ISSN 0005-1179, Automation and Remote Control, 2015, Vol. 72, No. 8, pp. 345–364. © Pleiades Publishing, Ltd., 2015.
Original Russian Text © I. S. Naumov and V. S. Vykhovanets, 2014, published in Upravlenie Bol'shimi Sistemami, 2014, No. 48, pp. 97–131.

LARGE-SCALE SYSTEMS CONTROL

Using Syntactic Text Analysis To Estimate Educational Tasks' Difficulty and Complexity

I. S. Naumov and V. S. Vykhovanets

Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: naigsa@gmail.com, valery@vykhovanets.ru

Received June 16, 2013

Abstract—We suggest a routine for automatic assessment of complexity and difficulty of educational tasks. This routine is based on text parsing, phrases' predicative structures identification and semantic network construction. Then we develop a mathematical model which employs a notion on semantic distance between words to calculate the volume of knowledge in a semantic network. We show that the volume of knowledge in a semantic network is a measure in the space of all semantic networks, and the semantic distance makes this space the metric one.

1. INTRODUCTION

Lately, a great number of investigations is oriented to automation and intellectualization of different activities. However, there are a lot of unsolved problems in learning automation. This situation is associated with the fact that education is a complex intellectual process poorly subject to formalization [3].

Development of methods, procedures, and devices is a noticeable tendency in learning automation implementing an individual approach to a learning process. Testing and assessment of knowledge, preparation of educational programs, an individual choice of educational tasks, etc. is a current problem of the sphere. An automatic choice of educational tasks is one of the most complicated as it is related to assessment of their subjective difficulty and objective complexity.

Acquiring knowledge, we can conditionally separate a source of knowledge, i.e., description of an object domain or an object of learning and a receiver of learning or a subject of learning. Knowledge of the object domain is admitted as reference knowledge which is constant during the learning process while knowledge of the student is changeable. The purpose of education is to transfer knowledge of the object domain to a learning subject as fully maximum as possible.

The main method of knowledge checking is solving educational tasks. The educational task can be considered as containing knowledge of the object which the student acquires as a result of receiving a right solution to the task. Owing to comparing knowledge of the educational task with corresponding knowledge of the object domain and current knowledge, we can estimate complexity and difficulty of educational tasks.

However, assessment of complexity and difficulty of educational tasks in many known procedures is carried out by indirect criteria, for example, by the number of solution steps, solution time, probability of solution, etc. [28, 29].

We also know the approach in which text complexity is defined as per concrete and abstract words, percentage of new words, length of sentences, complexity of text logical structure and difficulty is estimated after text reading based on experimental checking of text understanding by making questions to the text and analysing answers to the questions [20].

Today, methods based on the analysis of statistical regularities are famous in the process of text assessment. We can separate some papers devoted to the level of text readability [32, 33]. But these methods have a disadvantage: they do not consider concept interrelation of words which are crucial for “understanding” the text meaning.

Eventually, the known methods of assessment of difficulty and complexity of texts and educational tasks are not capable for constructing and using effective individual learning programs. Thus the problem of assessment of tasks’ complexity and difficulty is still topical.

In the paper, we will consider a method of assessment of complexity and difficulty of educational tasks based on automatic text parsing. As a result of educational material parsing, we construct a semantic network of the object domain which is a set of vertexes (notions of the object domain) and a set of links (relationships between notions). To construct a semantic network of the student, we construct a united semantic network of tasks rightly solved by the student. In this case, assessment of complexity and difficulty of a new educational task is conducted by comparing semantic networks of the object domain, a semantic network of the task, and a semantic network of the student.

2. COMPLEXITY AND DIFFICULTY

Notions “complexity” and “difficulty” are often used as synonyms in academic literature. However, these notions have a different meaning and they play crucial roles in the task under study. Usually, complexity means an objective assessment; difficulty, a subjective assessment [18, p. 86].

The notions “complexity” and “difficulty” characterize the link between a subject solving a task and an object, i.e., an educational task. Difficulty and complexity depend on different objective and subjective factors. For example, objective factors include a task subject, task requirements, task conditions; subjective factors may be a subject’s abilities and preparation, its motivation, mental and physical conditions, etc. [4, p. 62].

Finally, we have that each task can have two independent assessments: complexity and difficulty. Under complexity we will understand its objective characteristic which is determined by the content of domain knowledge sufficient for solving it. Under difficulty we will understand its subjective characteristic which can be obtained by comparing knowledge sufficient for solving the task with current knowledge which a student has.

It is to be noted that difficulty of the task cannot increase its complexity: task complexity is an achievable level of difficulty in the learning process. Quantitative assessments of complexity and difficulty are directly associated with processes of acquisition and comparison of knowledge.

3. KNOWLEDGE AND DATA

Generally speaking, knowledge is a result tested through practice of comprehension of objective reality represented in a subject’s consciousness in the form of notions and assertions confirmed by some sequence of inferences [10, art. “Knowledge”]. From the gnoseological point of view, knowledge is defined as a subjectively well-founded belief [7, p. 12]. Thus we admit the subjective (internal) nature of knowledge. In the objective meaning, knowledge is considered as a result, represented in the exterior form, of subjective knowledge admitted as objectively true at some historical moment [25, art. *Knowledge*]. It is believed that ideality of knowledge is an adequate consequence of those properties of the external world which they represent.

Knowledge is formed as a result of task-oriented pedagogical process, self-education, and life experience. In particular, this implies that knowledge need its objectivation, i.e., estrangement from the knowledge carrier in some external objective form. As it is impossible to carry out a transfer of knowledge in an internal (ideal) form, external forms in the type of specially constructed data are used. We can call a natural language (speech–language and writing–language) as a historical example of this.

To transfer knowledge, we use syntactically and semantically open forms of representation supposing the existence of some intended knowledge both of the object domain towards which the knowledge is expressed and knowledge in the form of expressing transferrable knowledge. Only after coordination of these background (intended) knowledge, structure knowledge, and rules of knowledge representation interpretation forms, we obtain an opportunity of adequate transfer and acquisition of new knowledge.

For representation and transfer of knowledge, we use data which are considered as a sequence of states (temporary or spatial) of some material object (a process, an event). Data are accepted by the subject as containing some information. Depending on the subject's emotional state, characteristics of its perception, and available background knowledge, the same data are perceived as containing different information. Consequently, information can be defined as a result of data interpretation conducted as certain, maybe, uncomplete knowledge of their structure and interpretation rules.

From the pragmatic point of view, knowledge is "data obtained at arranged place and at arranged time for solving a practical problem" [12, p. 50]. In this view, knowledge and data do not differ by their structure and storage method. Data become knowledge if and only if they are correspondingly interpreted by machines or human beings.

Another point of view on knowledge is used in "artificial intelligence" [19, p. 224]. Data are knowledge if they:

—are arranged in accordance with the conceptual model of the object domain (object ontology);

—are represented in one of the forms of knowledge representation (semantic networks, frames, scenarios, products, etc.);

—have a procedure of obtaining new (hidden) knowledge from the available one (an effective procedure of knowledge-based inference);

—are stored according to the technique which provides a high efficiency of typical knowledge operations (graph search, hierarchy analysis, logical inference, etc.).

Finally, we obtain that data should be stored, give an opportunity for searching, data should be checked, maintained, and updated. Knowledge, in its turn, is stored as data but as opposed to data it should be transformed from one representation form into another and have a procedure of knowledge-based inference. However, the main difficulty at working with data is not in its representation and handling but in its primary acquisition.

4. KNOWLEDGE ACQUISITION

Let us consider a known classification of knowledge acquisition methods [9, p. 100]. Communication methods of knowledge acquisition (Fig. 1) assume human interaction, namely, interaction of a knowledge engineer, i.e., a knowledge receiver and an object domain expert, i.e., a knowledge source.

Communication methods are divided into passive and active: in passive methods, a crucial role belongs to the expert; in active methods, to the knowledge engineer. Nowadays, full automation of communication methods of knowledge acquisition is questionable as the crucial roles in the process belong to carriers–subjects and knowledge receivers–subjects.

