Similarity Detection among Academic Contents through Semantic Technologies and Text Mining

Victor Saquicela[1], Fernando Baculima[1], Gerardo Orellana[2], Nelson Piedra[3], Marcos Orellana[2], and Mauricio Espinoza[1]

[1] Departamento de Ciencias de la Computación, Universidad de Cuenca, Ecuador
[2] Escuela de Ingeniería de Sistemas y Telemática, Universidad del Azuay, Ecuador
[3] Departamento de Ciencias de la Computación, Universidad Técnica Particular de Loja, Ecuador
{victor.saquicela, fernando.baculima, mauricio.espinoza}@ucuenca.edu.ec
{gorellana, marore}@uazuay.edu.ec
nopiedra@utpl.edu.ec

Abstract. Nowadays, the information of university courses is managed by means of syllabus based systems, this is the case of Ecuadorian Higher Education Institutions (IES for its Spanish acronym). However, the syllabus structure is not normalized among all universities, since there is a wide variety of formats and data models used for each IES which naturally, affects academic processes such as the students mobility or credits validation between IES. We have addressed these issues by presenting a proposal based on semantic technologies and text mining methods whose goal is to identify similarities among academic contents.

Keywords: Higher Education, Semantic Web, Ontologies, Text Mining

1 Introduction

In recent years, aiming to improve the common learning procedures in Higher Education Institutions, several researchers have proposed technological initiatives in the education field [2,19,44,17]. Specifically, Ecuadorian IES have been exposed to intensive changes through strict evaluation processes in the academic, administrative and structural points of view [8,26].

The 2011 Organic Law of Higher Education, disposed all IES undergo to an evaluation, acreditation and categorization process [7], and in consequence the tools to improve quality in IES gained special interest.

Regarding academic evaluation and with the aim to ensure quality in higher education, exhaustive analysis have been performed to every career in every IES. Nowadays, in Ecuador, IES are ruled by the Regulation of Academic Regime (RRA for its Spanish acronym), which specifies that each IES is composed by academic units, each academic unit by careers, and each career by subjects. A subject is defined as a set of contents about a specific area [9].

Although this hierarchical structure, each IES maintains the autonomy over its study plans, defining its own contents for each career, thus, this creates a large

heterogeneity over subjects contents. Also, there is a large variety of formats and careers program models in which syllabus are built and kept across the different IES.

Such heterogeneity isolates each IES from others, thus, the creation of a common national repository to store career related data and subject contents (syllabus), represents a very challenging task. A syllabus is a document in which contents and learning methods for students are defined by the members of academic units [21,44]. In addition, the lack of methods, processes and procedures to identify similarities among career contents in different IES makes students mobility through credits validation a difficult task [17]. With this background, students mobility among IES, at a national or international level, becomes a challenge, mainly due to the fact of not owning a systematized format for subject contents, hence, there is a large dependency of manual processes for homologations or equivalences matching between subjects.

Currently, efforts like [11,16,30] have introduced the benefits of introducing new information technologies into the education field. Knowledge representation has been applied to the learning context, specifically to represent learning resources, mainly implementing on-line learning modalities (e-learning) [10].

In this paper, we propose a new similarity detection approach for academic contents among Ecuadorian IES. This approach is aligned with the application of semantic and text mining technologies altogether, which will allow in the future, the construction of a computer system which will be able to provide solutions for the students mobility issue.

The remain of this paper is organized as follows: first, we describe the background and related works about semantic web and text mining, it pays special attention to the higher education field; next, we present the process for similarity detection among syllabus; finally, we present some conclusions and future works.

2 Background and Related Works

The need to improve processes, techniques and methodologies around activities such as learning and teaching has led knowledge representation researchers to improve how curricula are modeled and managed semantically [10,11,16,17,30]. The semantic web is an emerging technology to describe resources in the web aiming to make information understandable not only to humans but also to machines [6].

In the academic context, institutions produce information and store it in digital repositories. However, the lack of use of open standards and a semantic approach have caused difficulties when integrating and re-using contents through the Web [36]. Yet, important advances in the application of the Semantic Web in the educational context have been made. In [1] and [37] integration architectures and distributed interoperability of digital repositories are proposed, based on a Semantic Web approach, Linked Data technologies and federated queries. Those architectures have been successfully tested to integrate a group of institutional

repositories belonging to universities [35], and they have enabled a federated query environment within this context [40].

