ISSN 2007-9737

Construction of Conditional Probability Tables of Bayesian Networks using Ontologies and Wikipedia

Alan Ramírez Noriega[1], Reyes Juárez Ramírez[2], Juan J. Tapia[3], Víctor H. Castillo[4], Samantha Jiménez[3]

[1] Universidad Autónoma de Sinaloa,
Facultad de Ingeniería Mochis, Sinaloa,
Mexico

[2] Universidad Autónoma de Baja California,
Facultad de Ciencias Químicas e Ingeniería,
Mexico

[3] Instituto Politécnico Nacional,
Centro de Investigación y Desarrollo de Tecnología Digital,
Mexico

[4] Universidad de Colima,
Facultad de Ingeniería Mecánica y Electrónica,
Mexico

alandramireznoriega@uas.edu.mx, jtapiaa@ipn.mx, victorc@ucol.mx,
{samantha.jimenez, reyesjua}@uabc.edu.mx

Abstract. Building Bayesian Networks automatically serves to reduce time and effort by defining variables and the quantitative relation between variables. However, the quantitative part is the most complicated to solve because of it is statistical information. This research proposes a method to construct the quantitative part of a Bayesian Network based on text mining and ontologies for Intelligent Tutoring Systems. The network structure is built based on the variables and relations of an ontology. Conditional Probability Tables (CPT) are created from Wikipedia information. The constructed CPT reach a correlation of 0.895 against the experts' opinion. This correlation is good due to the subjectivity in the evaluations. We conclude that using the text mining in Wikipedia and ontologies, it is possible to construct CPT that adequately represents knowledge in an educative environment.

Keywords. Bayesian network, conditional probability table, ontology, text mining, wikipedia, intelligent tutoring system.

1 Introduction

An Intelligent Tutoring System (ITS) can be defined as a software system that uses artificial intelligence techniques to interact with students to teach them in the same way that a teacher does [37]. A significant problem in ITSs development is the assessment of student knowledge. ITSs must be able to determine accurately and quickly the students' learning level to decide what is important to teach them. Authors have proposed probability theory for handling the uncertainty in diagnosing student knowledge [37]. One of the probability-based techniques most used in ITSs is the Bayesian Network [7].

The Bayesian network (BN) adequately represents knowledge because it involves the use of reasoning in an environment of uncertainty [26]. A tutor needs to determine a student's level of knowledge, however, the teacher's assessment may have inaccuracies because of the difficulty

doi: 10.13053/CyS-23-4-2705

in measuring knowledge [25]. A BN uses probability theory as a framework to manage uncertainty. This utilizes a graphical description of a probabilistic distribution that efficiently combines the propagation of probabilities within a rigorous formalism [37].

The BN's construction, based on judgment of experts, is a complex task necessitating an interaction with the expert and the need to translate his knowledge into a structure of nodes, relations, and numerical values [42, 40]. While building a BN is completed in seconds, manual construction involves more time and more work [39, 42].

While BNs can be built manually or automatically, the process involves identifying variables, variables states, establishing relations between variables, and constructing the Conditional Probability Tables (CPT). However, the most complicated part of BN's generation is building the CPT because it requires statistical information representing the relation between variables [9, 21]. For this reason, constructing the CPT automatically would expedite completion.

Data mining is a way to build BNs automatically. This approach is based on the analysis of information to identify variables, relations and the CPT. However, this process has some problems [21]:

— Inability to cope with the missing data.

— Possibility in making wrong assumptions about the input or output data of the structure.

— It does not integrate structured knowledge sources of the current world resulting in inaccuracies.

— If there is a significant number of nodes in the networks, the learning algorithm search space becomes so large that it often leads to a decrease in learning.

Another way to automate the information is based on ontologies. A similar structure to the BN [36, 42, 6]. Ontologies represent the information semantically through nodes and links. BNs can be built based on ontologies, whereby a domain of knowledge to make probabilistic inferences can be represented. The ontological reasoning can then be applied to solve semantic problems. The qualitative part of the BN is easy to construct because it relies on a structure with a larger range of knowledge domains already represented by nodes and relations.

On the other hand, the main problem is the development of the quantitative part of a BN because ontologies do not handle probabilistic information necessary to BN inference.

There are efforts to generate BNs based on ontologies. They have been applied to the medical field as well as to other areas. These attempts can be classified into three types:

1. Manual methods: It focuses on the human execution of the processes of the creation of the BN by an analysis of knowledge formalized in ontologies [30, 32].

2. Semiautomatic methods: Automatically generate some part of the BN of a given ontology; These methods require user interaction to assist in the remaining parts [14, 9].

3. Automatic methods: These methods work with a particular form of input of the ontology and generate a simplified predefined structure from this input [43, 5].

CPT have been built in different ways, in [42, 44] is proposed an automatic construction of BNs extending the ontologies. However, manual labor is present when the ontology is built. Andrea [1] proposed ontology instance-based construction of CTP. However, this method could construct CPT that do not reflect the reality of the educational domain. Significantly, Andrea's approach would be useful only when the ontologies had sufficient instances, and the majority of available ontologies do not have this.

Bucci [5] automates the generation process of CPT with a BN template using a database of diseases and symptoms, however, the original domain is reduced, losing important information. The techniques of these works are not designed for the educational environment, and for this reason a new proposal is necessary to face this new domain.

The objective of this work is to construct the CPT by means of the relation between concepts of an ontology and Wikipedia information. Results of the method are focused on educational environments, therefore, results are compared with experts from the educational domain through the Pearson's test. The objective of this work is important because it helps to automate the construction of the ITS domain module.

This study is organized as follows: Section 2 establishes related work. Section 3 describes the theory that supports our research. Section 4 shows the used approach. Section 5 explains how the CPT are constructed and the relationship between variables is obtained. Section 6 describes the case study with an experiment and its results. Section 7 presents the discussion. The last sections describe conclusions and references.

2 Related Work

This section describes some works that have automated the process of BNs construction based on ontologies.