Textological methods mean different documents (teaching techniques, manuals, guidance, monographes, textbooks) by the knowledge source. There are three classes of textological methods oriented to the analysis of special literature, the analysis of textbooks, and the analysis of documents. These methods differ by the content of background knowledge which is necessary for knowledge acquisition from analyzed texts.

The simplest method is a method of analysis of textbooks. The analysis of documents, unlike the analysis of textbooks, is complicated because knowledge in them are very concise: documents

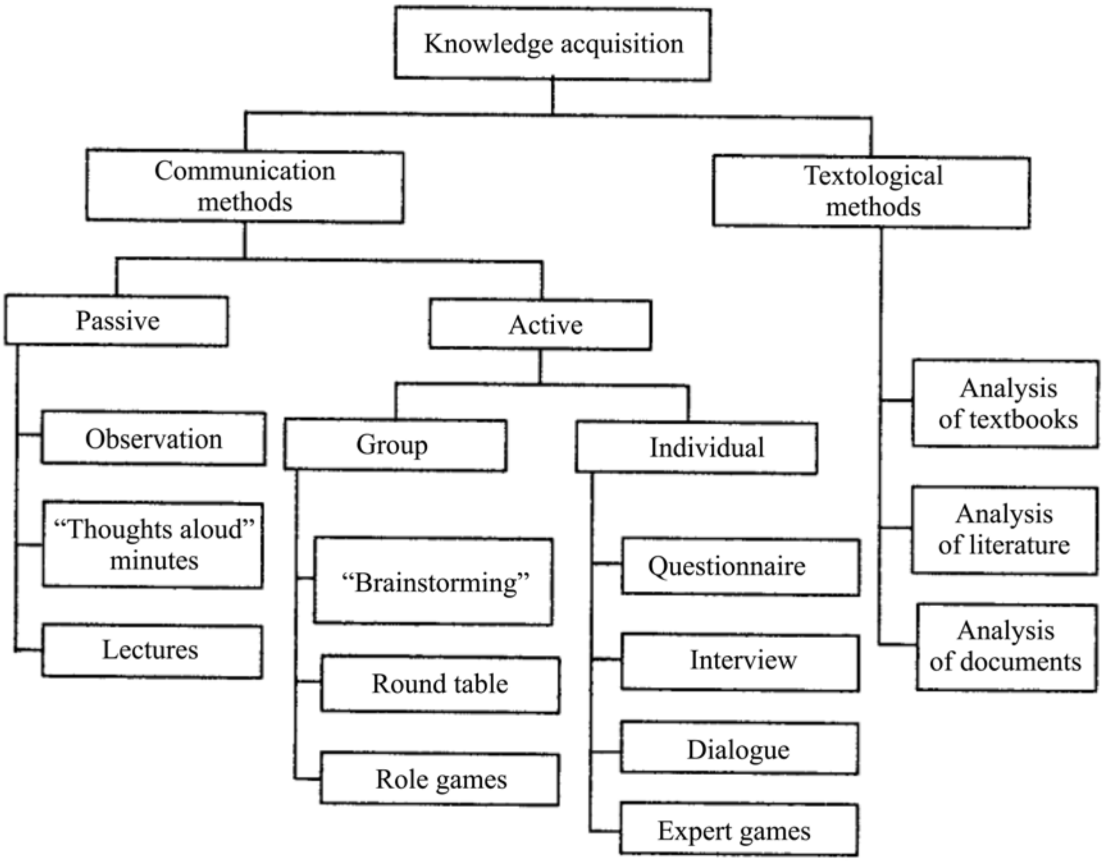


Fig. 1. Classification of knowledge acquisition methods.

have few discussions, notes, and comments. In its turn, special literature requires using a great amount of background knowledge and, therefore, is the most difficult for textological knowledge acquisition.

Unlike communication methods, the textological ones make it possible to introduce automation into the knowledge acquisition process. The reason is that texts have a certain single formal structure. At present, however, a group of textological methods is the least understood. Difficulties faced by researchers are as follows: a problem of text parsing which is in acquisition of meaning of the text and its transformation in one of the known forms of knowledge representation are still unsolved [5, 27].

Perception of any text is conducted at four levels of understanding [17]:

—understanding of contextual meaning of words and sentences (physical perception of the text and understanding of the direct, "superficial" meaning of words and sentences);

—understanding of figurative and symbolic meanings (correlation with the context, understanding of "deep-sea" meanings of words and sentences, correlation with background knowledge, perception of text meaning, its concept);

—understanding of characters and moods of personages (intellectual–emotional perception of the text);

—understanding of the author's attitude towards the recited material.

The first level is a language one, it requires knowledge of the language. At the level, by "uniting" meanings of words, we understand meaning of a sentence as a semantic unit of the text.

The second level is a meaning one or a level of understanding of meaning of the whole text. At this level, we analyze links between meaning units of the text, i.e., sentences.

The third level is a personal one. The reader correlates meaning of the read text with its life experience and knowledge. By thoughts the reader tries to understand the author's intention.

The fourth level is a reflexive one. The reader constructs an author's image and correlates its attitude to the read text with the author's attitude to the text.

It is evident that textological methods operate at the language level and do not apply for higher levels of text "understanding." However, qualitative identification, not full or partial acquisition of knowledge from texts, is sufficient for many applied spheres including the sphere of assessment of complexity and difficulty of educational tasks.

5. KNOWLEDGE IDENTIFICATION

It is known that any connected text generates some semantic space. "Semantic space of words is a domain of existence and functioning of their lexical meanings. Meanings of words do not exist singularly but they are linked with specific relationships and links with each other thanks to which words are also related as linguistic units" [31].

Semantic links of the text can be separated into explicit and implicit. Explicit links are links directly determined in the text and established at the parsing stage. Links between words defined by background knowledge and typical of any natural language are implicit links [27].

However, a great part of semantic links is detected at the parsing stage since at writing texts we are trying to minimize a necessary content of background knowledge. Otherwise, texts with a great number of implicit links are difficult for reading and understanding.

Lexical meaning of the word is a notion expressed by it. "A notion is an idea, which separates objects from some object domain and collects them into a class (generalizes) by indicating their general and distinguishing characteristics" [24].

It is to be noted that relationships between notions unlike relationships between words have some distinguishing characteristics: a notion can be expressed not with one but with a set of words; two notions can be linked with relationships which are different in various spheres of knowledge; a notion can have relationships towards itself. It is also important that notions unlike words–notions used in some text is not a final result of knowledge: notions are in constant concept specification and alteration [16].

As the main provision of the described approach, let us accept supposition that identification of knowledge containing in the text can be conducted by analyzing the sequence of its words. In this case, applied notions, which have a different meaning in the text, are considered as different notions.

Then the connected text can be represented as a sequence of words–notions connected with each other by explicit and implicit relationships; knowledge containing in the text can be identified with explicit relationships between its words accurate to implicit relationships generated by background knowledge. For example, such abstract notions as a "line" and a "point" have explicit relationships expressed by the following sentences: "The point lies on the line," "The line passes through the point," and "The point belongs to the line."

Unlike the known approaches, at which for identification of knowledge relationships between notions of the object domain are detected based on text parsing involving experts' participation, identification of knowledge contained in the text is conducted by establishing relationships between its words according to results of text parsing.

Therefore, by identification of knowledge we mean construction of a text semantic network based on its parsing. The provision of this approach is the follows: a text semantic network constructed according to results of text parsing is a subnetwork of its semantic network. We assume in this case that interrelations of notions obtained as consequence of text parsing shall be obligatorily entered into its semantic network.

6. SYNTAX OF SENTENCES

A simple sentence in the Russian language has a predicative structure and can be represented by a grammatical predicate the arguments of which are a grammatical subject and a grammatical object.

Usually, the subject expresses a grammatical object; the predicate, a grammatical predicate; the object, a grammatical subject. In its turn, a complex sentence consists of simple sentences and includes two or several predictive units constituting a unity in the terms of meaning, construction, and intonation [8].

Generally, the simple sentence in the Russian language is represented in the form of a tree where each arc goes from the main word to the dependent one and has a name of syntactic relationship [22, 38]. To simplify representation of sentences, strongly associated words are united into syntactic groups. In this case, one of parts of the group (a word) always acts as a group's representative and it subordinates other parts. The process of construction of such a tree is called parsing.

