ISSN 2007-9737

Wikipedia-based Learning Path Generation

Claudia Pérez Martínez[1], Gabriel López Morteo[1],
Magally Martínez Reyes[2], Alexander Gelbukh[3]

[1] Universidad Autónoma de Baja California, Instituto de Ingeniería, Mexico
[2] Universidad Autónoma de Baja California, Mexico
[3] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico

{claudia.perez92, galopez}@uabc.edu.mx,
mmreyes@hotmail.com, www.gelbukh.com.

Abstract. We describe a method for automatic generation of a learning path for education or self-education. As a knowledge base, our method uses the semantic structure view from Wikipedia, leveraging on its broad variety of covered concepts. We evaluate our results by comparing them with the learning paths suggested by a group of teachers. Our algorithm is a useful tool for instructional design process.

Keywords. Learning path, educational resources, Wikipedia, adaptive intelligent web-based educational systems, Spanish language

1 Introduction

The diversity of forms of accessing knowledge is one of the most important features of modern society [1]. Consequently, the process of transmission of knowledge turns into a relevant task. Instructional Design (ID) plays a relevant role by developing methods for creating learning experiences that help to develop and enhance skills and knowledge [2]. Such methods establish a set of phases of the process. One of them is curriculum sequencing. The main objective of curriculum sequencing is to select the most suitable individually planned sequence of knowledge and tasks [3].

The sequence of knowledge units that conform a learning path is called High Level Active Learning Path, or simply a Learning Path [4]; see Fig. 1.

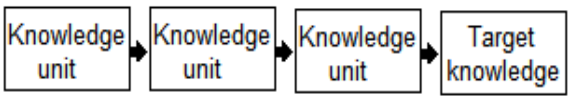

Fig. 1. Learning path

A learning path is designed for a new knowledge unit to be learned; this is the target knowledge unit for this path. The learning path to the target knowledge unit includes prior knowledge, which is necessary to understand the new knowledge.

Since the learning path must be adapted to the student's profile, the learning path building is challenging in web-based adaptive educational systems. The student profiles in the web environment can be more diverse than the profile of the students in the classroom. Different learning path generation approaches has been developed. Many of these developments are based on specific characteristics of each particular student, for example, the results of a pre-testing, the student's emotional current state, or his or her previous experience of use of educational resources.

However, to determine the learning path for a particular profile, one need first to know the desired final state of the student, i.e., what the student should know to understand a new knowledge unit, i.e., a generic learning path. After this, the generic learning path could be personalized by applying some learning strategy.

For each new knowledge unit, a new generic learning path needs to be built. The problem is how to construct at each moment a learning path for each new knowledge unit. To conduct the generic learning path building process, one should have a very complete knowledge base to extract the necessary information for each particular request at any time. In addition, such knowledge base must be machine-readable, that is, semantically understandable. This facilitates extraction of information from the knowledge base.

A useful and well-known structure for the knowledge base is ontology. An ontology is a formal structure for representing knowledge. It names and defines the types, properties, and interrelationships of the concepts in a specific domain of knowledge. Such characteristics make the ontology a tool convenient to build the learning path. The learning path is formed of related knowledge; from an ontology, it is possible to extract knowledge relationships relevant for learning planning.

Nevertheless, to build a new ontology usually results in high cost. In addition, usually ontologies are limited to one specific domain of knowledge. This problem has been confronted in the Natural Language Processing (NLP) area, which requires linguistic resources for different processes. As an output of such a collaborative effort, for example, we can mention WordNet [5], a large lexical database of English. Nevertheless, it cannot directly be used for other languages. A number of researchers found Wikipedia to be a good alternative for the needs related to lexical resources [6].

In spite of the fact that Wikipedia was designed to provide knowledge universally to human readers, NLP experts have founded in it interesting characteristics that turn Wikipedia into a valuable linguistic resource. It can be viewed as an enormous thesaurus, corpus, or ontology [6].

In view of this, in this paper we propose the use of Wikipedia as a linguistic resource for the development of a method to build a generic learning path for a specific target knowledge. This generic learning path is ready to be used by different learning strategies. Then different personalized learning paths can be created.

The main contributions of this work are consist in (i) providing a utility in the educational area: the learning path is useful as prior information for any learning strategy; (ii) proposing the possibility to use a well-studied and widely available linguistic resource for educational tasks.

The paper is organized as follows. In Section 2, different approaches to the learning path building are described. In Section 3, we present our proposal on the learning path building. Section 4 describes the validation of our results based on a survey of the teacher judgments. Finally, in Section 5, conclusions are drawn.

2 Related Work

2.1 Instructional Design

The term *instructional design* refers to the systematic and thoughtful process of converting learning principles and instruction overview into plans for instructional materials, activities, information resources, and evaluation resources [2].

This process is mainly based on the learning and instructional theories. Constructivism is currently the dominant theoretical basis for instructional design. It views the role of prior knowledge as an important element for the learning to take effect [7]. In particular, in the curriculum sequencing, the task is to build learning paths for given target knowledge. Based on the constructivist theory, the learning path for given target knowledge can be constituted for prior knowledge plus the target knowledge.

2.2 Approaches to Learning Path Generation

Adaptive and intelligent web-based educational systems (AIWBES) attempt to be more adaptive by building a model of the goals, preferences, and knowledge of each individual student and using this model throughout the interaction with the student in order to adapt the plan to the needs of that student. They also attempt to be more intelligent by incorporating and performing some activities traditionally performed by a human teacher, for example, building a personalized learning path [3].