An Ontology is an important technology which allows data to be represented in the Semantic Web [6]. Ontologies are models that can be used to structure knowledge, they may create a better interaction between teachers and students and improve the learning outcomes and teaching methods of the academic contents [11]. In [11], an ontology based system is presented to allow the integration and classification of heterogeneous systems, learning objects and curriculums, nevertheless, this work is not focused in solving similarities among syllabus or solving the students mobility problems. On the other hand, Demartini et al., in [17] describes the development of the Bowlogna ontology, this work reflects the interests of the Bologna project for the renewal and standardization of high education in Europe, authors mention the problem of student mobility regarding credits recognition, situation that motivated the development of such ontology and its applications.

Despite the mentioned efforts, similarity detection among syllabus is still an unsolved problem. To address this problem, some related works have been reviewed, [42] presents an approach to identify common research areas using semantic and data mining technologies. In [22], authors developed a software to help students of the Agder University to discover existing relations between computer science courses by using ontologies and semantic web. However, this approach is oriented to an unique field of science within the mentioned university. Finally, [32] presents an ontology model for the computer science area, using an extended version of the Wu & Palmer algorithm described in [46] to calculate semantic similarity between computer science courses, nevertheless, as the work presented in [22], this methodology is limited to one specific knowledge area and the use of an algorithm to measure conceptual distances based with the use of Wordnet service [18] as a knowledge representation approach for English words.

3 Similarity Detection among Academic Contents

With the aim to solve the students mobility issue among higher education institutions, this paper presents a process for academic contents similarity detection among IES. This proposal is inspired by the Linked Data Life Cycle described by Auer in [4] and the approach for the identification of common research areas proposed by Sumba et al. in [42]. The implementation of this process aims to develop a similarity detection software among IES syllabus. Figure 1, shows the proposed process which consists of the following components: i) Data Extraction, module designed to retrieving syllabus data from different IES through data extraction and cleaning processes, ii) Ontological Modeling Module, this is where the model for syllabus description is developed, adapted or extended based on existing methodologies, standards and recommendations for ontology development, iii) RDF-Ization Module, based on the ontological model, it generates syllabus data in RDF format, feeding the new semantic repository. RDF (Resource Description Framework) is a standard model for data interchange on

the Web [20], iv) Patterns Detection Module, through the use of text mining techniques, it discovers similarities and patterns among syllabus and v) Visualization, module designed to exploiting and displaying the semantic data in a comprehensible format to the final users.

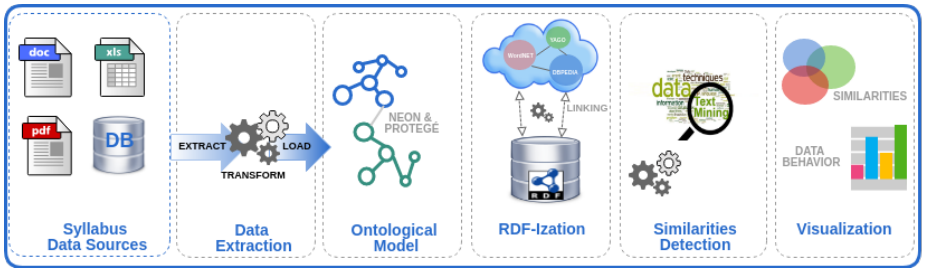

Fig. 1. Similarity detection process among syllabus.

3.1 Data Extraction

Syllabus data will be collected either manually or automatically. The access method will be defined according to the data format, since these can be presented in different formats such as PDF, HTML, WORD, Relational Database, among others. This module execute data pre-processing and cleaning techniques, which are necessary tasks that need special methods for their treatment for its later semantic annotation process [3].