We shall call the process of separation of syntactic groups in the sentence as sentence parsing; sentence parsing will be defined as comparison of its word groups with the predicative structure.

Figure. 2 represents an example of parsing and analysis of the sentence "Perpendicular lines form a right angle."

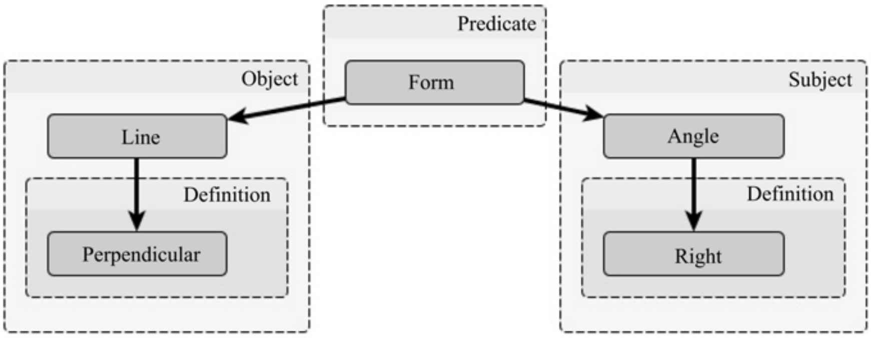


Fig. 2. Example of parsing and analysis.

The grammatical predicate is always at the top of the tree obtained after simple sentence parsing. The group of the grammatical predicate is the main in the sentence and it subordinates groups of the grammatical object and grammatical subject. Unlike the group of the grammatical predicate, presence of the groups of the grammatical object and grammatical subject is not obligatory.

Each of main syntactic groups of words may have other word groups. For example, there are defining, quantitative, circumstantial, continuous, coordinating, auxiliary, etc. groups. As a rule, separation into groups depends on a concrete implementation of the parser.

We shall assume that the grammatical object and grammatical subject describe notions of the object domain; the grammatical predicate, a link between them. By notion we mean the main word of the grammatical group of the object or subject complemented with definitions. For example, the notions "Triangle," "Isosceles triangle," and "Tight isosceles triangle" are different notions. By the predicate we mean the main word which belongs to the predicate group.

Sometimes, in the text there sentences without the object or subject. In this case, we introduce an "empty" notion for establishing a link between the object and the subject. For example, in the sentence "There are some types of shapes" there is an object expressed by the notion "types of shapes" but the subject is absent. Therefore, the notion "types of shapes" connects with the "empty" notion by the relationship "exist."

By means of simple sentence parsing we can make one judgement which reports interrelated notions and the character of their interrelation to us.

However, to identify knowledge of the text, it is necessary to consider all judgements. We shall unite judgements based on representation of a text in the form of a semantic network.

7. A SEMANTIC NETWORK

As a result of sentence parsing, we can obtain its semantic network (multigraph) consisting of nodes (vertexes) marked by names of grammatical subjects and arcs specifying relationships between them and marked by names of grammatical predicates. As the text consists of a set of sentences, we unite semantic networks of these sentences into a single semantic network.

Let two semantic networks are determined: a network of the text S and a network of the simple sentence S'. The network of the text S shall be determined in the form of an ordered set of three elements:

$$S = (N, E, P), \tag{1}$$

where $N = \{n_i | i = 1, \ldots, Q\}$ is a set network nodes and number of elements Q; $E = \{(n_i, n_j, p_k) | n_i, n_j \in N; p_k \in P\}$ is a set of its arcs; P is a set of binary predicates. The arcs are determined by the ordered sets of three elements $(n_i, n_j, p_k) \in N \times N \times P$ where $n_i \in N$ is an initial node, $n_j \in N$ is a final node, $p_k \in P$ is an arc name, $\times$ is an operator of Cartesian product of sets.

A network of the sentence S' is simple and it consists of two nodes n_1, n_2 and one arc marked by the name of some predicate p:

$$S' = (N', E', P'), \quad N' = \{n_1, n_2\}, \quad E' = \{(n_1, n_2, p)\}.$$

Then a union of networks $S = (N, E, P)$ and $S' = (N', E', P')$ will be a network $S'' = S \cup S'$ such that $S'' = (N'', E'', P'')$ and $N'' = N \cup N'$, $E'' = E \cup E'$, $P'' = P \cup P'$.

In a similar way, we can determine an intersection of networks. The intersection of networks $S' = (N', E', P')$ and $S'' = (N'', E'', P'')$ will be a network $S = S' \cap S''$ such that $S = (N, E, P)$ and $N = N' \cap N''$, $E = E' \cap E''$, $P = P' \cap P''$.

In its turn, a difference of networks $S' = (N', E', P')$ and $S'' = (N'', E'', P'')$ is a network $S = S' \backslash S''$ such that $S = (N, E, P)$ and $N = N' \backslash N''$, $E = E' \backslash E''$, $P = P' \backslash P''$.

To simplify semantic networks, we depict them in the form of a weighted directed multigraph. In this case, instead of names of grammatical predicates, we specify multiplicity of arcs equal to a number of different predicates connecting corresponding nodes. As we study not statistical properties of the text but knowledge expressible by it, at specifying multiplicity of the arcs we shall not consider frequency of repeating the same predicate. Finally, we have that multiplicity of the arc between two nodes of the network is equal to a number of different grammatical predicates connecting corresponding notions–words.

Figure 3 depicts a fragment of a semantic network obtained after parsing of the section "Main Properties of Geometric Shapes" of the geometry manual [26]. The semantic network contains 90 notions of the analyzed object domain and 215 relationships specified for them.

The network was obtained after performing the following actions:

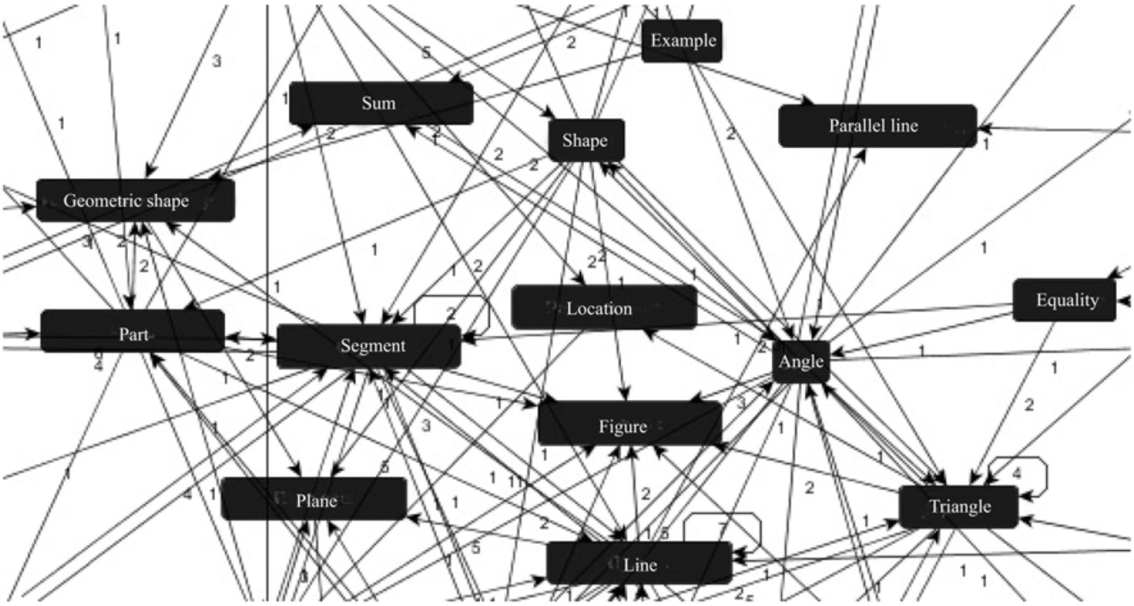


Fig. 3. Fragment of the semantic network "Planimetrics."

—search for notions expressed by singleword and multiword terms [6];

—establishment of syntactic relationships between detected notions;

—dereference of obtained reference links.

The constructed semantic network is a connected graph any node of which has at least one path connecting it with another node.