The goal of the curriculum sequencing technology, also referred to as instructional planning technology, is to provide the student with the most suitable individually planned sequence of knowledge units to learn and sequence of learning tasks (examples, questions, problems, etc.) to work with. In other words, it helps the student to find an "optimal path" through the learning material [4]. Automatic generation of learning path has been dealt with using different approaches, some of which are mentioned below.

Chen, Lee, and Chen [8] used the Item Response Theory and the difficulty parameter to model course material. The courses were previously labeled and selected. Huang, Huang, and Chen [9] used a genetic algorithm and an

algorithm of case-based reasoning. Their approach is based on a pre-test to students to gather incorrectly learning concepts through a computerized adaptive assessment; the different learning trajectories were pre-designed by experts in the domain of knowledge.

Chen [10] proposed constructing customized learning paths based on the simultaneous consideration of the level of difficulty of the courses and the concept of continuity of successive courses during the learning process. This approach also carries out an analysis of the textual content of educational resources. However, it requires a manual process in which educational resources are labeled with a difficulty level.

Chen and Duh [11] used the Fuzzy Item Response Theory and the fuzzy set theory to estimate the skills of students and recommend an appropriate course. The architecture of the proposed system includes an off-line module, in which a set of experts design the evaluation items for the construction of learning content.

Fazlollahtabar and Mahdavi [12] proposed to use a neuro-fuzzy approach to infer the characteristics of the student and to create and update the student profile taking into account the opinion of the professor. The algorithm selects a learning path j-th previously constructed, based on the result of the characterization of the student.

Chen, Peng, and Shiu [13] proposed a method based on a fuzzy clustering approach applied to a set of pre-designed evaluation items. They create a conceptual map based on ontologies, which is used to generate customized learning paths.

Katuk and Ryu [14] developed a system based on the flow system theory. The system determines two variables for the student: the student's capabilities and the level of difficulty of the tasks. Then, a learning path is generated for the student. In this approach, the topics and sequencing are previously built, and one of these is selected for sequencing engine based on the student model.

Fung, Tam, and Lam [15] proposed the creation of learning paths via correlation of information by grouping of concepts, which are sent to a rule-based genetic algorithm to find the best learning path. They used a statistical method of extraction of keywords to extract all relevant topics concerning the modules of a course. This proposal generates a learning path.

Therefore, there exist different approaches to deliver a learning path to the student, and in the majority of cases, the learning path is previously built. Then the algorithm has a limited operating domain. Only the last of the mentioned approaches effectively builds a learning path by analyzing directly the educative resource and applying statistical methods.

2.3 Semantic Similarity of Texts

Measuring similarity between texts has been actively addressed in frame of natural language processing. Typically, a text corpus is used for this task [16]. Soft measures taking into account similarity between individual words have been recently proposed [17]. In some works, rather sophisticated machine-learning methods have been used [18].

Traditionally, text similarity measures have been based on individual words; however, recently algorithms based on n-grams [19] and especially the newly proposed syntactic n-grams [20, 21], including non-continuous syntactic n-grams [22], have been proposed. While the majority of such algorithms are based on vector space model of the text, there exist graph-based approaches to text similarity [23], which resemble the approach described in this paper. Promising approaches to semantic text similarity are based on recognizing textual entailment [24], detection of concepts [25, 26, 27], and application of linguistic patterns [28]. Recently introduced neural models often significantly outperform traditional methods [29].

Semantic text similarity measures are used for classification [30] and clustering [31] of documents, as well as for search in the web [32], in social networks [33, 34], as well as in specialized document collections [35].

2.4 Wikipedia as a Linguistic Resource

Wikipedia is the largest online encyclopedia. It was created under the wiki paradigm. The articles are built asynchronously, in a collaborative manner, and are edited by the “wikipedians” in a distributed way. The structure of Wikipedia is defined by an ordered set of articles, which have internal references to other articles and external references to academic literature.

Each article describes only one concept. The article contains one or more sections, the first of them being a summary of the whole article. The articles are organized in sets called *categories*. The categories are built in a hierarchical structure of nodes, which can contain other categories or articles.

Wikipedia can be seen as a graph, where each leaf node is an article. The links that point to other articles are identified as targets, and the links that point to the given article are called anchors. The relationships between articles by links can represent semantic distance. Milne and Witten [42] established a numerical measure of semantic relatedness between two Wikipedia articles *a*, *b*:

$$relatedness(a,b) = \frac{\log(max(|A|,|B|)) - \log(|A \cap B|)}{\log(max(|W|)) - \log(\min(|A|,|B|))} \quad (1)$$

In this formula, *a* and *b* are two Wikipedia articles; *A* y *B* are the sets of articles linked with *a* and *b*, respectively; *W* is the entire set of Wikipedia articles.

Fig. 2 shows a representation of the articles in the form of a circle and the categories to which they belong (internal nodes in the figure). Links between articles are shown, too. It is plausible that if two articles point to the same internal destination, these two articles are related.

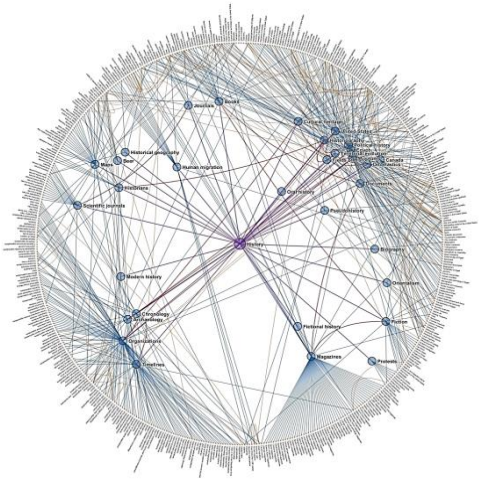

Fig 2. Structure of categories and articles from Wikipedia (Chris Harrison) [14]

The Wikipedia articles can be seen as nodes of an ontology. Each article has an identifier, which is its URI. There are relationships between articles and between categories, though relationship types are not specified.