Devitt [9] proposed an approach to harness the knowledge and inference capabilities inherent in an ontology model to automate the building of BNs to represent a domain of interest accurately. The method was implemented in the context of an adaptive, self-configuring network management system in the telecommunications area. Although Devitt adequately constructs the variables and their relations, the proposed method is considered semi-automatic because it does not fully generate CPT. Human effort is still necessary.

Ding [12] developed BayesOWL, this study proposed a modification to the IPFP (iterative proportional fitting procedure) algorithm to construct CPT. However, this method does not analyze any kind of knowledge domain information, so the method constructs CPT that do not reflect real aspects of the educational domain.

Fenz [14] used the ontology to provide the necessary knowledge about relevant influence factors, their relationships, their weights, and the scale which represents potential states of the identified influence factors. The developed method enables the semiautomatic generation and alternation of BNs. The limitations of the proposed method are the same as Devitt, in that CPT are not calculated and human effort is still necessary.

Andrea et al. [1] proposed to construct BN automatically. They argued that the information to make the automatic process is found on the same ontologies. The semantic information of the concepts can provide the relation between the nodes. The frequency of data of instances was used to construct the CPT.

This method may be adequate but requires that each ontology has sufficient instances to create the CPT, and the final CPT may not represent reality.

Yan [42] proposed a framework that provides a set of structural translation rules to map an OWL (Web Ontology Language) taxonomy into a BN directed acyclic graph. The framework makes an extension to the OWL standard to support the quantitative part to make an automatic conversion to the BN. Although the human effort disappears in the process of conversion, there is still human effort needed when quantitative information is added to the extended ontology.

These works have advanced in different areas, however, the educational area has not progressed. Related work can correctly define variables, states, and relations, however, they do not adequately face the construction of the CPT to the needs of this study. The main difference of our work with related work is to construct CPT automatically for an educational domain in a similar way as experts construct it. Our work impacts efforts such as [15]. These researchers used an ontology and a BN to evaluate knowledge in an adaptive educational system. However, the construction of both structures was manual; our method would support the building of CPT of the BN used in this article.

3 Fundamentals

3.1 Bayesian Network Theory

BNs handle a variety of definitions and concepts [29, 26]:

1. A node n is a random variable that can have multiple states.

2. An arc is a connection between two nodes and it represents the dependence between two variables. An arc is defined by an ordered pair of nodes (n_b, n_c).

3. The node n_b is a parent of the node n_c, if there is an arc (n_b, n_c) between the two nodes.

4. The node n_c is the son of node n_b, if there is an arc (n_b, n_c) between the two nodes.

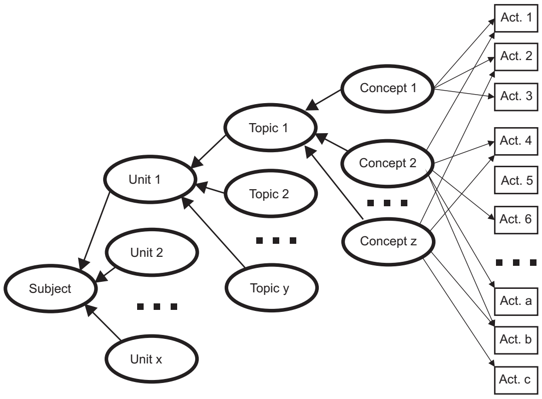

Fig. 1. Bayesian Network Structure for ITS

The states of a variable must meet with two properties:

1. To be mutually exclusive, i.e. a node can only be found in one of the states in a given time.

2. To be an exhaustive set, i.e. a node can have no value outside of the set.

The BN model [29] is represented as a 3-tuple in $BN = \{N, DB, P\}$. The variables N and DB represent the qualitative part of the BN. Elements in N represent random variables as $N = \{n_1, .., n_a\}$, a represents the total of elements in N. These variables are related to each other, these relations are represented as $DB = \{(n_b, n_c), .., (n_x, n_y)\}$. Where, the first element of the ordered pair represents the parent node (relation beginning), and the second element represents the child node (relation end).

The variable P represents the quantitative part of the BN, this variable works with Conditional Probability Tables (CPT). CPT shows the probability that an event will occur based on the combination of the nodes and the value of their states. The uncertainty of the causal relation is represented by the conditional probability table $P(n_q|\pi_q)$ associated with each node n_q, where π_q is the parent set of N. Under a conditional independence assumption, the graphic structure of BN allows an unambiguous representation of interdependency between variables. It leads to one of the most important feature of the BN: the joint probability distribution of $N = (n_1, ..., n_r)$ can be factored out as a product of the CPT in the network: $P(n_1, ..., n_r) = \prod_{q=1}^{r} P(n_q|\pi_q)$ [11].

3.2 Bayesian Network in ITSs

In recent years, ITSs has focused on three main approaches to student modeling [7] namely BNs, fuzzy logic, and ontologies. Other authors [38, 40, 41] concluded that BNs is the most commonly used approach in ITSs because of the advantages offered by their theory. Danaparamita [8] made an evaluation of two student models; fuzzy logic and BNs, and concluded there was a slight improvement of BNs over the fuzzy logic in predicting the learning needs of the student.

In this paper, the structure used to represent the knowledge is based on the Millán's model [27]. This model helps to determine the cognitive degree of the students through their interaction with the BN. Millán's model uses the next element to conform the BN:

— **Variables for measuring students attained knowledge:** This knowledge is broken down into pieces of information. Different levels of granularity can be obtained. Figure 1 shows how the knowledge of a subject is divided into units. The units are divided into topics and finally, the topics are divided into concepts.

— **Variables for gathering evidence:** These variables interact with the student to obtain evidence for inferring knowledge. These variables are represented at the right end of Figure 1. These can be tests, exercises, activities, lessons or the tracking of student interaction with the system [35].

— **Links between variables:** Dominate knowledge has a causal influence on learning preceding and immediate levels in the related granularity hierarchy. With regard to links between the nodes and the questions, it is considered that knowledge has a causal influence on correctly answering the variables for gathering evidence.