8. SEMANTIC DISTANCE

Two texts express close knowledge if the intersection of their semantic networks is commensurable with their union. At the same time, two texts have different knowledge if the intersection of their semantic networks is small as compared to their union. To receive a qualitative assessment of proximity of textological knowledge, it is necessary to be able to determine a distance between semantic networks and nodes of one semantic network.

Two main approaches to measure semantic distances between words are known: paradigmatic and syntagmatic [21]. The paradigmatic approach is based on measurement of semantic distances in a language lexicon where the lexicon is specified as a set of classes of words connected with paradigmatic links; by a paradigmatic link we mean a relationship between words which have generality of values. Paradigmatic relationships define ontological properties of the language on the whole and represent those links which exist between events of reality expressed by them without reference to any object domain.

The syntagmatic approach is based on measurement of semantic distances between words and texts using statistical methods. Two groups of methods are segregated: statistical and logical–and–statistical. Statistical methods are methods based on the "frequency–rank" model and they are described by Zipf, Mandelbrot, and Bradford laws. These laws express dynamic dependency of the word frequency in the text on its rank, where the rank of the word is a numerical order of the word in the frequency word-book of the language.

Logical–and–statistical methods include distributive–and–statistical, hypertext link, frequency–and–semantic, and componential analysis methods. The distributive–and–statistical method makes it possible to obtain a quantitative characteristic of words connexity in the text based on computing some previously given formula depending on statistical characteristics of words in the text. The hypertext link method establishes a link between words–notions based on general words in the definition of this notions and the frequency–and–semantic method in addition considers their frequency. In its turn, the componential analysis method relates words–notions by decomposition of their values into minimal semantic components, i.e., semes. The componential analysis method is based on the hypothesis that a value of each language unit consists of semes and a language word stock can be described using a limited number of them.

The first measure of semantic similarity of concepts–notions was created to assess similarity of nodes in taxonomies [2], which are a partial case of the semantic network. A measure of similarity *sim* of two nodes n_1 and n_2 in taxonomy is an inverse value of the length of the shortest path between these two nodes $d(n_1, n_2)$:

$$sim(n_1, n_2) = \frac{1}{d(n_1, n_2)}.$$

A normalized value of similarity was proposed in [35]; it is defined via a logarithm of the ratio of the shortest path in the taxonomy between two nodes n_1 and n_2 to its doubled diameter:

$$sim(n_1, n_2) = -\log \frac{d(n_1, n_2)}{2D},$$

where D is a diameter of taxonomy or a maximal distance between its nodes.

In another type of measures, we use semantic networks constructed on the basis of definitions of notions. The semantic network of the two words–notions n_1 and n_2 directly proportional to the number of words–notions which are a part of the definition of the first and second notions:

$$sim(n_1, n_2) = |gloss(n_1) \cap gloss(n_2)|,$$

where $gloss(n)$ is a set of words–notions in the definition of the word–notion n.

We have in a number of cases that there are no common notations at the intersection of definitions of comparable words–notions. Therefore, in [36] it is considered not only words–notions which participate in the definition of each notion but words–notions which connect these notions in the semantic network of the text.

The methods above of definition of the semantic distance between notions are based on particular types of relationships typical of taxonomies. Detection of such relationships in the text is a nontrivial task, which finally requires involvement of experts specialized in the object domain [23].

We define the semantic distance between notions for the tasks under study another way. Let there is a semantic network S. We fix its two arbitrary nodes n_i and n_j. We obtain $R(n_i, n_j)$ which is a set of paths without cycles (chain) of the length not more than M leading from the node n_i to the node n_j. Then the semantic distance L between the nodes n_i and n_j can be computed by the formula:

$$L(n_i, n_j) = \sum_{r \in R(n_i, n_j)} \frac{\min(w_1^r, w_2^r, \ldots, w_{d(r)}^r)}{d(r)}, \tag{2}$$

where $d(r)$ is a length of the path r, $d(r) \leq M$; M is a depth of the link; w_i^r is a weight of the arc i of the length r, $i = 1, 2, \ldots, d(r)$; min is a function returning a minimal value of its arguments.

It follows from formula (2) that two nodes are remote from each other if there are a lot of paths between them (notions are slightly intertwined). Remoteness of two nodes is larger, the greater weight of arc joining them is (more variative are links between notions). However, if on the path there is an arc with a small weight, then this node makes a smaller contribution to remoteness of nodes from each other. But not all paths are considered at counting the distance between nodes: those paths are excluded, the length of which is greater that the specified depth of the link (it is difficult to establish a link between notions as this requires a great number of sentences).

Is to be noted that semantic distance (2) is not metric as axioms of identity, symmetry, and triangle inequality are not fulfilled for it. This is because the notion cannot have a relationship to itself, links between two notions are not symmetric, and similarity of two notions depends not only on directly connecting them predicates but also on predicates specifying an indirect link through another notions.

Two notions connected with a long path are recognized as slightly intertwined. At computing semantic distances, we specify the depth of considered links M. For many practical applications, M can be chosen from a range from two till seven. A standard interpretation of M is a number of judgements which which a student can operate or a maximal number of judgements occurring in its references.

Thus, the grammatical predicate is a unit of measurement of semantic distances which can be used for expressing one judgement. It is to be noted that judgements are the main unit of measurement of knowledge content in didactic researches [20].

If two nodes of the network are not connected with one path, then calculation of the semantic distance between them produces a quantity equal to zero. Zero semantic distance denotes absence of a link between corresponding notions and confirms their meaningful indiscernibility. On the other hand, the greater quantity of semantic distance between nodes of the network is, the more remote are corresponding notions as for the content.

The semantic distance $L(n, n)$ between the same node n is illustrative. If $L(n, n)$ is equal to zero, then the corresponding notion can be recognized as simple. If the quantity $L(n, n)$ is great, then the corresponding notion is complex and can be recognized as not unveiled by content.

9. MEASUREMENT OF KNOWLEDGE

Besides the semantic distance between nodes of the network for assessment of difficulty and complexity of educational tasks, it is necessary to compute the semantic distance between semantic networks and define a content of knowledge contained in it.

Measurement of knowledge contents is still conducted by methods based on expert assessments of students' educational activities (subjective methods) an testing of students (objective methods). These methods are based in the modern theory of pedagogical measurements [1] where the learning process is considered as students' constant overcoming of an edge between the achievable area of knowledge (a level of relevant development) and potentially possible (a zone of proximal development). A teachers' aim is to find difficult but adequate tasks promoting the level of relevant development [13].

We also know an approach according to which measurement of knowledge is conducted on the basis of measurement of capacity of notions where by capacity of notion we mean a number of links of this notion with other notions; a unit link is a unit of measurement [15]. In this case, measurement of knowledge content in a text, topic, or an educational discipline is reduced to detection of notions of the object domain and calculation of the number of links between them by expert methods.

To determine similarity of two semantic networks, we search for homomorphisms transforming one network into another. However, the search for a homomorphism makes it possible to determine only a qualitative "similarity" of networks but it does not allow to measure knowledge contents containing in these networks.

It is obvious that the listed methods are incapable for definition of a knowledge content in a text semantic network obtained on the basis of parsing as in one case we need involvement of experts (an expert opinion, testing, calculation of capacity of notions); in another case, there is no effectively calculated distance between semantic networks (the search for homomorphisms).

By a knowledge content contained in a semantic network $S = (N, E, P)$ we mean a quantity calculated using the following formula:

$$K(S) = \sum_{n_i, n_j \in N} L(n_i, n_j), \tag{3}$$

where $K(S)$ is a knowledge content in the semantic network S, $L(n_i, n_j)$ is a semantic distance between the nodes n_i and n_j calculated using formula (2).

Formula (3) affirms that the knowledge content in the network S is a sum of semantic distances between all pairs of its nodes.

As with the semantic distance, a measurement unit of the knowledge content is a grammatical predicate.

