »

[3] Desmontils, E., Jacquin, C., & Morin, E. (2002). Indexation sémantique de documents sur le Web: application aux ressources humaines. Journées de l'AS-CNRS Web sémantique.

[4] Bodain, Y. (2006, April). Logiciel d'annotation pour la conception de cours sur le web sémantique. In Proceedings of the 18th International Conference of the Association Francophone d'Interaction Homme-Machine (pp. 249-252). ACM

[5] Baziz, M. (2005). Indexation conceptuelle guidée par ontologie pour la recherche d'information (Doctoral dissertation, Toulouse 3).

[6] Khelif, K., & Dieng-Kuntz, R. (2004). Annotations sémantiques pour le domaine Biopuces. In 15èmes Journées francophones d'Ingénierie des Connaissances(pp. 273-284). Presses universitaires de Grenoble

[7] B. Pouliquen, D. Delamarre, and P. Beux. Indexation de textes médicaux par extraction de concepts, et ses utilisations. In Proceedings of JADT'02, 6`emes Journ´ees internationales d'Analyse statistique des Donn´ees Textuelles, mars 2002.

[8] Staab, S., Maedche, A., & Handschuh, S. (2001). An annotation framework for the semantic web. Inst. AIFB, University

[9] Soo, V. W., Lee, C. Y., Li, C. C., Chen, S. L., & Chen, C. C. (2003, May). Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques. In Digital Libraries, 2003. Proceedings. 2003 Joint Conference on (pp. 61-72). IEEE.

[10] Al-Bukhitan, S., Helmy, T., & Al-Mulhem, M. (2014). Semantic Annotation Tool for Annotating Arabic Web Documents. Procedia Computer Science, 32, 429-436.

[11] Al-Yahya, M., Al-Shaman, M., Al-Otaiby, N., Al-Sultan, W., Al-Zahrani, A., & Al-Dalbahie, M. (2015). Ontology-Based Semantic Annotation of Arabic Language Text. International Journal of Modern Education & Computer Science, 7(7).

[12] Amardeilh, F. (2007). Web sémantique et informatique linguistique (Doctoral dissertation, institut de recherche en informatique de toulouse)

[13] Kessler, M. M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American documentation*, *16*(3), 223-233.

[14] E. Garfield. Co-citation analysis of the scientific literature: Henry small on mapping the collective mind of science. Essays of an Information Scientist : Of Nobel Class, Women in Science, Citation Classics and Other Essays, 15(19), 1993

[15] Abrouk, L. (2006). Annotation de documents par le contexte de citation basée sur une ontologie (Doctoral dissertation, Université Montpellier II-Sciences et Techniques du Languedoc)

[16] Boudebza, S., Berkani, L., Azouaou, F., & Nouali, O. (2015). Using an ontological and Rule-based approach for contextual semantic annotations in online comMunities. In E-Learning Systems, Environments and Approaches(pp. 95-115). Springer International Publishing

[17] Vaddi, R. S., Yalamanchili, B. S., & Anne, K. R. (2015). Focus Question based Inquiry Guided Learning for the Attainment of Course Learning Outcomes.International Journal of Modern Education and Computer Science (IJMECS),7(7), 48.

[18] H. Luhn. The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2) :159–165 and 317, April 1958.

[19] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523, 1988.

[20] Tchechmedjiev, A. (2012). État de l'art: mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances. JEP-TALN-RECITAL 2012, 295.

[21] WU, Z. et PALMER, M. (1994). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on ACL, volume 2 de ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

[22] RADA, R., MILI, H., BICKNELL, E. et BLETTNER, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1):17–30.

[23] LEACOCK, C. et CHODOROW, M. (1998). Combining local context and wordnet similarity for word sense identification. WordNet : An Electronic Lexical Database. C. Fellbaum. Ed. MIT Press. Cambridge. MA.

Authors' Profiles

Benyahia kadda: PhD student at EEDIS laboratory, Djillali Liabes University, and also an Assistant at the Computer Science Department, Taher Moulay university, Algeria. Interested by semantic web

Lehireche ahmed: He is a full Professor at the computer science department of sidi belabbes University –Algeria- and a researcher at the Evolutionary Engineering & Distributed Information Systems Laboratory of Djillali Liabes University-algeria. As a professor, his lectures include the various facets of artificial intelligence and semantic in IT. As a researcher, a Director of research, head of the Knowledge Engineering Team at the Evolutionary Engineering & Distributed Information Systems Laboratory. He is author or co-author of more than 40 papers in the field of artificial intelligence, Knowledge Engineering and computer science theory.