In spite of the fact that Wikipedia was created for encyclopedic use, researchers in natural language found Wikipedia useful as a linguistic resource. They have identified in it characteristics of useful linguistic resources such as a thesaurus, an ontology, and a concept network structure [6]. Such kind of useful linguistic resources, generally, is built for a specific knowledge domain, due to high cost of its construction. The most attractive characteristic of Wikipedia as a linguistic resource is that it covers very diverse knowledge domains. In addition, it is not necessary to spend any effort on its construction, since it already has been collaboratively constructed, and keeps growing.

Wikipedia has been used a linguistic resource for semantic analysis [36], information retrieval [33], sense disambiguation [37, 38] and named entity disambiguation [39], building semantic hierarchies [40], and machine translation [41].

3 Building a Learning Path

In this section, we describe a model that is able to construct a learning path for a target knowledge; see Fig. 3.

In the sequel, we describe the proposed algorithm to generate a learning path without any pre-existing educational resource. We use only the Wikipedia as an external linguistic resource.

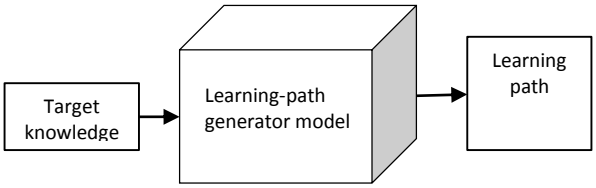

Fig. 3. Learning path generation model

3.1 Considerations for the Algorithm

The key idea of our approach is to consider topological ordering of the graph of Wikipedia articles: given a node in the graph,

- find nodes to which it links (which are probably pre-requisites of learning),
- expand this set recursively by adding the nodes to which they link (with a suitable termination or selection condition), and
- find topological ordering on this set (or "as much topological as possible" if it contains cycles, that is, remove some links, as few as possible, chosen in an optimal way, to make the subgraph acyclic).

The topological order means that all pre-requisites for each topic will be learnt before the given topic.

A learning path for specific target knowledge (tk) can be seen as an organized set of knowledge units (ku), each one representing the prior knowledge for a new knowledge unit, called then target knowledge. The last element in the learning path is precisely the final target knowledge. The learning path can be visualized as an acyclic directed graph, where each node represents a concept (knowledge unit) that is to be learned by the student before learning the next one; see Fig. 4.

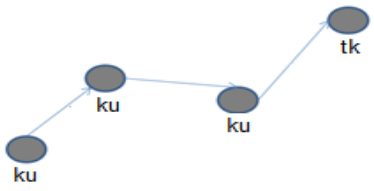

Fig. 4. Graphic representation of learning path

Our approach to building a learning path in an automatic way by using Wikipedia as an additional knowledge resource is based on existing learning theories. Constructivism, as the dominant theoretical basis in the current scheme of instructional design, is based in turn on theories of cognitive learning. One of its principles is related to the cumulative nature of learning. The role of the prior knowledge is seen as an important element for the learning to take place. This principle is based on the theory of information processing, considering the processes occurring in the brain when the learning takes place. The information passes from short-term memory to long-term memory as long as it has a relationship with the existing knowledge. This is a result of the need for the information stored to be significant (integrated with previous knowledge). Therefore, from the point of view of instructional design, in order for students to be able to acquire new knowledge, they must first acquire prior knowledge. Hence, the learning path for a new knowledge is designed with respect to the relevant background related to the new knowledge.

As to Wikipedia, a view of the architecture of Wikipedia is its structure of links between articles and the categories, which establish semantic relationships; see Fig 2. The greater the coincidence, the stronger the semantic relationship. The maximum possible value is 1, the coincidence of the article with itself. In addition, each time in an article there is a link, it is because a significant and geographically diverse set of people (wikipedians) have decided that this concept is relevant in that article. This means that in Wikipedia, there is a universal opinion, by generating links, about which information is relevant and meaningful about some concept (article). This view has been measured, evaluated, and recognized [42].

Yet another relevant fact is that the structure of Wikipedia articles is elaborated by wikipedians under the instructions of the manual of style of Wikipedia. An important specification for our proposal is the formation of sections of an article. The manual specifies for the first section: "The lead serves as an introduction to the article and a summary of its most important aspects" and "it should be able to stand alone as a concise overview. It should define the topic, establish context, explain why the topic is notable, and summarize the most important points, including any prominent controversies."

Then, the first section is the definition of the concept, the most relevant and meaningful information on the concept. It contains links that point to the most important concepts related to this one.

Then, the learning path for the new knowledge is formed of previous or meaningful concepts. They can be found in the definition of the new knowledge. For new knowledge corresponding to a Wikipedia article (the one that has the title similar to this new knowledge), its definition is in the first section. In addition, the first section contains the most meaningful concepts related to the new knowledge by the links. It is relevant to know how it is possible to obtain a numerical measure of the semantic relatedness between the links and the new knowledge.

3.2 Our Proposal

We propose a generation algorithm for the learning path for the given new knowledge. This new knowledge is named target knowledge. For some target knowledge, its learning path is a set of important concepts, or knowledge units, for the target knowledge, with the ending node to be the given target knowledge.