Theorem 1. *The knowledge content* (3) *is an additive measure on a set of semantic networks.*

To prove Theorem 1, let us first show that the measure of an empty network is equal to zero. Indeed, if the semantic network S is empty, $S = (\varnothing, \varnothing, \varnothing)$, then from formulae (3) and (2) we immediately have $K(S) = 0$. From formulae (3) and (2) we also have a statement that the measure of union of two networks S_1 and S_2 such that $S_1 \cap S_2 = (\varnothing, \varnothing, \varnothing)$, is equal to the sum of these measures: $K(S_1 \cup S_2) = K(S_1) + K(S_2)$. ⧫

Unlike another known approaches, formula (3) makes it possible to measure the knowledge content contained in an arbitrary text.

We also know some approaches to define distances between graphes. This means using a height of a directed graph, which is equal to the greatest length of the path from the root to a leaf in a multilevel structure of its representation [30]. We also use a distance obtained on the basis of computing a diameter of the graph, a maximal number of edges connecting its two vertexes [11]. We also know a distance between graphes obtained by computing edge density, i.e., a numerical quantity characterizing similarity of the graph to a fully connected one [14].

It is obvious that the listed approaches are inapplicable for definition of distances between semantic networks. A semantic distance between the networks S_1 and S_2 is defined as a knowledge content contained in the symmetric difference of these networks:

$$D(S_1, S_2) = K(S_1 \backslash S_2 \cup S_2 \backslash S_1). \tag{4}$$

Theorem 2. *The semantic distance* (4) *is metric on the set of semantic networks.*

Ro prove Theorem 2, it is sufficient to show that axioms of identity, symmetry, and triangle are fulfilled on the set of semantic networks. Let S_1 and S_2 are semantic networks. If $S_1 = S_2$, then from (4) it follows $D(S_1, S_2) = 0$. Let now $D(S_1, S_2) = 0$. From (4) we have that $S_1 = S_2$. Finally, $D(S_1, S_2) = 0$ if and only if $S_1 = S_2$. The axiom of identity is proved. The axiom of symmetry also immediately follows from (4): $D(S_1, S_2) = D(S_2, S_1)$. The axiom of triangle follows from formula (4) and Theorem 1:

$$K(S_1 \backslash S_2 \cup S_2 \backslash S_1) + K(S_2 \backslash S_3 \cup S_3 \backslash S_2) \geq K(S_1 \backslash S_3 \cup S_3 \backslash S_1).⧫$$

Therefore, a set of text semantic networks is metric space; a semantic distance between two networks is equal to a cumulative knowledge content contained in them.

10. ASSESSMENT OF KNOWLEDGE

Assessment of knowledge assumes comparison of student's available knowledge with reference knowledge. In our case, reference knowledge is knowledge on the object domain; student's knowledge is determined by tasks solved by it. If solving of educational tasks is successful, a student shows knowledge acquired by it.

Let us consider an approach proposed in the paper for assessment knowledge contents as well as difficulty and complexity of educational tasks. Let there are following semantic networks:

—S is a semantic network of the object domain obtained by text parsing describing this object domain;

—T is a semantic network of the educational task obtained by parsing of its text;

—O is a student's semantic network obtained by a successive union of semantic networks of tasks successfully solved by students.

A result of comparing the network of the object domain S and the network of the current task T determines complexity of the task (an objective characteristic of the task); comparison of the network of the current task T and the student's network O, determines difficulty of the task (a subjective characteristic of the task). Comparison of the network of the object domain S and the student's network O allows determining the content of knowledge still unlearned by it (Fig. 4).

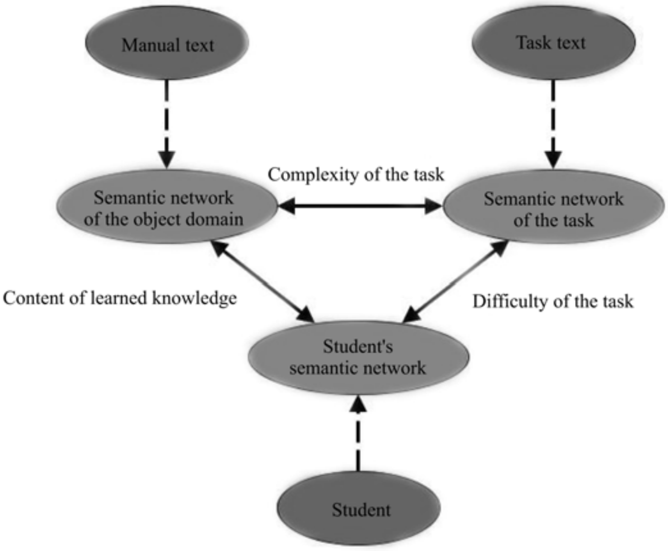

Manual text
Task text
Complexity of the task
Semantic network of the object domain
Semantic network of the task
Content of learned knowledge
Difficulty of the task
Student's semantic network
Student

Fig. 4. Semantic characteristics.

Considering specific properties of the task being solved, we can show that the networks S, T, and O are matched, i.e., $T \subseteq S$, $O \subseteq S$ where the ratio of entry of the networks $\subseteq$ is determined as follows: if $S' = (N', E', P')$, $S'' = (N'', E'', P'')$ and $S' \subseteq S''$, then $N' \subseteq N''$, $E' \subseteq E''$ and $P' \subseteq P''$.

Finally, we have the following formulae allowing computing complexity and difficulty of the educational task T and the content of still unlearned knowledge:

$$U(O) = K(S\backslash O), \quad u(O) = K(S\backslash O)/K(S); \tag{5}$$

$$C(T) = K(T \cap S), \quad c(T) = K(T \cap S)/K(S); \tag{6}$$

$$H(T) = K(T\backslash O), \quad h(T) = K(T\backslash O)/K(O); \tag{7}$$

$$\bar{H}(T) = K(T), \quad \bar{h}(T) = K(T)/K(O). \tag{8}$$

The content of unlearned knowledge $U(O)$ and relative content of unlearned content $u(O)$ are computed by formula (5) based on the content of knowledge contained in the semantic network of the object domain S and the student's semantic network O.

The complexity C of the educational task T is determined by formula (6) and it is equal to the content of knowledge from the semantic network of the object domain S contained in the semantic network of the task T. The relative complexity of the task c is a share of knowledge of the object domain contained in the task T.

The difficulty H of the educational task T is determined by formula (7) as a content of new knowledge contained in the semantic network of the task T with respect to the student's semantic network O. The relative difficulty of the task h is a ration of new knowledge in the task T to knowledge which a student has.

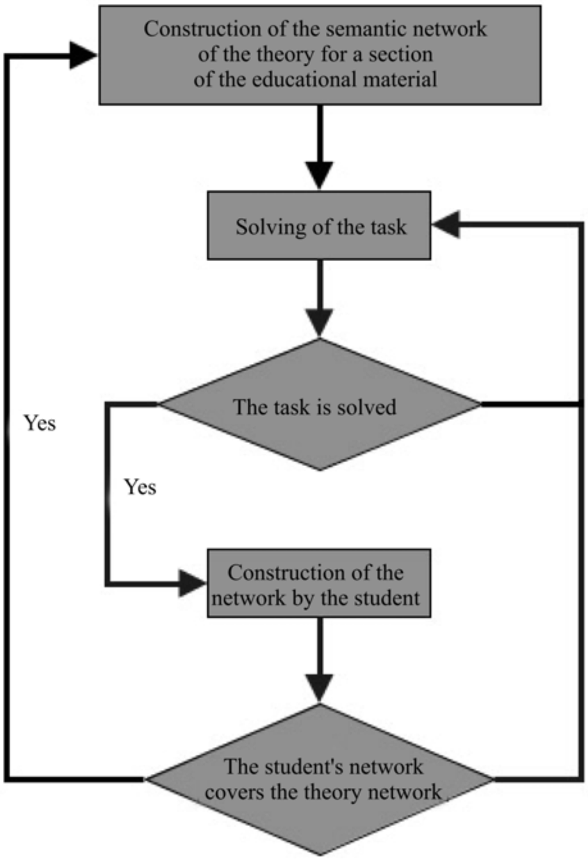

Construction of the semantic network of the theory for a section of the educational material
Solving of the task
The task is solved
Yes
Yes
Construction of the network by the student
The student's network covers the theory network

Fig. 5. Learning process.