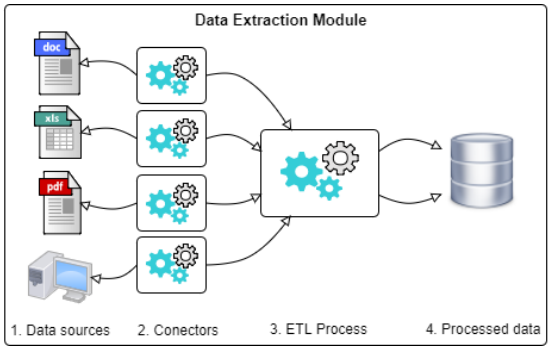

Fig. 2. Data Extraction Process.

Figure 2, shows data extraction process and its tasks: 1) Data sources, refers to documents or databases where syllabus of the IES reside; 2) Connectors, the

software components that must be developed to connect with data sources; 3) ETL process, pre-process and cleaning the data that comes from data sources using specialized software; and 4) Processed data, a new temporary repository containing clean data for the subsequent semantic annotation.

3.2 Ontological Model

One of the main goals of this paper is the use of semantic technologies in order to make syllabus data understandable by both, humans and computers. It is important to reuse existing ontological models in the best possible way to facilitate the inclusion and interoperability of the new data in the web of data [23]. In the background and related works section, several ontological models were presented, which were developed with the aim of solving education and learning related issues. As part of the proposed process, this work will reuse, integrate and extend some ontologies such as BOWLOGNA, FOAF, BIBO, among others, following the NEON ontology development methodology [43], which is widely used in the ontological engineering field. In addition, to apply NEON in the ontology development, this work will use the software PROTÉGÉ, which, due to its ease of use and community support, is widely used in many projects in the field of ontology development [31]. Finally, with the aim of identifying inconsistencies or problems during the ontology development, this work will use the OOPS! Platform, an OntOlogy Pitfall Scanner, which allows an easy ontology evaluation on-line [38].

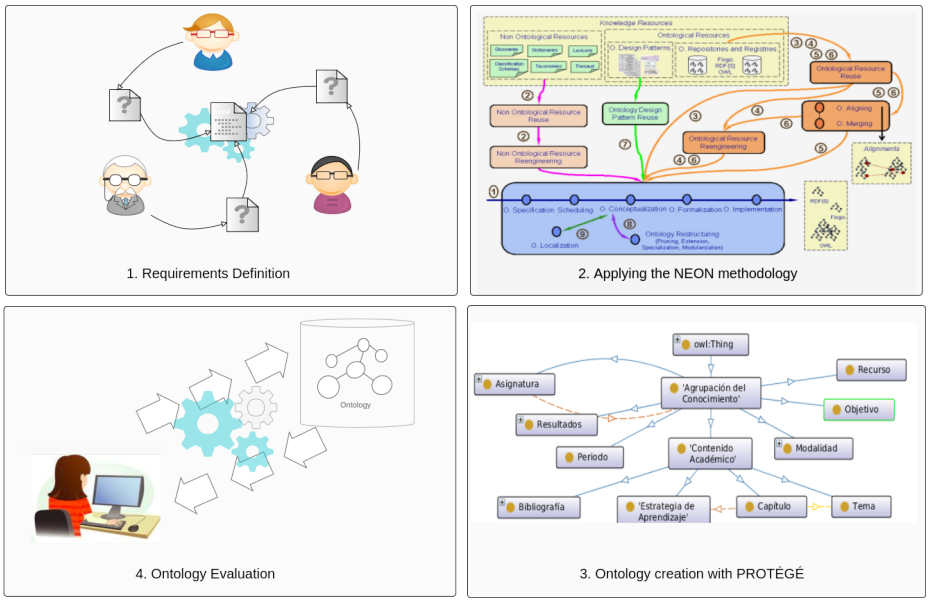

Fig. 3. Ecuadorian Syllabus Ontology Definition Process.

Figure 3 shows the process to reuse, adapt and extend an ontology for Ecuadorian IES syllabus, such process is composed by: 1) Requirements definition, it identifies the requirements of the problem and assigns a team of people for the specific tasks, 2) Application of the NEON methodology, which through the selection of a set of scenarios allows planning the extension/creation of ontology networks, 3) Ontology creation, it uses the PROTÉGÉ software to reuse, unify, extend and generate the required ontology and 4) Ontology Evaluation, it verifies the effectiveness and functionality of the ontology through the data instantiation and query execution.