For given target knowledge, the learning path is constructed by finding the corresponding Wikipedia article. The algorithm then proceeds to extract all the links of the first section and compute the semantic relatedness between the target knowledge and each link. Finally, based on this measure, the algorithm orders the knowledge units in descending form.

The algorithm can include a parameter, *k*, to set the number of links in the learning path. These links define the set of elements of the learning path. To describe the learning path building process, we state some assumptions as follows.

In Wikipedia content, unlike the categories structure that is a hierarchy, the article link structure is a cyclic graph. This can be seen resembling the structure of the human brain. We associate an event or object with some ideas or concepts. Depending the situation (context), the same ideas can be evoked from another context.

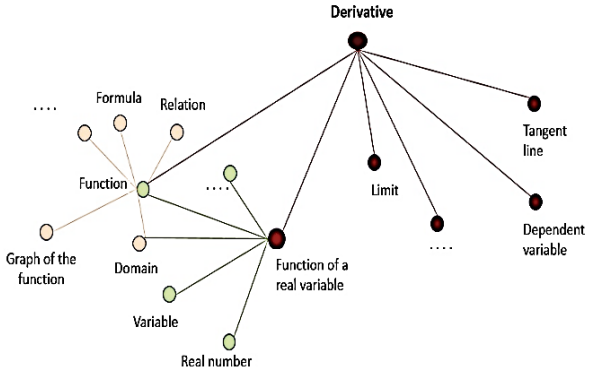

Fig. 5. Knowledge units semantically related to *derivative*

Fig. 5 shows a snapshot of the Wikipedia article "Derivative" and its anchors. "Derivative" has nodes that point to different articles and at the same time, these articles point to other articles or the same articles. This article points to "Function" and "Function of a real variable," which in turn links to "Function."

Therefore, it is possible for a teacher in the classroom to teach the concept "Derivative" after addressing the concept "Function" and then the concept "Function of a real variable." Alternatively, the teacher can select only the concept "Function of a real variable" before teaching the concept "Derivative." What is the correct ordering? Which other of the concepts must be select to build the learning path for "Derivative"? Does the selection depend only on the learning strategy? To answer these questions, it is necessary to know the structure of concepts.

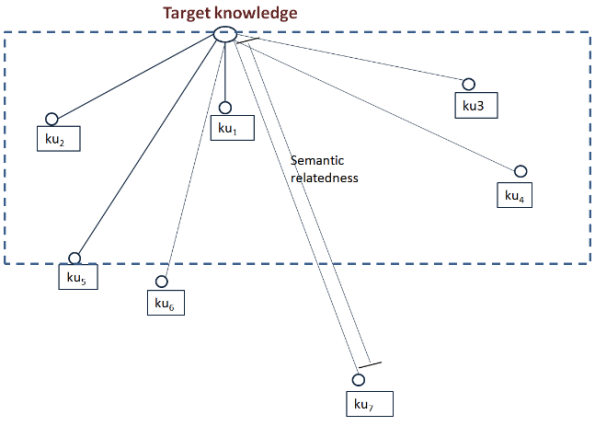

Fig. 6. Graphical representation of a learning path

The teacher undoubtedly knows this structure, but in an automatic system, it is necessary to provide or generate this information. Once the system possesses this information, it can select the appropriate concepts to build the learning path, as described below.

3.3 The Algorithm

In view of the special interest in providing the adequate learning path to each different student, one method for learning path construction first should know which knowledge units are necessary to understand a new concept, or a new knowledge unit. Then, it will choose an adequate strategy to decide which knowledge units to select, basing on the student's profile.

Table 1. Ordered inks for target knowledge "Derivada" (Spanish for *derivative*)

Wikipedia ID	Title	Translation	Relatedness (*derivada*, art$_i$)
52621	*límite_(matemáticas)*	limit (mathematics)	0.8030
158144	*función_matemática*	mathematical function	0.7708
303646	*cálculo*	calculus	0.7537
26334	*derivada_parcial*	partial derivative	0.7473
2867502	*posición*	position	0.7198
1596719	*pendiente_de_una_recta*	slope of a straight ine	0.7080
3833233	*variable_independiente*	independent variable	0.6975
628276	*recta_tangente*	tangent line	0.6777
39365	*movimiento_(física)*	movement (physics)	0.6417
346854	*aproximación_lineal*	linear approximation	0.6354
151680	*gráfica*	plot	0.6212
4846085	*velocidad_instantánea*	instantaneous velocity	0.6083
26743	*diferencial*	differential	0.6076
3857488	*velocidad*	velocity	0.5822
3857488	*velocidad_media*	average velocity	0.5822
1763	*matemáticas*	mathematics	0.5791
2762	*tiempo*	time	0.5389

The process to select the necessary knowledge units to build the learning path consists in selecting a subset of knowledge units from a complete universe of possibilities appropriate for the particular student profile. Our algorithm builds a complete set of knowledge units surrounding the objective knowledge unit. It proceeds as follows:

1. Find a Wikipedia article that matches the target knowledge, art_{target}.
2. Identify the articles linked from art_{target}, denoted $outputLinks(art_{target})$, only from the first, summary, section of the article art_{target}.
3. Compute semantic relatedness between each outgoing link art_i and art_{meta}, denoted $relatedness(art_{meta}, art_i)$.
4. Sort the articles in descending order by their relatedness with the art_{meta}.
5. Select the first *k* articles art_i and assign them to the knowledge units to form the learning path, $ku_i = correspondent\ concept\ from\ art_i$.