Besides the difficulty of the educational task T connected with availability of new knowledge in it, we can put to use integral difficulties of the educational task $\bar{H}$ and $\bar{h}$ (8) considering the total content of knowledge in the task. Integral difficulties make it possible to take a number of solution steps (a number of judgements) into account which a student will conduct so that to follow links between all notions of the task.

Let us now consider construction of the above semantic networks in the learning process. As a rule, a course of studies is split into parts (sections). Testing of knowledge is carried out after each part; it is implemented through solving corresponding tasks and their checking (Fig. 5).

The first stage is construction of a semantic network of the object domain S as per the current part of educational material. The second stage is solving check tasks by the student and their checking. Different strategies of choosing check tasks can be used; for example, in increasing their difficulty or complexity. After learning the material by the student (the content of unlearned knowledge is equal to zero or is relatively small), we construct the semantic network of the object domain for a new part of the learning material. The network is not constructed from the beginning as it is completed.

11. A WORKING EXAMPLE

Let us now consider the above method of knowledge assessment by the following example. Let we have a text of the task: "A point D is taken on the side AB of the triangle ABC. What is the length of the side AB of the triangle, if the side AD is equal to 5 cm and the side BD is 6 cm?"

In this task we have three notions and two ratios, one of which is met twice. Figures 6 and 7 represent a network of the task as well as fragments of the student's network and network of the object domain. A fragment of the object domain includes all notions of the network of the task. It is to be noted that it follows from the student's network that the student does not know how the notions "Side" and "Point" are correlated.

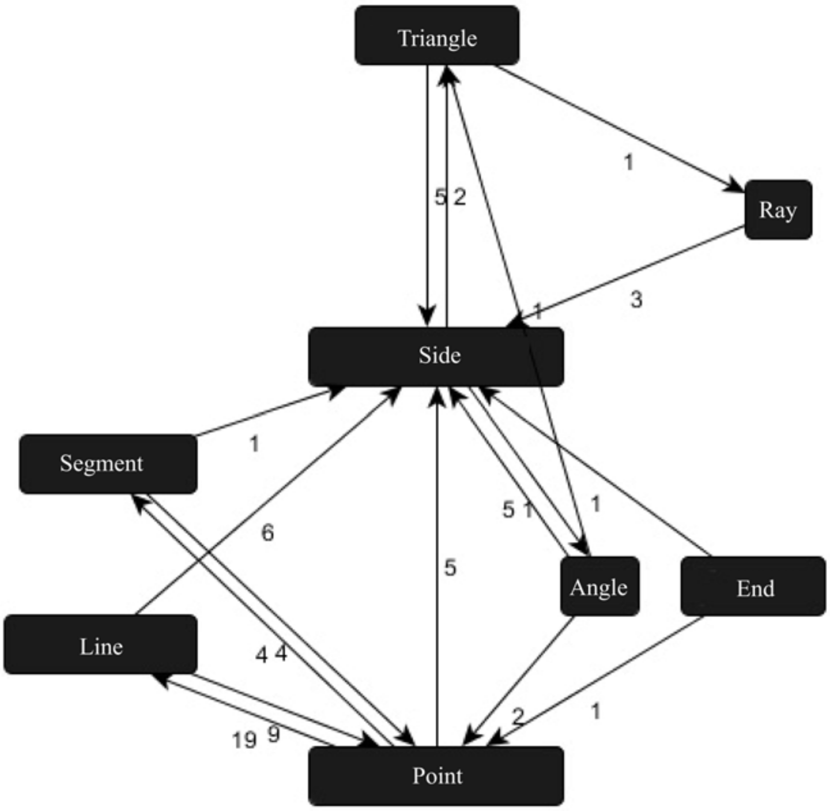


Fig. 6. Semantic network of the object domain S.

We shall preliminary show how computation of semantic distances between notions is conducted. We consider the notions "Point" and "Side" in a semantic network of the object domain S. To simplify computations, we shall specify a depth of retraceable links M equal to two. We write a set of paths R (point, side) connecting the chosen nodes: $\{$(Point, Side), (Point, Line, Side), (Point, Segment, Side)$\}$. Using formula (2), we find

$$L(\text{Point, Side}) = \frac{\min(5)}{1} + \frac{\min(19,6)}{2} + \frac{\min(4,1)}{2} = 8.5.$$

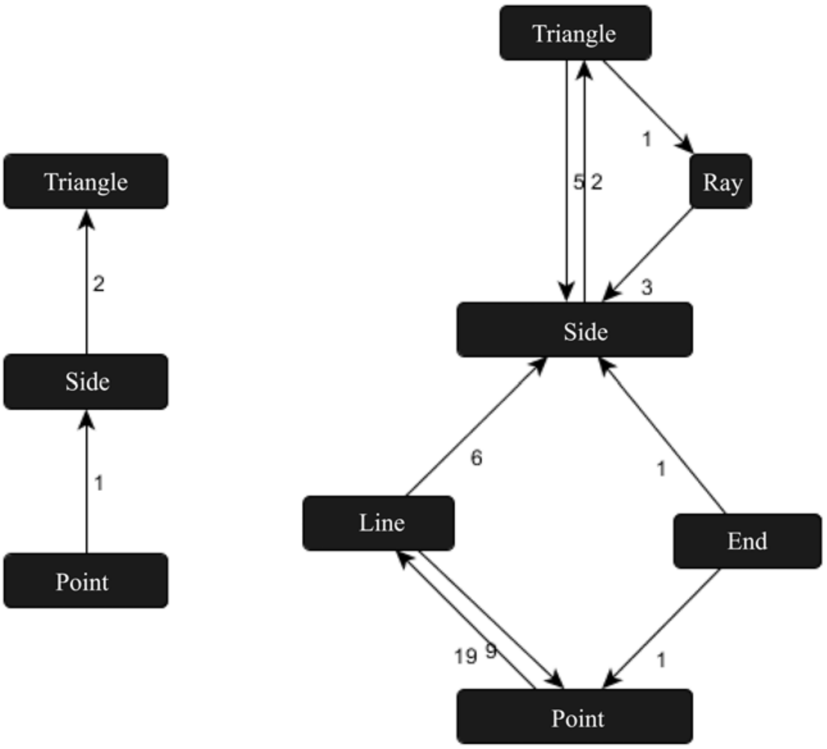


Fig. 7. Semantic network of the task T and student O.

The obtained results can be interpreted as follows: the notion "Point" is connected with the notion "Side" 8.5 with different predicates. The semantic distance between the notions "Side" and "Point" is equal to another value: L (Side, Point) = 0.5.

We shall show computation of the content of knowledge using the definition of difficulty $H(T)$ in the educational task T. A content of new knowledge contained in the semantic network of the task T with respect to the student's semantic network O is determined by a semantic network $T\backslash O$ which in our example consists of the nodes "Side", "Point," and an arc connecting them. The content of knowledge contained in this network is equal to unit. Hence, the difficulty of the task $H(T)$ is also equal to unit. The obtained quantity $H(T)$ determines a content of knowledge which the student will receive if it rightly solves the task T.

12. COMPUTATIONAL EXPERIMENT

We chose an ETAP-3 parser developed at the Institute for Information Transmission Problems, Russian Academy of Sciences, for conducting a computational experiment as a language processor [22]. As compared to other parsers, the ETAP-3 parser has a series of advantages: there is a generally available service for automatic text parsing, results of the analysis are transformed in xml format (UNL), sentence parsing is conducted without limiting complexity.

Using ETAP-3, we conducted parsing of two chapters of the object domain “Planimetrics”: “Main Properties of the Simplest Geometric Shapes” and “Geometrical Construction.” Each chapter is separated into three parts: the object domain, questions, and tasks. For example, we obtained after parsing of the chapter “Geometrical Construction” 242 words–notions and 817 relationships between them.