The development process of the MillÃ¡n's model is divided into six phases [33]:

1. Definition of a knowledge domain: the work area is chosen based on the needs of the problem.

2. Development of a hierarchical scale of knowledge: Classification of the knowledge in different levels.

3. Construction of the Bayesian Network: Create nodes and establish dependence relations between them.

4. Design of the CPT: Assign probabilities to nodes, according to relationships with parents.

5. Design of activities to gather evidence: Create a bank of activities and assign relations with the concepts contained in nodes.

6. Creation of the CPT for the activities: Assign probabilities to the question nodes according to their parents.

The phases 1, 2, 3, and 5 create the network structure (structural or qualitative section), and the phases 4 and 6 calculate the estimated probability values for each node (parametric or quantitative part). The phases 4 and 6 could be combined into one. However, they were divided into two phases to allow better organization and clarity. This work focuses on the phase 4 of the methodology.

3.3 Ontology

In computer science area, Gruber [17] stipulated that an ontology is an explicit specification of a conceptualization. The conceptualization refers to an abstract model of some phenomenon that has identified its relevant concepts. The word explicit means that all concepts employed and the constraints on their use are explicitly defined in the model [3].

In the simplest case, an ontology describes a hierarchy of concepts (i.e. classes) related by taxonomic relationships (is-a, part-of). In more sophisticated cases, an ontology defines domain classes, properties (or attributes) for each class, class instances (or individuals), and also the relationships that hold between class instances. It is also possible to add some logical axioms to constrain concept interpretation and express complex relationships between concepts [3].

4 Calculation of the Relation between Concepts

Wikipedia has a vast knowledge in the structure and articles, this encyclopedia is recognized as an enabling knowledge base for a variety of intelligent systems [18]. Wikipedia articles are related to other concepts representing other articles. Knowledge is formed through relations.

Therefore, if the concept x is based on the concept y to be explained, a relation of dependence exists between these concepts. Measuring the dependence between concepts, concept y with concept x, indicates the importance of the first concept in understanding the second one. If we have a quantitative way to measure this relation, we could establish weights to concepts that represent their importance.

The semantic relation of concepts depends on multiple factors and the correct combination of them [24]. A combination of factors was considered to measure the importance of concepts. Our proposal considers three main factors:

- **Frequencies:** The basic idea of this factor is that keywords in a text are those terms that are most repeated. Therefore, a concept is important if it is repeated several times in a document [4, 24]. This factor is useful in various techniques based on the ID-TFD algorithm [47]. This aspect is called factor fr for our purposes. This factor involves the following cleaning process:
  - Tokenize sentences: The tokenization breaks down the sentences into a set of words [20] called tokens. It is the minimal unit represented.
  - Delete stop-words: The stop-words list includes the most frequently occurring words in a text [16]. Stop Words are words which are not significant to our method (e.g. a, the, of, etc.). These words are eliminated from the original text.
  - Apply stemming: The stemming technique utilizes the root form of a word. The main objective is to assign equal importance to words having the same root. Thus, words in their different forms are considered to be the same [20]. The most common method used to do this process is Porter's algorithm [31].
- **Hops:** This measure is calculated based on Wikipedia structure. In a hierarchical tree structure, the concepts that involve fewer jumps between the nodes are the most closely related [23, 24, 46]. The relation between two concepts, A and B, is wide if the content of concept A has a direct link to concept B. The relation between two concepts will decrease according to the number of hops that exist between the two concepts. This aspect is called factor s for our purposes.
- **Similarity of concepts:** If two terms are related, they will share words within their meaning [46]. This is based on the idea of obtaining the frequency of words of two texts; there will be more relation as they share more common words between them. This aspect is called factor j for our purposes.

Variables and equations have been defined based on the aforementioned factors to measure the similarity between concepts in an educational environment.

4.1 Variables in the Ontology

The ontology is used to represent the structure of the network. However, this article does not focus on building the qualitative part of the network. It focuses on building the quantitative part. Nonetheless, to construct the quantitative part, it is important to define the structure to represent the mathematical model.

The variables representing the problem and how their relationship are shown Figure 2. The root class of the ontology is represented by the variable r. The variable h represents a class that is composed of other classes. Finally, the variable p is class that form another class.

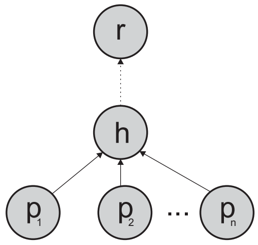

Fig. 2. Relations between variables in the ontology

The set P represents the parents that influence a child ($P = \{p_1, p_2, ..., p_n\}$). The set X includes the set P and other two elements root and child, the set X is formed as $X = \{r, h, \{P\}\}$. Where x_y represents an element of X and P represents a subset of X.

The elements of the set X are related in the following way (see Figure 2): (1) The element $root$ is influenced by the element $child$ (h, r), in this case r is a child and h acts as the parent. (2) The element $child$ is influenced by the element p_i (p_i, h), in this case h is a child and p_i acts as a parent.

The objective is to find the numerical relation between the child and the parents that influence it. Each element of X corresponds to a Wikipedia page that describes a concept. The element x_y

can be treated by the title (concept representing the page) or the content (the text of the article including links, references, and other attributes). When we refer to the concept, it will be denoted as $ct(x_y)$ and for the content as $cd(x_y)$. The same applies to the elements of P, for the concept as $ct(p_i)$ and for the content as $cd(p_i)$.

4.2 Quantitative Relation between Variables

The previous foundation was considered in this section. The equation $rel(p_i, h) = \frac{1}{3}(fr_{norm}(p_i) + j_{norm}(p_i) + s_{norm}(p_i))$ was defined to obtain the numerical relationship between the variables.

This equation considers three factors: the factor fr_{norm} obtains the normalized relation measure considering the frequencies, the factor j_{norm} obtains the normalized measure of relation taking into account the coefficient of Jaccard, and the factor s_{norm} obtains the normalized measure of relation considering the hierarchical scale of Wikipedia. Each one of these factors is explained below.