This process can be summarized in the following manner:

$$\boldsymbol{learningPath}(tk,k) = \{\, art_i | art_i \in outputLinksFirstSection(tk), i\,\{1,\dots,k\} \wedge \forall i_{\{1,..k-1\}}\, relatedness(tk, art_i) > relatedness(tk, art_{i-1}).$$

The learning path consist of *k* knowledge units, each knowledge unit corresponding to titles of articles *art$_i$*.

To evaluate the effectiveness of the chosen criteria, we also implemented a baseline technique that randomly selects the links from the entire Wikipedia article set correspondent to the target knowledge.

4 Results

4.1 Algorithm

The algorithm was implemented using the JWPL [43] library to extract information from Wikipedia database. We used the version of November 2008 of Wikipedia in Spanish language. While our algorithm, based on link analysis, does not depend on language, Spanish was used to compare the results with the paths manually compiled by Mexican teachers through a survey.

The input data are the target knowledge and the size of the learning path, *k*. In our experiments, we used *k* = 5.

To evaluate the usefulness of the characteristics used for our algorithm (such as using only the links in the first, summary, section of the article and

ordering the links based on the semantic relatedness with the target knowledge), we compared the results with a baseline randomized algorithm.

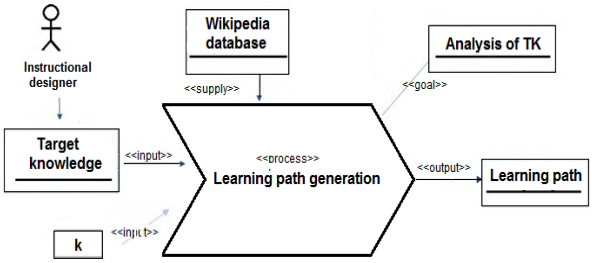

Fig. 7. Learning path generation process

The baseline algorithm takes all links in the article corresponding to the target knowledge and randomly selects k links that conform to the learning path.

Our algorithm generated the following learning path for the concept “Derivada” (derivative), $k = 5$:

$$\boldsymbol{learningPath}(\text{"Derivada"}, proposed\ algorithm) = \{\text{“límite\_(matemáticas)”, “función\_matemática”, “cálculo”, “derivada\_parcial”, “posición”}\}$$

(derivative, limit (mathematics), mathematical function, calculus, partial derivative, position).

4.2 Human Teachers’ opinion

A survey to 32 college math teachers gave 32 different learning paths for the target knowledge “Derivada” (derivative). Table 2 shows the ID of the knowledge units forming the learning path.

4.3 Baseline Algorithm

The baseline algorithm was implemented under the same principle (using the links from the correspondent Wikipedia article to the target knowledge) but using the links indistinctly. The learning paths generated by 32 independent executions of this algorithm are shown in Table 3.

Since the proposed algorithm generates always the same response, we considered 32 similar learning paths to count the coincidence with the teacher opinions.

The frequency of the use de each generated knowledge unit is showed in Fig. 8.

Table 2. Learning paths generated by teachers for the concept “Derivada” (derivative) in the survey

Teacher	Suggested learning path								
1	20	2	11	1	6				
2	1	3	4	6	8				
3	2	11	1	6	8	10	14	16	12
4	2	5	6	11	8				
5	2	11	6	8	1				
6	1	2	7	8	12				
7	7	1	4	10	11				
8	9	11	5	17	18	19	6	8	1
9	2	1	11	6	7				
10	2	7	6	1	12				
11	2	7	1	17	6	36			
12	20	7	2	1	18				
13	29	2	1	6	18	8			
14	21	6	2	8	1				
15	2	1	7	22	23	11			
16	2	1	24	25	20				
17	26	20	27	9	28	30	8	2	1
18	2	1	20	28	31	32	33	34	
19	1	6	2	20	29				
20	20	28	6	8	1	18			
21	36	7	37	29					
22	29	3	1	22	20				
23	29	2	1	35	34				
24	1	2	20	11	28	27			
25	1	2	28	20					
26	2	1	27	22	20				
27	2	7	37						
28	7	37	36	2					
29	2	11	6	20	1				
30	20	28	6	2	11				
31	29	1	20	22	28				
32	2	7	6	1	14	12			

As it can be seen from that figure, the baseline algorithm has a high variation and does not seem to strongly agree with the teachers’ opinions.

The teachers’ opinions have a strong preference for two knowledge units that match the rows 2 and 5, selected by our algorithm of size 5.

4.4 Similarity

As it can be seen from the data, the most significant answers are the links ID 1, 2, 6, 20, 7, 11, 8, 28, 29, 12, 18, 22.

In addition, we found that there was high similarity between two learning paths, when they have at least two matching knowledge units.

Table 3. Learning paths generated by the baseline algorithm for the concept “Derivada” (derivative)

Run	Generated learning path (baseline)								
1	68	35	44	17	84	27	89	38	86
2	79	74	38	6					
3	42	19	13	84	38				
4	13	22	62	29	8	28	76	81	
5	53	82	7	75	13				
6	63	41	69	46	6	57	54	83	69
7	61	83	71	66	86				
8	68	28							
9	74	92	2	50	73	40	26	33	70
10	13	45	65	86					
11	3	69							
12	30	62	49	42	15	14			
13	72	3							
14	65	82	29	83	2	35	35	4	
15	70	90	64						
16	84								
17	47	91	5	57	32				
18	90	69	27	12					
19	43	20	47	27	5				
20	87	12	32	76	10	4			
21	87	67	48						
22	44	49	31	59					
23	31	27	71	66	84	53			
24	73	16	10	4	70	47	8		
25	68	7	57	54					
26	6	20							
27	60								
28	26	53	43	76	58	46	80		
29	29	49	52	12	68	37			
30	14								
31	85	75	24						
32	11	22	33	17	53				

We have evaluated whether there was significant similarity between the most frequent answers in the teachers’ opinions and our algorithm, as well as similarity between the teachers’ opinions and the baseline algorithm. The statistical account is shown in Table 5.