For conducting the computational experiment, we developed a program (Fig. 8) which loads files with results of parsing in UNL format and produces semantic networks from them. The following tools were implemented for working with semantic networks in the program:

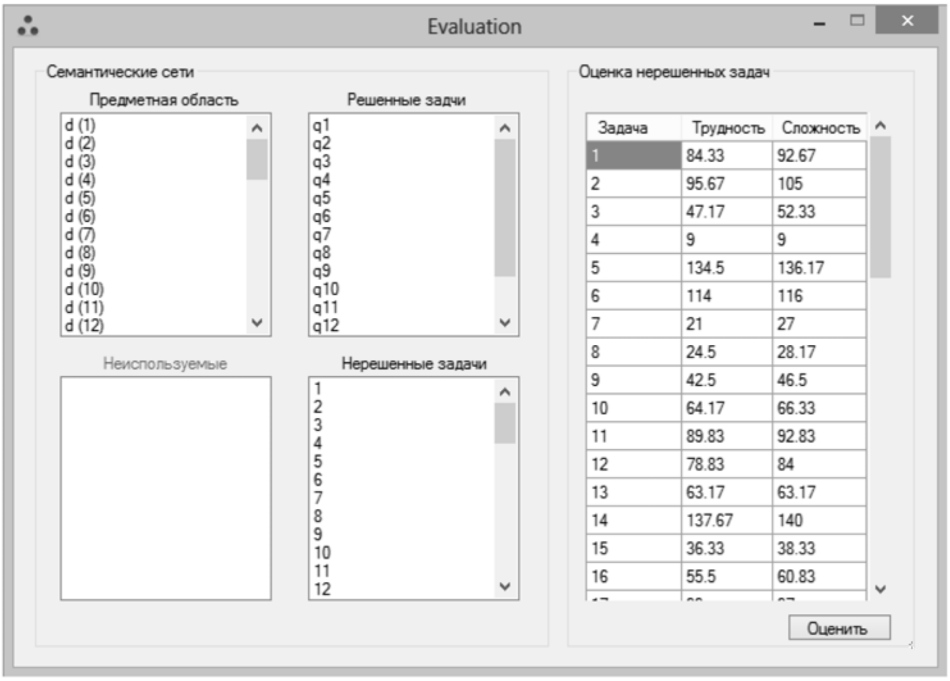
—addition, deletion, union of semantic networks;

—transfer of a semantic network in different groups (the object domain, solved tasks, unsolved tasks);

—viewing of statistic data of semantic networks (frequency of words and relationships, parts of speech);

—viewing of semantic networks in the graphic mode.

A process of solution and assessment of tasks was modeled using the chapter “Geometrical Construction.” A network made on the basis of semantic networks of questions was taken as a student’s basic network; namely, it was supposed that the student has answered all questions and it has to solve 30 tasks.



Предметная область

d (1)
d (2)
d (3)
d (4)
d (5)
d (6)
d (7)
d (8)
d (9)
d (10)
d (11)
d (12)

Решенные задачи

q1
q2
q3
q4
q5
q6
q7
q8
q9
q10
q11
q12

Неиспользуемые

Нерешенные задачи

1
2
3
4
5
6
7
8
9
10
11
12

Задача	Трудность	Сложность
1	84.33	92.67
2	95.67	105
3	47.17	52.33
4	9	9
5	134.5	136.17
6	114	116
7	21	27
8	24.5	28.17
9	42.5	46.5
10	64.17	66.33
11	89.83	92.83
12	78.83	84
13	63.17	63.17
14	137.67	140
15	36.33	38.33
16	55.5	60.83

Fig. 8. Program interface.

Figure 9 displays diagrams of step change of difficulty of tasks during their solving. The least complicated task was chosen for solving at each step. After choosing the task, its semantic network is added into the student’s semantic network, which caused less difficulty while solving other tasks at the following steps of learning.

Figure 9 also depicts that solving tasks with little difficulty insignificantly influences the change of difficulty of other tasks at the following steps of learning. However, while choosing for solving more complex tasks we observe a considerable decrease in difficulty of still unsolved tasks.

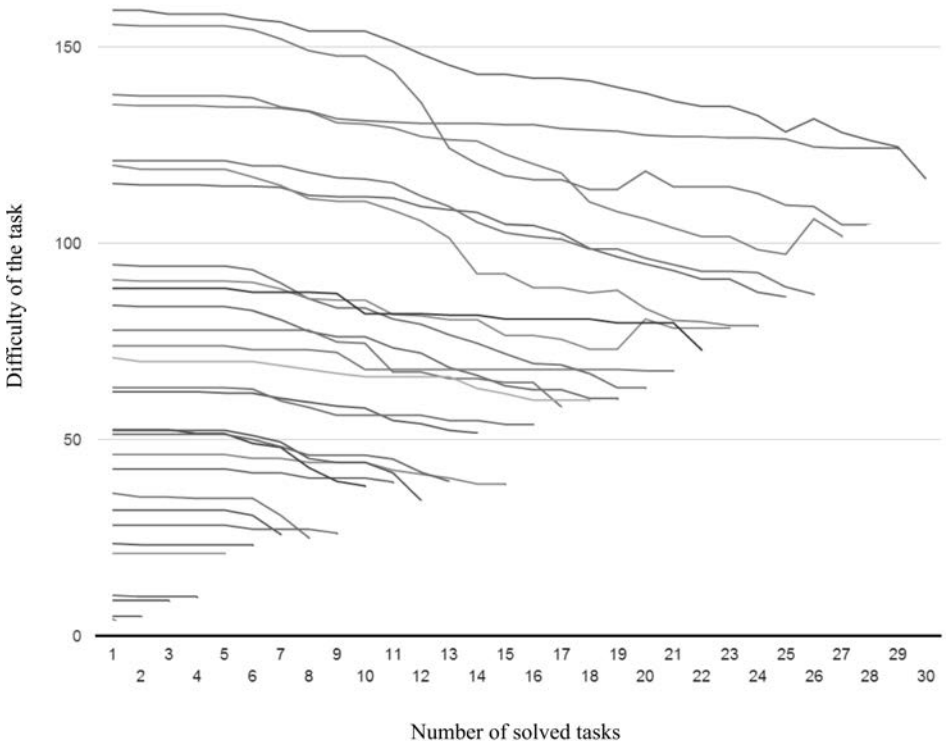


Fig. 9. Changes in difficulty of tasks while solving.

13. CONCLUSIONS

Assessment of complexity and difficulty of educational tasks is one of the most crucial problems in the sphere of learning automation. In the paper, we propose a solution to this task based on text parsing oriented to detection of the predicative structure in components composing its sentences and construction of semantic networks of texts according to results of this analysis.

We developed a mathematical model which employs a notion of the semantic distance between words–notions detected at the process of automatic parsing to calculate the content of knowledge in a semantic network.

We showed that the content of knowledge contained in semantic networks is a measure on the set of semantic networks and the introduced distance between semantic networks makes the set of semantic networks the metric space.

Unlike other existing methods of detection contents of knowledge based on using ontology and thesaurus, the developed method is universal as it is not related to any concrete object domain and involvement of experts for its primary description is not required. The main condition for applicability of the method is provision of description of the object domain in the set of texts in a natural language and availability of an algorithm detecting the predicative structure of sentences.

In the proposed method of measuring the content of knowledge as in Shannon–Hartley information theory [34, 37], we abstracted from the psychical character of the testing events and found such a material object by the properties of which we can judge on the rate of modelled psychical processes.

In information theory, this is probability of event or frequency of its reporting to the person subject: the more frequently some message appears, the more it is unexpected for the person subject and it is perceived as containing less content of information. For considering specific features of perception of messages, the content of information in the message is determined as a negative logarithm of its probability.

In the developed method, measurements of the content of knowledge is a text submitted to a person subject as a set of sentences with a predicative structure and perceived by it as having some knowledge: the greater is the number of notions and predicates connecting them in the text, the greater is the content of knowledge that can be perceived from the text. For considering an active character of knowledge, the depth of traceable links at counting the content of knowledge is limited by an ability of the person subject to establish conceptional links between notions by conducting inferences with a certain number of initial judgements in them.

REFERENCES

1. Avanesov, B.S., Knowledge as a Studies Course of Pedagogical Measuremnts, *Pedagogich. Izmeren.*, 2005, no. 3, pp. 43–52.

2. Anisimov, A.V., Liman, K.S., and Marchenko, A.A., Methods of Computing Measures of Words Semantic Similarity in a Natural Language, URL: http://lingvoworks.org.ua (Cited April 9, 2013).