4.3 Quantitative Relation Considering the Frequencies (Factor fr)

This factor looks for a value that represents the parent p_i through its frequency. It considers doing a frequency search on all directly related nodes; sibling nodes, the child node, and the root node.

The equation $fr(p_i) = \sum_{y=1}^{m}(fr(ct(x_y), cp(p_i)))$ was used to find this numeric relation, where $fr(ct(x_y), cp(p_i))$ is the frequency search of the term p_i in the domain x_y and m is the total elements of the set X. The previous process will be repeated for each element of X, obtaining a sum of each result. $ct(x_y)$ must be different from $ct(p_i)$ because P is a subset of X.

The results are normalized through equation $fr_{norm}(p_i) = fr(p_i)/\sum_{i=1}^{n} fr(p_i)$.

4.4 Quantitative Relation Considering Coincidence of Concepts (Factor j)

This factor establishes a quantitative relation between two sets, c_1 and c_2. The words found in the article p_i represent c_1 and the words found in the article h represent c_2. The relation is obtained through the coefficient of Jaccard [22], this coefficient is represented in equation $j(c_1, c_2) = |c_1 \cap c_2| \,/\, |c_1 \cup c_2|$.

The equation represents the intersection of both sets divided by the union of both sets. The results are normalized through equation $j_{norm}(p_i) = j(p_i)/\sum_{i=1}^{n} j(p_i)$.

The coefficient of Dice [10] was also used. However, there were no differences when normalizing the values to assign the weights to the relations. The most common coefficient was used to find the numeric relation [19].

4.5 Quantitative Relation Considering Hops Between Concepts (Factor s)

This factor refers to the number of hops between two concepts, c_1 and c_2. That is, the number of links that must be followed to go from one Wikipedia page to another. The concept c_1 is represented by $cp(p_i)$ and the concept c_2 is represented by $cp(h)$. The measure relation is calculated through the equation $s(c_1, c_2) = (w - d(c_1, c_2)) * t$.

The equation is supported by the equations $x = round(x_{mean})$, $w = x + 1$, and $t = 1/x$. Where x represents the rounded average of hops to go from one page to another in Wikipedia [28][2]. The variable w represents a hops limit. Any number of hops from one page to another greater than w will be considered a null relation.

The variable t represents a factor applied to the hops to learn the final value between 0 and 1. The results are normalized through equation $s_{norm}(p_i) = s(p_i)/\sum_{i=1}^{n} s(p_i)$.

doi: 10.13053/CyS-23-4-2705

5 Construction of Conditional Probability Table

This section shows how the CPT are constructed with the weight of relations.

The conditional probability table of each node is constructed by the $Build\_CPT$ function in the algorithm 1. The function receives as input an array of elements $weights[]$ and it constructs a support matrix ($matrix$). The length of the input array helps to calculate the rows of the matrix. The number of columns is obtained through the operation 2^{rows}.

The design of the BN contemplates two states or scales of the nodes; $present$ or $absent$. The $present$ part is calculated because the $absent$ one is obtained by the complement of the state $present$ (considering 1 as the maximum value). The probability of each possibility must be calculated to create the CPT. Thus, a node can be known or unknown. If the node is known, it is represented by the weight of its relation given the array $weights$, if not it is represented by 0. The total values to be obtained are 2^{rows}. Where $rows$ represents the number of nodes that influence another. Thus, if we have 2 nodes that influence another, the total value of the state $present$ will be $2^2 = 4$, plus 4 values for the state $absent$.

Code from line 7 to line 28 in algorithm 1 builds the support matrix to find the probability for each combination given by the parents and their states. The completed support matrix constructs the conditional probability table (From line 30 to 40). This is done by adding each column of the support matrix, thus obtaining the value of the state $present$ (line 36) as well as the $absent$ value (line 37). The built conditional probability table is the result returned by the function (line 41).

6 Experiment and Results

A case study is designed for testing the correlation of the method and experts. Firstly, We selected an ontology of the Object Orientation (OO) domain from the article [34]. This serves as a basis for obtaining the variables and relations for the BN. A questionnaire survey was made to teachers who teach Object Oriented Programming (OOP).

Algorithm 1 Process for building Conditional Probability Table

```
1:  function BUILD_CPT($weights[]$)
2:      $rows \leftarrow lenght(weights)$
3:      $columns \leftarrow 2^{rows}$
4:      initialize $matrix[rows][columns]$ with zeros
5:      initialize $CPT[2][columns]$ with zeros
6:      $cont \leftarrow 0; x \leftarrow 0; flag \leftarrow true$
7:      while $x < rows$ do
8:          $jumps \leftarrow (2^{rows-x})/2$
9:          $y \leftarrow 0$
10:         while $y < columns$ do
11:             if $flag$ then
12:                 $matrix[x][y] \leftarrow weights[x]$
13:             else
14:                 $matrix[x][y] \leftarrow 0$
15:             end if
16:             $cont \leftarrow cont + 1$
17:             if $cont >= jumps$ then
18:                 if $flag$ then
19:                     $flag \leftarrow false$
20:                 else
21:                     $flag \leftarrow true$
22:                 end if
23:                 $cont \leftarrow 0$
24:             end if
25:             $y \leftarrow y + 1$
26:         end while
27:         $x \leftarrow x + 1$
28:     end while
29:     $sum \leftarrow 0; y \leftarrow 0$
30:     while $y < columns$ do
31:         $x \leftarrow 0$
32:         while $x < rows$ do
33:             $sum \leftarrow sum + matrix[x][y]$
34:             $x \leftarrow x + 1$
35:         end while
36:         $CPT[present][y] \leftarrow sum$
37:         $CPT[absent][y] \leftarrow 1 - sum$
38:         $sum \leftarrow 0$
39:         $y \leftarrow y + 1$
40:     end while
41:     return $CPT$
42: end function
```

The results of the survey helped to quantitatively determine the association between one concept to learn another. Then, the method evaluated the same relations between two concepts as the experts did. Finally, a correlation test between the obtained data was made to measure the accuracy of the proposed method.