It is observable that our approach, which only considers the summary section of the Wikipedia article and considers the relatedness between the target knowledge and each knowledge unit, shows an increase in the probability of closeness with the learning path generated by human experts, the teachers’ opinion. Finally, we obtain:

$$P_{proposedAlgorithm} = 1,$$

$$P_{baselineAlgorithm} = \frac{5}{32} = 0.1562.$$

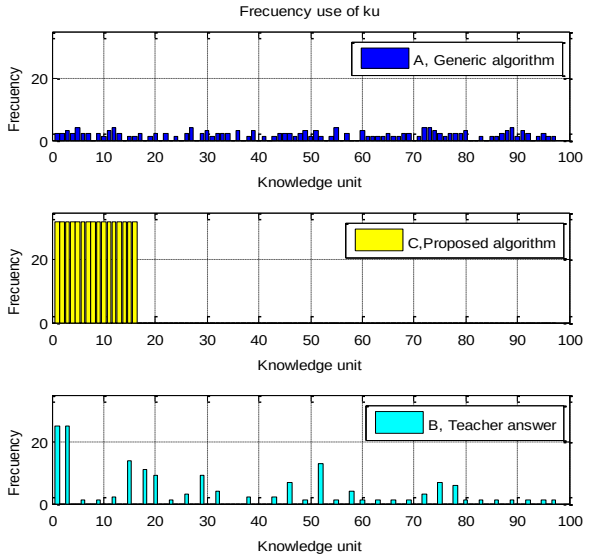

Fig. 8 Frequency of use of the knowledge units (the top graph corresponds to the baseline (generic) algorithm

Table 4. Learning path generated by proposed algorithm

Id	Generated learning path
1–32	1 2 3 4 5

Table 5. Similarity with the teacher’s opinions and the responses of the algorithm

	Similar to teachers’ opinion?	
	Yes	No
Our algorithm	32	0
Baseline algorithm	5	27

5 Conclusions and Future Work

We have presented a methodology to automate the learning path building process. Our technique is based on the use of Wikipedia as a linguistic resource by (i) using the first (summary) section of the Wikipedia article as a definition of a concept; (ii) assuming that its links point to the most relevant related information on the concept; and (iii) considering the semantic relationship measure between two concepts defined by the links [42].

The advantage of using this source of knowledge over other approaches is that in our approach it is not necessary to build dedicated resources, because the Wikipedia is a database

already built and constantly growing. It is possible to obtain a learning path for as many concepts as there are in Wikipedia. Wikipedia represents a broad domain of human knowledge.

A case study has been analyzed with the concept "Derivative" (derivative) from the Spanish Wikipedia. A group of college mathematics teachers was surveyed. They built learning paths for the concept "Derivative" (derivative).

The teachers' opinions are not very homogeneous, but there are some coincidences in the knowledge units selected by the teachers for the learning path for this concept.

Stronger coincidence with the expert opinions was observed for the learning path generated by our algorithm than for that generated by the baseline algorithm. Greater similarity we found between the teachers' opinions and our proposal that between the teachers' opinions and the baseline algorithm.

The generated learning path can be used in the instructional design for educational virtual environments where it is necessary to build learning path in real time for different types of the target knowledge.

In our future work, we will refine the algorithm to obtain a greater similarity with the teachers' opinions and will try to adapt the constructed learning plan to the personality traits of the student [44]. We also expect to consider interactive constructions of the learning plan, which will probably require analysis of human-computer dialog [45].

Acknowledgment

The fourth author acknowledges the support of the Instituto Politécnico Nacional via the grants SIP 20152095 and SIP 20152100.