3. Borisov, I.K., *Optimizatsiya protsessa obucheniya: Obshchedidakticheskii aspekt* (Optimization of Education Process. Common Pedagogical Aspect), Moscow: Pedagogika, 1977.

4. Ball, G.A., *Teoriya uchebniykh zadach: psichologo-pedagogicheckii aspekt* (Educational Tasks Theory: Psychological-and-Pedagogical Aspect), Moscow: Pedagogika, 1990.

5. Belonogov, G.G., *Komp'yuternaya lingvistika i perspektivnye informatsionnye tekhnologii* (Computational Linguistics and Advanced Information Technology), Moscow: Russkii Mir, 2004.

6. Braslavski, P.I. and Sokolov, E.A., Comparison of Five Methods for Variable Length Term Extraction, *Trudy mezhd. konf. "Dialog 2008"* (Proc. Int. Conf. "Dialog 2008"), 2008, pp. 67–74.

7. Vagin, V.N., Golovina, E.Yu., Zagoryanskaya, A.A., and Fomina, M.V., *Dostovernyi i pravdopodobnyi vyvod v intellektual'nykh sistemakh* (Exact and Plausible Inference in Intelligent Systems), Moscow: Fizmatlit, 2008.

8. Valgina, N.S., Rosenthal', D.E., and Fomina, M.I., *Sovremennyi russkii yazyk: uchebnik* (Modern Russian Language: Textbook), Valgina, N.S., Ed., Moscow: Logos, 2002.

9. Gavrilova, T.A. and Khoroshevsky, V.F., *Bazy znanii intellektual'nykh sistem* (Knowledge Bases of Intelligent Systems), St.-Petersburg: Piter, 2000.

10. Danil'yan, O.G. and Panova, N.I., *Sovremennyi slovar' po obshchestvennym naukam* (The Modern Dictionary on Social Studies), Moscow: Eksmo-Press, 2007.

11. Evstigneev, V.A., *Primenenie teorii grafov v programmirovanii* (Application of Graph Theory to Programming), Moscow: Nauka, 1985.

12. Ermakov, A.E., Knowledge Extraction from Text and Its Processing: Current State and Prospects, *Inform. Tekhn.*, 2009, no. 7, pp. 50–55.

13. Efremova, N.F., *Testovyi kontrol' v obrazovanii: uchebnoe posobie* (Test Control in Education: Textbook), Moscow: Universitetskaya Kniga, 2007.

14. Karpenko, A.P. and Sokolov, N.K., The Semantic Net Complexity Measures in Learning System, *Vestn. MGTU, Ser. Priborostr.*, 2009, no. 1(74), pp. 50–66.

15. Karpenko, M.P., The Problem of Measurement of Knowledge and Education Technologies, *Zh. Prakt. Psikhol.*, 1997, no. 4, pp. 74–79.

16. Kravtsov, L.G., Methodological Problems in Psychologic Analysis of Thought in Notions, *Mater. I Ros. konf. kogn. nauke* (Mat. 1st Rus. Conf. Cogn. Sci.), Kazan: Kazan. Gos. Univ., 2004, URL: http://www.ksu.ru/ss/cogsci04/science/cogsci04/sod.php3 (Cited November 22, 2013).

17. Krasnykh, V.V., *Osnovy psikholingvistiki i teorii kommunikatsii* (Fundamentals of Psycholinguistics and Theory of Communication), Moscow: Gnosis, 2012.

18. Lerner, I.Ya., Complexity Factors in Cognitive Problems, *Nov. Issled. Pedagog. Nauk.*, 1970, no. 1, pp. 86–91.

19. Luger, G.F., *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, London: Addison-Wesley, 2002. Translated under the title *Iskusstvennyi intellekt: strategii i metody resheniya slozhnykh problem*, Moscow: Vil'yams, 2005.

20. Mikk, Ya.A., *Optimizatsiya slozhnosti uchebnogo teksta: v pomoshch' avtoram i redaktoram* (Optimization of Complexity of the Educational Text: For the Aid to Authors and Editors), Moscow: Prosveshchenie, 1981.

21. Mitrofanova, O.A., Semantic Distances: Problems and Prospects, *XXXIV Mezhd. filolog. konf: Vypusk 21. "Prikladnaya i matematicheskaya lingvistika"* (34 Int. Phil. Conf.: Issue 21. "Applied and Mathematical Linguistics"), St.-Petersburg, 2005, pp. 59–63.

22. *Multifunctional Linguistic Processor ETAP-3*, URL: http://www.iitp.ru/ru/science.

23. Naikhanova, L.V., *Tekhnologiya sozdaniya metodov avtomaticheskogo postroeniya ontologii s primeneniem geneticheskogo i avtomatnogo programmirovaniya* (Technology of Creating Methods for Automatic Construction of Ontology Using Genetic and Automata-Based Programming), Ulan-Ude: BNTs,, 2008.

24. *Novaya filosofskaya entsiklopediya* (New Encyclopaedia of Philosophy), 4 vols., Moscow: Mysl', 2010, URL: http://iph.ras.ru/enc.htm (Cited April 9, 2014).

25. *Noveishii filosofskii slovar'* (Newest Philosophical Dictionary), Minsk: Knizhnyi Dom, 2003, 3rd ed.

26. Pogorelov, A.V., *Geometriya: uchebnik dlya 7–11 klassov obshcheobrazovatel'nykh uchrezhdenii* (Geometry: Textbook for Secondary School), Moscow: Prosveshchenie, 1995.

27. Popov, E.V., *Obshchenie c EVM na estestvennom yazyke* (Natural Language Interaction with Computers), Moscow: Nauka, 1982.

28. Romadina, O.G. and Rakitina, E.A., Procedure of the Estimation of Complexity of Educational Problems on Computer Science, *Vopr. Sovrem. Nauki Prakt.*, 2010, no. 10–12(31), pp. 146–151.

29. Ryzhenko, N.G., Graphic Modeling as a Method for Defining Complexity of Solutions to School Text Problems, *Matem. Inform.*, Omsk, 2001, issue 1, pp. 99–103.

30. Fedotov, I.E., *Nekotorye priemy parallel'nogo programmirovaniyz: uchebnoe posobie* (Some Methods of Parallel Programming: Textbook), Moscow: MGIREA, 2008.

31. Filippovich, Yu.N. and Prokhorov, A.V., *Semantika informatsionnykh tekhnologii: opyty slovarno-tezaurusnogo opisaniy* (Semantics of Information Technologies: Experience of Dictionary-Thesaurus Description), Moscow: MGUP, 2002.

32. Collins-Tompson, K., Bennett, P., White, R., Chia, S., and Sontag, D., Personalizing Web Search Results by Reading Level, *Proc. 20th ACM Int. Conf. on Information and Knowledge Management*, New York, 2011, pp. 403–412.

33. Dubay, W., *The Principles of Readability*, Costa Mesa: Impact Information, 2004.

34. Hartley, R.V.L., Transmission of Information, *Bell Syst. Tech. J.*, 1928, July, pp. 535–563.

35. Leacock, C. and Chodorow, M., Combining Local Context and WordNet Similarity for Word Sense Identification, in *WordNet: An Electronic Lexical Database*, Fellbaum, C., Ed., Boston: MIT Press, 1998, pp. 265–283.

36. Patwardhan, S., Banerjee, S., and Pedersen, T., Using Measures of Semantic Relatedness for Word Sense Disambiguation, *Proc. 4th Int. Conf. on Intelligent Text Processing and Computational Linguistics*, 2003, pp. 241–257.

37. Shannon, C.E., A Mathematical Theory of Communication, *Bell Sys. Tech. J.*, 1948, vol. 27, pp. 379–423, 623–656.

38. Anisimovich K.V., Druzhkin K.Ju., Minlos F.R., Petrova M.A., Selegey V.P., and Zuev K.A., Syntactic and Semantic Parser Based on ABBYY Compreno Linguistic Technologies, in *Kom'yuternaya lingvistika i intellektual'nye technologii* (Computer Linguistics and Intellectual Technologies), 2012, pp. 810–822, URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Anisimovich.pdf (Cited December 2, 2013).