6.1 Ontology for Object Orientation Domain

For this study, the ontology from [34] was reduced into classes and relations (see Figure 3), we considered only those classes represented by an article in Wikipedia. The developed content of the concept is necessary for its mining. The domain modeled in the ontology is the OO, this is a field in computer science. The OO is a programming paradigm organized around objects and data. It is an important topic in programming-related careers.

6.2 Experiment Description

The experiment was designed and executed to verify that the value of the relations used to construct the CPT are similar to the expert opinions for this domain. Our method for building BN for educational environments is novel and there are no datasets with which to compare results.

The opinion of experts is the best way to validate these types of structures that try to represent an often subjective domain. Fenz [13] validated his results in this manner utilizing experts for their evaluation. This study confirmed the importance of experts and established a statistical test to check correlation.

6.2.1 Generation of Weights by Experts

The goal of this segment of the experiment is to determine the importance of a particular concept in learning another general concept. The instrument is intended for teachers who teach the OOP course. The ontology represents a section of the knowledge of this course. The type of sampling used is non-probabilistic because the representation of the population is not required in this experiment. The sample considered professors who teach the course at universities of Mexico.

The survey determines to what extent a concept is important in understanding another. For example: (1) To what extent do you believe that the concept $abstraction$ is important for understanding the concept $object - orientation$? (relation $abstraction\_object - orientation$). (2) To what extent do you believe that the concept $constructor$ is important for understanding the concept $class$? (relation $constructor$-$class$).

Each ontology relation established a question. The scale evaluated values from one to seven, where one was considered as least important and seven as the most important. The process of the experiment consisted of the following steps:

1. A list of universities that teach OO was made to identify candidate teachers to answer the questionnaire.
2. The instrument was created based on ontology.
3. An on-line version of the survey was created using Google Forms[1].
4. Teachers answered the questionnaire over Internet. No time limit was considered, and personal information was not required from the participants.
5. (5) Finally, the obtained data was analyzed and interpreted.

6.2.2 Generation of Weights by the Method

The relations obtained from the ontology were the input of the method. The same relations as the experts were evaluated to determine the weight of each class.

The proposed method was implemented in the Java programming language with the support of the JWPL (Java Wikipedia Library) programming interface. JWPL Provides access to all Wikipedia information in different languages in a structured way. This includes a MediaWiki tag analyzer for in-depth page content analysis. JWPL offers methods of accessing properties such as links, templates, categories, text and other properties [45].

6.3 Results

The results of the experts and the method are displayed in Table 1. The relations are described in the column with the same name, an identifier (column Id) was added for later references. The opinion of the experts with non-normalized values is shown in the column $Value$. The column $Weight$ ($experts$) displays the relation weight taking into account the equation $weight(P, c_i) = c_i / \sum_{i=1}^{m} c_i$.

[1] https://www.google.com/forms/about/

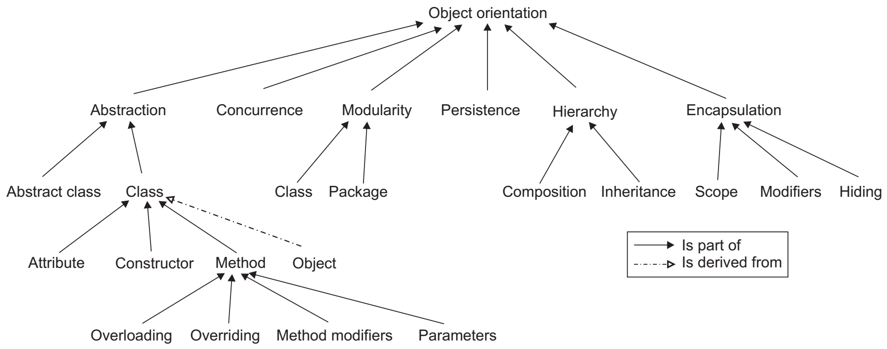

Fig. 3. Ontology part of the Object Orientation

Here c_i is the i element value of the set C $(C = \{c_1, ..., c_m\})$. c refers to concepts that compose other concept, where all the elements of C compose the same main concept. In other words, The variables $composition$ and $inheritance$ share the same child ($hierarchy$).

Table 1. Results of the experts and the method

		Experts		Algorithm
Id	Relation	Value	Weight	Weight
1	OO_abstraction	6.80	0.19	0.22
2	OO_concurrency	4.70	0.13	0.12
3	OO_encapsulation	6.45	0.18	0.13
4	OO_hierarchy	6.50	0.19	0.07
5	OO_modularity	5.90	0.17	0.08
6	OO_persistence	4.60	0.13	0.04
7	abstraction_abstract-class	6.40	0.48	0.36
8	abstraction_class	6.95	0.52	0.64
9	modularity_class	6.55	0.51	0.61
10	modularity_package	6.35	0.49	0.39
11	hierarchy_composition	6.05	0.47	0.44
12	hierarchy_inheritance	6.85	0.53	0.56
13	encapsulation_scope	6.25	0.33	0.24
14	encapsulation_hiding	6.40	0.34	0.43
15	encapsulation_modifiers	6.25	0.33	0.33
16	class_attribute	6.90	0.25	0.14
17	class_constructor	6.75	0.25	0.20
18	class_method	6.90	0.25	0.29
19	class_object	6.70	0.25	0.38
20	method_overloading	6.30	0.25	0.23
21	method_overriding	6.05	0.24	0.26
22	method_method-modifiers	6.25	0.25	0.22
23	method_parameters	6.80	0.27	0.30

6.4 Construction of a Conditional Probability Table

Therefore, the maximum value is divided between these two variables according to the proportion of the value of the expert.

The values returned by the method appear in the column $Weight$ ($Algorithm$). These values do not require any processing; the method already shows them normalized.

A visual interpretation is represented in Figure 4. This chart displays the results of the Table 1. The x-axis represents the relations considering its identifier. The y-axis represents the weight obtained by each relation. The solid line represents the normalized results of the experts. The dotted line represents the results of the method.