3.3 RDF-Ization

Once the ontological model has been defined and created, the RDF triplets composed of syllabus data will be created and stored in a repository (TripleStore). This process depends on the data extraction and ontological modelling modules. The Linked Data principles proposed by Berners-Lee in [5] will be followed, also specialized frameworks will be used to generate the data in RDF format.

The RDF-Ization process is shown in Figure 4, where the resulting data from the extraction module will be mapped with the ontology model by using the LOD-GF framework, which consists of a friendly graphical interface with drag & drop functionalities allowing the RDF generation in an easy and intuitive way [13]. Finally, the new RDF data will be stored in a TripleStore for the subsequent exploitation through SPARQL queries.

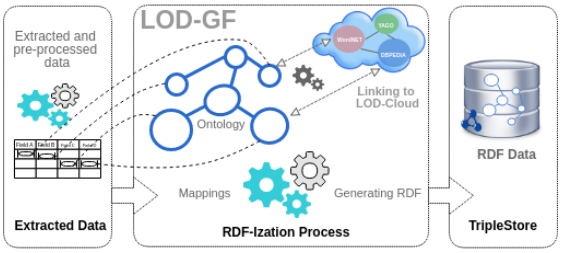

Fig. 4. RDF Generation Process.

3.4 Similarity Detection Techniques

As described in the background and related works section, there are many projects implementing data mining techniques to discover behavioral patterns from raw data. In this paper, we propose the implementation of Text Mining Techniques to analyze unstructured texts, aiming to discover similarities among syllabuses of different Ecuadorians IES careers. Sailaja et al., presented in [39] a "Text Mining Framework", shown in figure 5, which is described by the following 3 stages:

• Stage I: *Text Pre-processing*, including the data cleaning process such as text lemmatization, stopwords removal, dimensionality reduction, among others.

• Stage II: *Text Mining Techniques*, indicates the criteria to select the proper algorithms to process documents.

• Stage III: *Text Analysis*, indicates the use of several tools for information discovery; it means unstructured texts will be converted into meaningful information that helps decision-making.

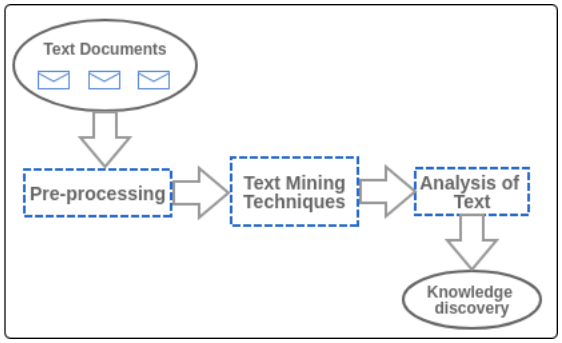

Fig. 5. Text Mining Framework [39].

Based on the Text Mining Framework described above, this paper presents the process shown in Figure 6 to confront the application of text mining on syllabus data. This process will allows patterns discovering, similarities or differences among syllabus, helping to solve the students mobility problem and address the problems discovered in academic content.

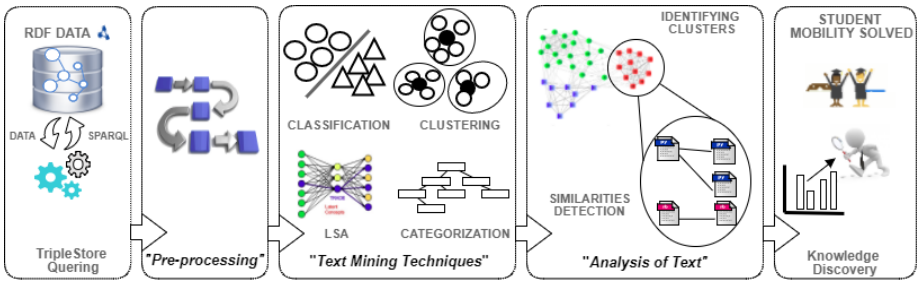

Fig. 6. Text Mining Implementation Process.