References

1. **UNESCO, (2005).** *Towards Knowledge Societies*. http://unesdoc.unesco.org/images/0014/001418/141843e.pdf.
2. **Smith, P. & Ragan, T. (1999).** *Instructional design second edition*. Wiley.
3. **Brusilovsky, P & Peylo, C. (2003).** Adaptive and Intelligent Web-based Educational Systems. *Int. J. Artif. Intell. Ed.*, Vol. 13, pp. 2–4.
4. **Brusilovsky, P. (1999).** Adaptive and Intelligent Technologies for Web-based Education. *Künstliche. Intelligenz*, Vol. 4, 19–25.
5. **WordNet. (2010).** Princeton University. https://wordnet.princeton.edu.
6. **Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009).** Mining Meaning from Wikipedia. *International Journal of Human-Computer Studies*, Vol. 67, No. 9, pp. 716–754.
7. **Belloch, C. (2013).** *Diseño Instruccional*. http://www.uv.es/~bellochc/pedagogia/EVA4.pdf
8. **Chen, C., Lee, H., & Chen, Y. (2005).** Personalized e-learning system using Item Response Theory. *Comput. Educ.* Vol. 44, No. 3, pp. 237–255.
9. **Huang, M., Huang, H., & Chen, M. (2007).** Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach. *Expert Syst. Appl.* Vol. 33, No. 3, pp. 551–564.
10. **Chen, C. (2008).** Intelligent web-based learning system with personalized learning path guidance. *Comput. Educ.* Vol. 51, No. 2. pp. 787–814.
11. **Chen, C. & Duh, L. (2008).** Personalized web-based tutoring system based on fuzzy item response theory. *Expert Syst. Appl.* Vol. 34, No. 4, pp. 2298–2315.
12. **Fazlollahtabar, H. & Mahdavi, I. (2009).** User / Tutor Optimal Learning Path in E-Learning Using Comprehensive Neuro-Fuzzy Approach. *Educational Research Review.* Vol. 4, No. 2, pp. 142–155.
13. **Chen, C., Peng, C., Shiu, J. (2008).** Ontology-based concept map for planning personalized learning path. *2008 IEEE Conference on Cybernetics and Intelligent Systems*, pp. 1337–1342. doi: 10.1109/ICCIS.2008.4670870.
14. **Katuk, N., Ryu, H. (2010).** Finding an optimal learning path in dynamic curriculum sequencing with flow experience. *2010 International Conference on Computer Applications and Industrial Electronics (ICCAIE)*, pp. 227–232.
15. **Fung, S. T., Tam, V., & Lam, E. Y. (2011).** Enhancing learning paths with concept clustering and rule-based optimization. In *Proceedings of the 2011 IEEE 11th International Conference on Advanced Learning Technologies* (ICALT'11). IEEE Computer Society, 249–253.
16. **Torres-Moreno, J.-M., Sierra, G., Peinl, P. (2014).** A German Corpus for Similarity Detection Tasks. *International Journal of Computational Linguistics and Applications*, Vol. 5, No. 2, pp. 9–22.
17. **Sidorov, G., Gelbukh, A., Gómez-Adorno, E., Pinto, D. (2014).** Soft Similarity and Soft Cosine

Measure: Similarity of Features in Vector Space Model. *Computación y Sistemas*, Vol. 18, No. 3, pp. 491–504.

18. **Huynh, D., Tran, D., Ma, W., Sharma, D. (2014).** Semantic Similarity Measure Using Relational and Latent Topic Features. *International Journal of Computational Linguistics and Applications*, Vol. 5, No. 1, pp. 11–25.
19. **Çelebi, A., Özgür, A. (2013).** N-gram Parsing for Jointly Training a Discriminative Constituency Parser. *Polibits*, Vol. 47, pp. 5–12.
20. **Sidorov, G. (2014).** Should syntactic n-grams contain names of syntactic relations? *International Journal of Computational Linguistics and Applications*, Vol. 4, No. 2, pp. 169–188.
21. **Sidorov, G. (2013).** Syntactic dependency based n-grams in rule based automatic English as second language grammar correction. *International Journal of Computational Linguistics and Applications*, Vol. 5, No. 2, pp. 23–46.
22. **Sidorov, G. (2013).** Non-continuous Syntactic N-grams. *Polibits*, Vol. 48, pp. 69–78.
23. **Das, N., Ghosh, S., Gonçalves, T., Quaresma, P. (2014).** Comparison of Different Graph Distance Metrics for Semantic Text Based Classification. *Polibits*, Vol. 49, pp. 51–57.
24. **Pakray, P., Poria, S., Bandyopadhyay, S., Gelbukh, A. (2011).** Semantic textual entailment recognition using UNL. *Polibits*, Vol. 43, pp. 23–27
25. **Poria, S., Agarwal, B., Gelbukh, A., Hussain, A., Howard, N. (2014).** *Dependency-based semantic parsing for concept-level text analysis. Computational Linguistics and Intelligent Text Processing. Proceedings of the 15th International Conference, CICLing 2014, Nepal, Part I*, pp. 113–127
26. **Cambria, E., Poria, S., Bisio, F., Bajpai, R., Chaturvedi, I. (2015).** The CLSA Model: A Novel Framework for Concept-Level Sentiment Analysis. *Computational Linguistics and Intelligent Text Processing. Proceedings of the 16th International Conference, CICLing 2015, Egypt, Part II,* pp. 3–22
27. **Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., Hussain, A. (2015).** Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach. *Cognitive Computation*, Vol. 7, No. 4, pp. 487–499
28. **Chikersal, P., Poria, S., Cambria, E., Gelbukh, A., Siong, C. E. (2015).** Modelling Public Sentiment in Twitter: Using Linguistic Patterns to Enhance Supervised Learning. *Computational Linguistics and Intelligent Text Processing. Proceedings of the 16th International Conference, CICLing 2015, Egypt, Part II*, pp. 49–65
29. **Poria, S., Cambria, E., Gelbukh, A. (2015).** Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-Level Multimodal Sentiment Analysis. Proceedings of EMNLP 2015, pp. 2539–2544
30. **Schnitzer, S., Schmidt, S., Rensing, C., Harriehausen-Mühlbauer, B. (2014).** Combining Active and Ensemble Learning for Efficient Classification of Web Documents. *Polibits*, Vol. 49, pp. 39–45.
31. **Cobos, C., Mendoza, M., León, E., Manic, M., Herrera-Viedma, E. (2013).** TopicSearch—Personalized Web Clustering Engine Using Semantic Query Expansion, Memetic Algorithms and Intelligent Agents. *Polibits*, Vol. 47, pp. 31–45.
32. **Alonso-Rorís, V. M., Santos Gago, J. M., Pérez Rodríguez, R., Rivas Costa, C., Gómez Carballa, M. A., Anido Rifón, L. (2014).** Information Extraction in Semantic, Highly-Structured, and Semi-Structured Web Sources. *Polibits*, Vol. 49, pp. 69–75.
33. **Jia, L., Yu, C., Meng, W., Zhang, L. (2013).** Facet-Driven Blog Feed Retrieval. *International Journal of Computational Linguistics and Applications*, Vol. 4, No. 1, pp. 175–194.
34. **Neunerdt, M., Reyer, M., Mathar, R. (2013).** A POS Tagger for Social Media Texts Trained on Web Comments. *Polibits*, Vol. 48, pp. 61–68.
35. **Ordoñez, H., Corrales, J. C., Cobos, C. (2014).** MultiSearchBP: Environment for Search and Clustering of Business Process Models. *Polibits*, Vol. 49, pp. 29–37.
36. **Haralambous, Y., Klyuev, V. (2013).** Thematically Reinforced Explicit Semantic Analysis. *International Journal of Computational Linguistics and Applications*, Vol. 4, No. 1, pp. 79–94.
37. **Melara Abarca, R., Perez-Martinez, C., Gelbukh, A., López Morteo, G., Martinez Reyes, M., Pérez López, M. (2014).** Wikification of Learning Objects using Metadata as an Alternative Context for Disambiguation. *Computación y Sistemas*, Vol. 18, No. 4, pp. 755–765.
38. **Henrich, V., Hinrichs, E., Vodolazova, T. (2012).** An Automatic Method for Creating a Sense-Annotated Corpus Harvested from the Web. *International Journal of Computational Linguistics and Applications*, Vol. 3, No. 2, pp. 47–62.
39. **Reddy B., K., Kumar, K., Krishna, S., Pingali P, Varma, V. (2010).** Linking Named Entities to a Structured Knowledge Base. *International Journal of Computational Linguistics and Applications*, Vol. 1, No. 1–2, pp. 121–136.