This part shows how a conditional probability table is constructed based on experts knowledge. The variable $encapsulation$ was considered to the example. The algorithm 1 and its weights shown in the Table 1 are considered to this process. The variable $encapsulation$ is formed by three variables; $scope$, $hiding$, and $modifiers$. So, the created relations are $encapsulation\_scope$, $encapsulation\_hiding$, and $encapsulation\_modifiers$.

The variable $encapsulation$ has three parents, therefore there are 3 relations with their weights. This represents the input variable $weight$ in the algorithm 1. The value for state $present$ is 8 (2^3), plus 8 values for state $absent$. The support matrix ($matrix$) represented in the algorithm 1 is shown in the Table 2.

The values 0.24, 0.43, and 0.33 in Table 2 are taken from Table 1. These values can be appreciated when one variable is $present$ but the other two are $absent$. For instance, when $scope$ and $hiding$ variables are $absent$ but $modifiers$ variable

Table 2. Conditional Probability Table to encapsulation variable

scope	**presente**				**absent**			
	0.24	0.24	0.24	0.24	0.0	0.0	0.0	0.0
hiding	**present**		**absent**		**present**		**absent**	
	0.43	0.43	0.0	0.0	0.43	0.43	0.0	0.0
modifiers	**p**	**a**	**p**	**a**	**p**	**a**	**p**	**a**
	0.33	0.0	0.33	0.0	0.33	0.0	0.33	0.0
present ($\sum$)	1.0	0.67	0.57	0.24	0.76	0.43	0.33	0.0
absent ($1 - \sum$)	0.0	0.33	0.43	0.76	0.24	0.57	0.67	1.0

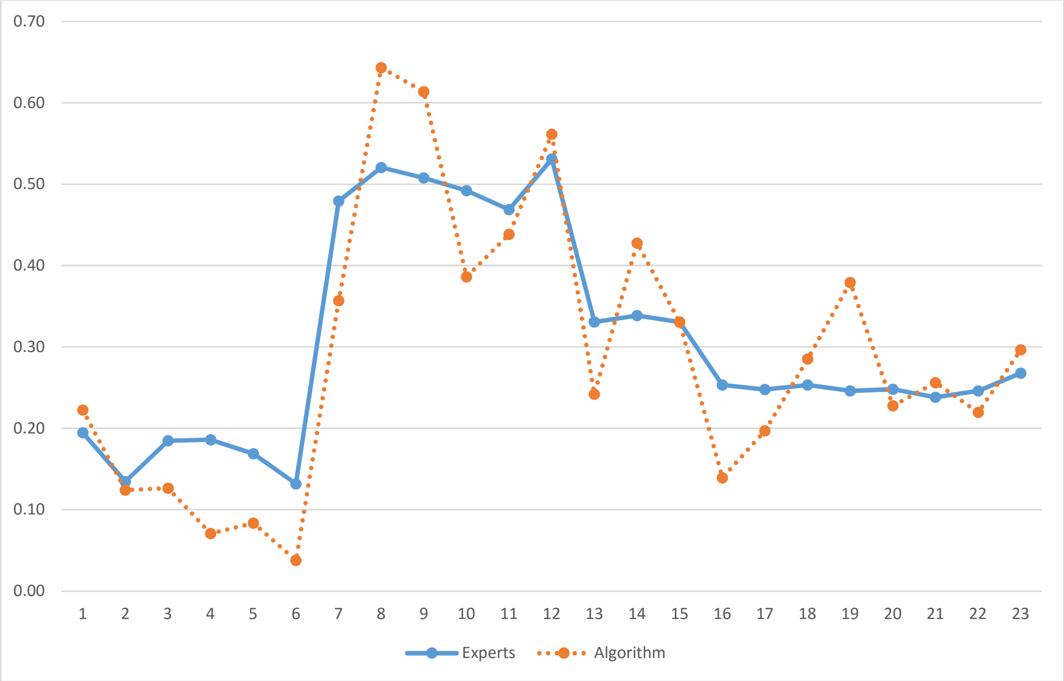

Fig. 4. Comparison between the experts and the method

is $present$, then $encapsulation$ value is 0.33 ($0+0+0.33$), equal than the relation 15 in Table 1.

The values to obtain the state $present$ of the CPT are obtained through a summation of columns as shown in Table 2. The value $absent$ is obtained through the complement of the value $present$. A conditional probability table is constructed for each network variable following the structure of the ontology.

6.5 Hypothesis Testing

The test is based on the hypothesis that the combination of the three factors proposed will show results close to the estimated values by the group of teachers. This is verified by a correlation test where it is expected to obtain a correlation greater than 0.8 with a confidence level of 95%.

Two variables were considered, $method$ and $experts$. The variable $method$ is defined as the degree of numerical relation between the classes generated through our method. The variable $experts$ is defined as the degree of numerical relation between classes obtained through the group of university professors.

The $method$ and $experts$ approved the Levene and the Kolmogorov-Smirnov test with a confidence level greater than 95%. Therefore, we use the Pearson test. The hypothesis are:

— H_0: the variables $method$ and $experts$ have no correlation.

— H_1: the variables $method$ and $experts$ have correlation.

The Pearson test showed a correlation between variables of 0.895. The test yielded a $P-value$ below the significance level ($0.001 < 0.05$). Therefore, we accept H_1; the variables $method$ and $experts$ have correlation.

7 Discussion

A method faces the difficult task of accurately determining values in the same way as experts. The decision of the experts is based on subjectivity and often does not show agreement in their opinions. For this reason, their opinions were averaged to obtain a consensus in general.

A positive correlation of 0.895 between the results of the experts and the method was obtained applied the Pearson test. This represents a high positive correlation in the range [-1,1]. A reliability greater than 99% confidence was obtained with the test. A greater correlation 0.8 is acceptable considering the subjectivity that appears in the estimates of the relations.

The used method is justified because of the implicit knowledge within Wikipedia. The articles have knowledge represented by its textual information. The hierarchical structure provides information support. The proper combination of these factors provides a measure of the relation between classes. This knowledge is generated and evaluated by people with experience in the domain who reached a consensus on the content of the articles. Therefore, a degree of relation can be established between what an expert thinks (experts surveyed) and what others write (Wikipedia information).