Text Pre-processing: The mining process of unstructured texts, consists of several stages; first, there is a data pre-processing [3,33], this stage depends on the specific domain to be analyzed, and involves the elimination of non-relevant words or terms (Stopwords), transformation of text to binary matrices, etc. The platform that will be developed, will extract RDF data from the TripleStore then use NLTK tool to delete insignificant terms. Second, Text Mining algorithms need information in a format they can process, thus, the text will be transformed to numerical matrices using the TF-IDF method recommended in [3].

Text Mining Techniques: There are several techniques and algorithms in the Artificial Intelligence field which perform text mining. As described in [3], clustering is one of the most popular text mining technique and is widely used in classification, visualization and document organization applications. Clustering allows the determination of groups of documents that share common features or have some similarity within the documents collection [14]. Amid the different techniques of clustering, this approach proposes the implementation of the K-means algorithm which is widely used in the field of Data-Mining and Text-Mining, this algorithm divides n documents into k different clusters. In addition, we propose the use of semantic similarity techniques to measure the relatedness of the documents within each group (cluster) [28]. Among some techniques for analyzing the semantic content of a text, we can refer to Genetic Algorithms described in [15], Latent Semantic Analysis (LSA) described in [45], Machine Learning algorithms described in [3], and Word Embeddings techniques that have gained popularity in the text mining community [25]. The algorithms based on WordNET and distance measurement between terms have presented good results [32,29] as well as those based on the information contents [24]. In the process proposed by this paper, we intend to use a combination of the methods, techniques and algorithms described above which will allows the development of a similarity detection platform to solve the students mobility issue.

Analysis of Text: First, with the aim of reducing search spaces, clustering algorithms will be used to identify relatedness among syllabus. Subsequently, on each identified cluster, different text mining algorithms or methods will be executed to compare resulting texts among the different documents (syllabus). Different methods such as cosine similarity will be used to determine the similarity degree between two syllabuses. In addition, new triplets will be created and load on the TripleStore that will allow to identify and query the syllabus clusters as well as the existing similarity among them. In general, this stage indicates the knowledge extraction over the data set and finally, with the help of visualization methods, the detected knowledge and similarities will be exploited.

3.5 Visualization

Recently, visualization techniques have gained interest in the data management field. Visualization of information is an important tool to tell the hidden stories about the data [41]. For example, in [27] and [12] the authors propose techniques related with visualization of the uncertainty in a data set, while in [34] the author

propose an algorithm for visualization of word clouds and semantic relationships existing between these.

The generation of RDF data has several advantages: standardization, interoperability, structured data, etc., however, knowledge of the RDF query language (SPARQL) is needed when accessing this information. Therefore, as part of the similarity detection process, visualization models modules will be developed in order to exploit the RDF data. Text mining algorithms allow extracting valuable information (knowledge) from raw data, so, we propose to implement models or techniques to visualize this information with the purpose of help to final users to envision the students mobility issue and credits recognition in a improved manner.

4 Conclusions

In this work, we analyzed several proposals that want to improve related aspects of syllabus data management for the subjects taught in different careers of several Ecuadorian IES. In all these works the implementation of new technologies are proposed to improve teaching and learning methods and techniques.

The aim of this new approach is to propose a methodological solution for the creation of a common semantic syllabus repository and to discover behavioral patterns and similarities among them.

In the future, with the implementation of the proposed methodology, we intend to create a syllabus platform based on semantic technologies, as well as to implement a series of text mining techniques that will help solving the students mobility issue among IES.

5 Acknowledgment

This work is part of the "Detección de similitudes entre contenidos académicos de carrera a través de la aplicación de tecnologías semánticas y minería de datos" project, supported by the Ecuadorian Corporation for Research and Academia (CEDIA for its Spanish acronym).