40. **Vor Der Brück, T. (2010).** Hypernymy Extraction Using a Semantic Network Representation. *International Journal of Computational Linguistics and Applications*, Vol. 1, No. 1–2, pp. 105–119.
41. **Homola, P., Kuboň, V. (2010).** Exploiting Charts in the MT Between Related Languages. *International Journal of Computational Linguistics and Applications*, Vol. 1, No. 1–2, pp. 185–199.
42. **Witten, I.H. & Milne, D. (2008).** An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, AAAI Press, USA, pp. 25–30.
43. **Zesch, T., Gurevych, I., Mühlhäuse, M. (2014).** Java-based Wikipedia API. https://www.ukp.tu-darmstadt.de/ukp-home.
44. **Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., Howard, N. (2013).** Common sense knowledge based personality recognition from text. *Advances in Soft Computing and Its Applications. Proceedings of the 12th Mexican International Conference on Artificial Intelligence, MICAI 2013, Mexico, Part II*, pp. 484–496
45. **Sidorov, G., Kobozeva, I., Zimmerling, A., Chanona-Hernández, L., Kolesnikova, O. (2014).** Computational Model of Dialog Based on Rules Applied to a Robotic Mobile Guide. *Polibits*, Vol. 50, pp. 35–42.

Claudia Perez-Martinez received her Master degree in Computer Science from the Mexico State Autonomous University (Universidad Autónoma de Estado de México). She received her Master degree in Educational Technology from the Monterrey Institute of Technology (Instituto Tecnológico de Monterrey). She is pursuing a PhD at the Institute of Engineering, Baja California Autonomous University (Instituto de Ingeniería, Universidad Autónoma de Baja California).

Gabriel López Morteo received his PhD degree in Computer Science from the Ensenada Center of Scientific Research and Higher Education (Centro de Investigación Científica y de Educación Superior de Ensenada). He has academic experience in computer assisted education, collaborative systems, process engineering, and Internet application development. He is a full time researcher at the Department of Computing and Informatics of Institute of Engineering of Baja California Autonomous University (Departamento de Computación e Informática del Instituto de Ingeniería de la Universidad Autónoma de Baja California). He carries out research in the field of computer science and applied informatics, oriented to electronic learning environments, interactive learning objects, and the behavior of learning virtual communities. He is coordinator of the Physics Engineering Area of the Baja California Institute of Engineering in Information Technologies (Instituto de Ingeniería en Tecnologías de la Información de Baja California, A.C.).

Magally Martinez Reyes received a PhD degree in Mathematics Education from Centro de Investigación y Estudios Avanzados of Instituto Politécnico Nacional. She has a Bachelor degree and a Master degree in Mathematics by Universidad Nacional Autónoma de México. She is a member of the Sistema Nacional de Investigadores. Her research interest is development of educational software for learning mathematics. She is currently a full professor, with the PROMEP profile, of the Universidad Autónoma del Estado de México Valle de Chalco, associated with the program of Computer engineering and MSc in Computer Science. She has published several specialized articles in national and international journals on mathematics and computing.

Alexander Gelbukh received his MSc degree in Mathematics from the Moscow State Lomonosov University, Russia, and a PhD degree in Computer Science from VINITI, Russia. He is currently a Research Professor and Head of the Natural Language Processing Laboratory of the Center for Computing Research (Centro de Investigación en Computación, CIC) of the Instituto Politécnico Nacional (IPN), Mexico, invited professor of the Universidad Nacional, Colombia, and visiting professor of the Moscow Sholokhov University for the Humanities. He is former President of the Mexican Society of Artificial Intelligence (SMIA). He is a Member of the Mexican Academy of Sciences and National Researcher of Mexico (SNI) at Excellence level 2. He is author or coauthor of more than 500 research publications in natural language processing and artificial intelligence.

Article received on 10/12/2014; accepted 15/05/2015. Corresponding author is Claudia Perez-Martinez.