The obtained results indicate that the quantitative part of a BN can be constructed based on ontologies and Wikipedia information. This structure is useful for educational purposes with similar results to the experts.

The most outstanding contributions provided by the method proposed in this paper are:

1. The reduction of effort and time in the BN creation is achieved with the automatic identification of the weights of the relations.

2. The automatic construction of one of the four models of an ITS; Domain Model.

3. The automatic construction of the phase four of the Millán model. It is the most complicated phase of the process. It also serves as a basis for automating the phase six.

4. This method helps to increase the teaching domains not yet represented in BNs. This gives the possibility of making inferences about new domains in ITS.

5. The same domain of knowledge in two different structures is obtained. This provides the possibility of doing reasoning on two levels; ontological and probabilistic.

There are some points to overcome such as:

1. The model only works with concepts that have an article with content in Wikipedia.

2. A usable BN for an ITS needs six phase. These activities create one phase.

3. The construction of the quantitative part of BN is contemplated in this work. However, the qualitative part is missing. It requires solving other problems related to structure (cycles, exponential complexity of relations, among others).

Regarding the first point. Wikipedia has a large number of articles from different areas. However, some very specialized topics are not yet defined. Few encyclopedias have more articles than Wikipedia. This makes it a complicated problem. And only a search on the Internet could overcome this type of disadvantage. The approach used would have to be modified.

In relation to points two and three, the most complicated part of the six phases to create BN is addressed in this article. The proposed method reduces the work to construct to BN, remaining the qualitative part to be constructed, which is simpler to be based on the ontology. A BN considering the two basic parts, quantitative and qualitative, is proposed in a next article.

8 Conclusions

This work presents a method to construct the quantitative part of a BN. The method applies text mining for finding the relation between the variables of a BN. The case study of this article considers an ontology of OO as the basis for testing the method and locate the weight of the relations.

This work established a relation between domain experts and the proposed method. The experts were teachers of the OOP who gave their opinion on the importance of concepts in learning others. The method consisted in mining the structure of Wikipedia and its articles based on the classes and relations of an ontology.

According to results, this study can affirm that the frequencies, hops, and coincidence between concepts allow obtaining measures between concepts similar to experts of the same domain. The Pearson's test showed an accuracy of 0.895 between results of the proposed method and results of experts, this correlation allows to build CPT with high degree of reliability to be used in ITSs.

Education has very theoretical phases such as information in Wikipedia articles. Therefore, this information can be used to obtain the weight of the relations between concepts in a BN. However, assuming we want to diagnose a disease or assign a bank credit, the Wikipedia information is inadequate because there is no type of knowledge for these kinds of problems. It is necessary to find other information sources to solve it.

The future work contemplates constructing the qualitative part of the BN to have a completely automated structure. In this way, based on an ontology and Wikipedia we can build a BN for a educational environment. We will also test in other domains to prove the effectiveness of our proposal.

Acknowledgements

We appreciate the support of Consejo Nacional de Ciencia y Tecnología (CONACYT) and Universidad Autónoma de Baja California for resources provided to develop this research.

References

1. **Andrea, B. & Franco, T. (2009).** Extending ontology queries with Bayesian network reasoning. *International Conference on Intelligent Engineering Systems, INES 2009*, pp. 165–170.
2. **Arola, K. & Wysocki, A. (2012).** *Composing Media Composing Embodiment*. Utah State University Press.
3. **Ben Messaoud, M., Leray, P., & Ben Amor, N. (2013).** Active learning of causal Bayesian networks using ontologies: A case study. *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.
4. **Bhowmik, R. (2008).** Keyword extraction from abstracts and titles. *Conference Proceedings - IEEE SOUTHEASTCON*, pp. 610–617.
5. **Bucci, G., Sandrucci, V., & Vicario, E. (2011).** Ontologies and Bayesian networks in medical diagnosis. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 1–8.
6. **Chang, Y. S., Hung, W. C., & Juang, T. Y. (2013).** Depression diagnosis based on ontologies and bayesian networks. *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pp. 3452–3457.
7. **Chrysafiadi, K. & Virvou, M. (2013).** Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, Vol. 40, No. 11, pp. 4715–4729.
8. **Danaparamita, M. & Gaol, F. L. (2014).** Comparing Student Model Accuracy with Bayesian Network and Fuzzy Logic in Predicting Student Knowledge Level. *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 9, No. 4, pp. 109–120.
9. **Devitt, A., Danev, B., & Matusikova, K. (2006).** Constructing Bayesian Networks Automatically using Ontologies. *Applied Ontology*, Vol. 1, No. 1.
10. **Dice, L. R. . (1945).** Measures of the Amount of Ecologic Association Between Species. *Ecology*, Vol. 26, No. 3, pp. 297–302.

11. **Ding, Z. & Peng, Y. (2004).** A probabilistic extension to ontology language OWL. *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, Vol. 00, No. C, pp. 1–10.

12. **Ding, Z., Peng, Y., & Pan, R. (2006).** BayesOWL: Uncertainty modeling in semantic web ontologies. *Studies in Fuzziness and Soft Computing*, Vol. 204, pp. 3–29.

13. **Fenz, S. (2012).** An ontology-based approach for constructing Bayesian networks. *Data & Knowledge Engineering*, Vol. 73, pp. 73–88.

14. **Fenz, S., Tjoa, a. M., & Hudec, M. (2009).** Ontology-based generation of bayesian networks. *Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2009*, , No. i, pp. 712–717.

15. **Ferreira, H. N. M., Brant-ribeiro, T., Ara, R. D., Dorc, F. A., & Cattelan, R. G. (2016).** An Automatic and Dynamic Student Modeling Approach for Adaptive and Intelligent Educational Systems using Ontologies and Bayesian Networks. *2016 IEEE 28th International Conference on Tools with Artificial Intelligence*.

16. **Fox, C. (1989).** A stop list for general text. *ACM SIGIR Forum*, Vol. 24, No. 1-2, pp. 19–21.