References

1. Abad, K., Carvallo, J.P., Espinoza, M., Saquicela, V.: Towards the creation of a semantic repository of istar-based context models. In: Mejia, J., Munoz, M., Rocha, Á., Calvo-Manzano, J. (eds.) Trends and Applications in Software Engineering. pp. 125–137. Springer International Publishing, Cham (2016)
2. Afros, E., Schryer, C.F.: The genre of syllabus in higher education. Journal of English for Academic Purposes 8, 224233 (2009)
3. Allahyari, M., Pouriyeh, S.A., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: A brief survey of text mining: Classification, clustering and extraction techniques. CoRR abs/1707.02919 (2017), http://arxiv.org/abs/1707.02919

4. Auer, S., Bhmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Williams, H.: Managing the life-cycle of linked data with the lod2 stack. The Semantic Web – ISWC 2012: 11th International Semantic Web Conference (2012)
5. Berners-Lee, T.: Linked data (2009), https://www.w3.org/DesignIssues/LinkedData.html
6. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American (2001)
7. CEAACES: Reforma al reglamento transitorio para la tipologa de universidades y escuelas politcnicas y de los tipos de carreras o programas que podrn ofertar cada una de estas instituciones (2012), http://www.ceaaces.gob.ec/sitio/wp-content/uploads/2013/10/REFORMA-AL-REGLAMENTO-TRANSITORIO-PARA-LA-TIPOLOGI%CC%81A-DE-UNIVERSIDADES-Y-ESCUELAS-POLITECNICAS.pdf
8. CEAACES: Suspendida por falta de calidad. el cierre de catorce universidades en ecuador (2013), http://www.ceaaces.gob.ec/sitio/wp-content/uploads/2013/10/CIERRE-DE-UNIVERSIDADES-placas-ok.pdf.
9. CES: Reglamento de rgimen acadmico (2013), http://www.snna.gob.ec/wp-content/themes/institucion/dw-pages/Descargas/regimen_academico.pdf
10. Chung, H., Kim, J.: Semantic model of syllabus and learning ontology for intelligent learning system. Computational Collective Intelligence. Technologies and Applications: 6th International Conference, ICCCI 2014 p. 175183 (2014)
11. Chung, H., Kim, J.: An ontological approach for semantic modeling of curriculum and syllabus in higher education. International Journal of Information and Education Technology 6, 365369 (2016)
12. Collins, C., Carpendale, S., Penn, G.: Visualization of uncertainty in lattices to support decision-making. IEEE VGTC Conference on Visualization p. 5158 (2007)
13. de Cuenca, U.: Lod-gf: An integral open data generation framework (2017), https://ucuenca.github.io/lodplatform/
14. Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval p. 318329 (1992)
15. Dahiya, R., Singh, A.: A survey on text mining using genetic algorithm. International Journal Of Innovative Research And Development 3(5) (2014)
16. Deliyska, B., Manoilov, P.: Ontologies in intelligent learning systems. Intelligent Learning Systems and Advancements in Computer-Aided Instruction: Emerging Studies pp. 31–48 (2012)
17. Demartini, G., Enchev, I., Gapany, J., Cudr-Mauroux, P.: The bowlogna ontology: Fostering open curricula and agile knowledge bases for europes higher education landscape. Semantic Web 4(1), 5363 (2013)
18. Fellbaum, C.: Wordnet and wordnets. Encyclopedia of Language and Linguistics p. 2665 (2005)
19. Garrido, A., Morales, L., Serina, I.: On the use of case-based planning for e-learning personalization. Expert Systems with Applications 60, 115 (2016)
20. Group, R.W.: Resource description framework (rdf) (2004), https://www.w3.org/RDF/
21. Habanek, D.: An examination of the integrity of the syllabus. College Teaching 53(2), 62–64 (2005), http://www.jstor.org/stable/27559222