17. **Gruber, T. R. (1995).** Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, Vol. 43, No. 5-6, pp. 907–928.

18. **Hadj Taieb, M. A., Ben Aouicha, M., & Ben Hamadou, A. (2013).** Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems*, Vol. 50, pp. 260–278.

19. **Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., & Montmain, J. (2014).** A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of Biomedical Informatics*, Vol. 48, pp. 38–53.

20. **Hingu, D., Shah, D., & Udmale, S. S. (2015).** Automatic text summarization of Wikipedia articles. *Proceedings - International Conference on Communication, Information and Computing Technology, ICCICT 2015*, pp. 15–18.

21. **Hu, X.-X., Wang, H., & Wang, S. (2007).** Using Expert's Knowledge to Build Bayesian Networks. *International Conference on Computational Intelligence and Security Workshops (CISW 2007)*, pp. 220–223.

22. **Jaccard, P. (1912).** The Distribution of the Flora in the Alpine Zone. *New Phytologist*, Vol. 11, No. 2, pp. 37–50.

23. **Jindal, V., Bawa, S., & Batra, S. (2014).** A review of ranking approaches for semantic search on Web. *Information Processing and Management*, Vol. 50, No. 2, pp. 416–425.

24. **Li, Y., Bandar, Z. a., & McLean, D. (2003).** An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 4, pp. 871–882.

25. **Lima, S. & Stump, S. (2010).** How to obtain knowledge evidences in a student model based on Bayesian Network. *Information Systems and Technologies (CISTI), 2010 5th Iberian Conference on*, pp. 1–4.

26. **Liu, H., Zhou, S., Lam, W., & Guan, J. (2017).** A New Hybrid Method for Learning Bayesian Networks: Separation and Reunion. *Knowledge-Based Systems*, Vol. 121, pp. 185–197.

27. **Millán, E. & Pérez-De-La-Cruz, J. L. (2002).** A Bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adapted Interaction*, Vol. 12, pp. 281–330.

28. **Milne, D. & Witten, I. H. (2008).** Learning to link with Wikipedia. *Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*, pp. 509–518.

29. **Neapolitan, R. E. (1990).** *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. John Wiley & Sons, Inc., New York, NY, USA.

30. **Pilato, G., Augello, A., Missikoff, M., & Taglino, F. (2012).** Integration of ontologies and Bayesian networks for maritime situation awareness. *Proceedings - IEEE 6th International Conference on Semantic Computing, ICSC 2012*, pp. 170–177.

31. **Porter, M. F. (1980).** An algorithm for suffix stripping. *Program*, Vol. 40, No. 3, pp. 211–218.

32. **Pshenichny, C. a. (2014).** Knowledge engineering in volcanology: practical claims and general approach. *Journal of Volcanology and Geothermal Research*, Vol. 286, pp. 78–92.

33. **Ramírez-Noriega, A., Juárez-Ramírez, R., Huertas, C., & Martínez-Ramírez, Y. (2015).** A Methodology for building Bayesian Networks for Knowledge Representation in Intelligent Tutoring Systems. *Congreso Internacional de Investigación*

e Innovación en Ingeniería de Software 2015, San Luís Potosí, pp. 124–133.

34. **Ramírez-Noriega, A., Juárez-Ramírez, R., Jiménez, S., Martínez-Ramírez, Y., & Armenta, J. (2017).** *Building a Bayesian Network for Object Oriented Programming with Experts' Knowledge*. Springer International Publishing, Cham, pp. 267–276.

35. **Ramírez-Noriega, A., Juárez-Ramírez, R., & Martínez-Ramírez, Y. (2016).** Evaluation module based on Bayesian networks to Intelligent Tutoring Systems. *International Journal of Information Management*.

36. **Rodrigues, F. H., Bez, M. R., & Flores, C. D. (2013).** Generating Bayesian networks from medical ontologies. *2013 8th Computing Colombian Conference, 8CCC 2013*.

37. **Santhi, R., Priya, B., & Nandhini, J. (2013).** Review of intelligent tutoring systems using bayesian approach. *CoRR*, Vol. abs/1302.7.

38. **Satar, A. (2012).** Using of Intelligence Tutoring Systems For Knowledge Representation in Learning & Teaching Process. *Kufa for Mathematics and Computer*, Vol. 1, No. 5, pp. 1–13.

39. **Thirumuruganathan, S. & Huber, M. (2011).** Building Bayesian Network based expert systems from rules. *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3002–3008.

40. **Ting, C.-Y., Cheah, W.-N., & Ho, C. C. (2013).** Student Engagement Modeling Using Bayesian Networks. *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2939–2944.

41. **Victorio-Meza, H., Mejia-Lavalle, M., & Ortiz, G. R. (2014).** Advances on Knowledge Representation of Intelligent Tutoring Systems. *2014 International Conference on Mechatronics, Electronics and Automotive Engineering*, pp. 212–216.

42. **Yan, L. & Wei, C. H. (2013).** Development of a novel asset management system for power transformers based on Ontology. In *2013 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*. pp. 1–6.

43. **Yang, K.-a., Yang, H.-j., Yang, J.-d., & Kim, K.-h. (2005).** Bio-ontology Construction Using Object-oriented Paradigm 1. *Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05)*.

44. **Yang, Y. Y. & Calmet, J. (2005).** OntoBayes: An Ontology-Driven Uncertainty Model. *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, Vol. 1.

45. **Zesch, T., Müller, C., & Gurevych, I. (2008).** Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco, pp. 1646–1652.

46. **Zhang, X., Asano, Y., & Yoshikawa, M. (2012).** Mining and explaining relationships in Wikipedia. *IEICE Transactions on Information and Systems*, Vol. E95-D, No. 7, pp. 1918–1931.

47. **Zhiqiang, L., Werimin, S., & Zhenhua, Y. (2009).** Measuring Semantic Similarity between Words Using Wikipedia. *2009 International Conference on Web Information Systems and Mining*, pp. 251–255.

Article received on 05/03/2017; accepted on 05/05/2018.
Corresponding author is Alan Ramírez Noriega.