22. Hokstad, T.: Ontology based study planning and classification of university subjects (2015), https://brage.bibsys.no/xmlui/bitstream/handle/11250/299458/Tor-Erik%20Hokstad.pdf
23. Hyland, B., Atemezing, G., Villazon-Terrazas, B.: Best practices for publishing linked data (2014), http://www.w3.org/TR/ld-bp/
24. Jeong, Y., Song, M.: Applying content-based similarity measure to author cocitation analysis. iConference 2016 Proceedings p. 17 (2016)
25. Jiang, Z., Li, L., Huang, D., Jin, L.: Training word embeddings for deep learning in biomedical text mining tasks. IEEE International Conference on Bioinformatics and Biomedicine (BIBM) pp. 625–628 (2015)
26. Joffre, C.P., Delgado, B., Kosik, R.O., Huang, L., Zhao, X., Su, T.P., ... Fan, A.P.: Medical education in ecuador. Medical Teacher 35(12), 979–984 (2013)
27. Johnson, C.R., Sanderson, A.R.: A next step: Visualizing errors and uncertainty. IEEE Computer Graphics and Applications 23(5), 6–10 (2003)
28. Lee, M., Chang, J., Hsieh, T.: A grammar-based semantic similarity algorithm for natural language sentences. The Scientific World Journal (2014)
29. Meng, L., Huang, R., Gu, J.: A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology 6(1), 1–12 (2013)
30. Miltenoff, P., Keengwe, J., Schnellert, G.: Technological strategic planning and globalization in higher education. Learning Tools and Teaching Approaches through ICT Advancements pp. 348–358 (2013)
31. Musen, M.A., the Protg Team: The protg project: A look back and a look forward. AI Matters 1(4), 4–12 (2015)
32. Nuntawong, C., Namahoot, C.S., Brckner, M.: A semantic similarity assessment tool for computer science subjects using extended wu & palmers algorithm and ontology. Information Science and Applications pp. 989–996 (2015)
33. Patel, R., Sharma, G.: A survey on text mining techniques. International Journal Of Engineering And Computer Science 3(1), 5621–5625 (2014)
34. Paulovich, F.V., Toledo, F.M.B., Telles, G.P., Minghim, R., Nonato, L.G.: Semantic wordification of document collections. Computer Graphics Forum 31(3pt3), 1145–1153 (2012)
35. Piedra, N., Chicaiza, J., Lopez-Vargas, J., Caro, E.T.: Guidelines to producing structured interoperable data from open access repositories pp. 1–9 (2016)
36. Piedra, N., Chicaiza, J., Lpez, J., Caro, E.T.: A rating system that open-data repositories must satisfy to be considered oer: Reusing open data resources in teaching pp. 1768–1777 (2017)
37. Piedra, N., Chicaiza, J., Quichimbo, P., Saquicela, V., Cadme, E., Lpez, J., ... Tovar, E.: Framework for the integration of digital resources based-on a semantic web approach. RISTI-Revista Ibrica de Sistemas e Tecnologias de Informao pp. 55–70 (2015)
38. Poveda-Villaln, M., Surez-Figueroa, M., Garca-Delgado, M., Gmez-Prez, A.: Oops! (ontology pitfall scanner!): supporting ontology evaluation on-line. International Journal on Semantic Web & Information Systems pp. 7–34 (2014)
39. Sailaja, N.V., Padmasree, L., Mangathayaru, N.: Survey of text mining techniques, challenges and their applications. International Journal of Computer Applications 146(11), 3035 (2016)
40. Segarra, J., Ortiz, J., Espinoza, M., Saquicela, V.: Integration of digital repositories through federated queries using semantic technologies. In Computing Conference (CLEI), 2016 XLII Latin American pp. 1–9 (2016)
41. Segel, E., Heer, J.: Narrative visualization: Telling stories with data. IEEE Transactions on Visualization and Computer Graphics 16(6), 11391148 (2010)

42. Sumba, X., Sumba, F., Tello, A., Baculima, F., Espinoza, M., Saquicela, V.: Detecting similar areas of knowledge using semantic and data mining technologies. Electronic Notes in Theoretical Computer Science 329, 149–167 (2016)
43. Surez-Figueroa, M.: Neon methodology for building ontology networks: Specification, scheduling and reuse (2010), http://oa.upm.es/3879/2/MARIA_DEL-_CARMEN_SUAREZ_DE_FIGUEROA_BAONZA.pdf
44. Tokatl, A.M., Keli, Y.: Syllabus:how much does it contribute to the effective communication with the students? Social and Behavioral Sciences 1(1), 1491–1494 (2009)
45. Tu, H.T., Phan, T.T., Nguyen, K.P.: An adaptive latent semantic analysis for text mining. In 2017 International Conference on System Science and Engineering (ICSSE) pp. 588–593 (2017)
46. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. 32Nd Annual Meeting on Association for Computational Linguistics pp. 133–138 (1994)