CSITA AUTOMATION

ISSN 2414-9055

UDC 004.82:322(075):621.311.004.13 https://doi.org/10.31721/2414-9055.2019.5.1.24

EFFICIENCY OF THESAURUS OF FACTUAL COLLOCATIONS OF PROFESSIONAL ONTOLOGY OF LINGUISTIC CORPUS OF POWER SYSTEM ACCIDENTS

Kotov I.A., PhD, Associate Professor,
Tron V.V., PhD, Associate Professor,
Serdiuk O.Y., Assistant,
Pylypenko O.V., Assistant
Kryvyi Rih National University

Abstract. The object of the research is functioning of smart systems of decision-support in accident control of power systems on the basis of professional ontologies of factual models and specialized thesauruses. Methods of efficiency indices of the professional ontology thesaurus of factual knowledge of accident elimination in the power systems are under study.

The following *methods* have been used in the course of the research: analysis of national and foreign experience and systemization of available approaches and methods, methods of mathematical statistics, the graph theory, formal systems and languages, methods of system analysis, methods of processing experiment results, methods of synthesis and analysis of mathematical models, methods of artificial intelligence, methods of computer simulation and power system control.

The scientific novelty consists in developing and applying the efficiency factor of the professional thesaurus ontology based on factual representation of knowledge of emergency management in a power system. The obtained factor allows assessing the expertise rate of the linguistic corpus in various scientific and technical areas and building efficient factual knowledge bases. Despite current approaches to building knowledge bases, the suggested approach allows unifying development of decision-support systems and reducing their implementation time.

The practical significance of the studies involves assessing efficiency of applying professional knowledge in a factual form to building unified smart systems of decision-support systems for emergency control of power systems.

The results of the research include elaboration of the formal model of professional ontology of factual knowledge and the thesaurus of professional terms and slang. The model of assessing efficiency of the professional thesaurus is suggested on the basis of the ratio between intensities of growing general and specific professional vocabularies. The conclusion is drawn that it is expedient to use professional thesauruses to increase efficiency of knowledge bases of smart systems. In case of computer implementation, it is recommended to switch from alphabetical knowledge representation to the professional-hieroglyphic one. The research results demonstrate increased efficiency resulted from transition to more specific specialized scientific and technical areas.

Key words: ontology, thesaurus, vocabulary, power system, dispatch control, collocation, formal language.

Introduction

Emergencies occurring in complex industrial systems are always accompanied by considerable damages [1-3]. The human factor is of great importance here as employees are under huge psycho-physiological pressure and not always able to react to an emergency in a proper way [1-3].

A modern power-system (PS) is a large complex hierarchical control object characterized by simultaneous generation, distribution and consumption of power. Thus, computer-aided smart decision-making support of operation dispatch personnel (ODP) becomes of particular significance. Decision-support systems (DSSs) are integrated into operational data management complexes (ODMCs) of the automated dispatch control system (ADCS) - SCADA [4].

One of the targets of building and introducing the DMSS is to develop an adaptive model of representing professional vocabulary [5, 6]. The ontology apparatus can serve as a unified model of professional knowledge [7]. Professional thesaurus is the ontology platform.

Capacity of the DSS is mostly determined by efficiency of professional thesaurus application. This arouses the problem of studying and elaborating criteria for assessing efficiency of applying professional vocabulary thesaurus in DSS ontologies to control emergency conditions in power systems. Considering specific features of the professional area, factual collocations are chosen as a form of knowledge presentation, and the linguistic corpus of accident elimination and prevention in the power system is taken as a professional vocabulary source [8].

The research is aimed at developing a mathematical model of professional thesaurus of factual collocations and elaborating assessment criteria for its efficiency within the ontology model. The developed criteria can reveal efficiency of application of professional thesauruses to reduction of knowledge bases and increase of the

DSS rates.

The research aim conditions the necessity of solving the following tasks:

1. developing a formal-logical model of factual collocation ontology;
2. building a factual knowledge base determined by the subset of the linguistic corpus of accident elimination and prevention in the power system;
3. creating a general thesaurus of professional vocabulary;
4. building a specialized thesaurus of professional terms and slang;
5. providing statistic processing of initial lexical sampling and professional thesauruses;
6. developing assessment criteria for efficiency of the factual collocation thesaurus of professional ontology of the linguistic corpus of accident elimination in the power system;
7. demonstrating practical applicability, value and significance of the developed models and criteria for assessing efficiency of thesauruses.

Many national and foreign scholars have accumulated experience of developing theoretical models and implementing forms of knowledge-bases representation including A.A. Bashlykov, V.N. Vagin, A.G. Vendelin, V.A. Gelovani, L.G. Yevlanova, V.S. Kretov, O.I. Larichev, Yu.Ya. Lyubarskiy, Zh.L. Loryer, Dzh.F. Lyugger, D.A. Pospelov, V.D. Samoylov, K. Taunsend, P.V. Terelyanskiy, Dzh. Ulman, D. Waterman, P.K. Fishbern, F. Forsayt, Yu.P. Chaplinskiy. The works by S.A. Barkalov, Ya.D. Barkin, P.I. Bartolomey, A.A. Bashlykov, R.N. Berdnikov, A.F. Butkevich, A.M. Glazunova, S.O. Grishanov, Yu.Ya. Lyubarskiy, M.Sh. Misrikhanov, D.A. Panasetskiy, G.Ye. Pospelov, V.M. Cheban, etc. deal with intellectualization of controlling power-system conditions [9-15].

Analysis of researches and publications confirms the topicality of the problem chosen. It can be concluded that there are no unified solutions for presenting professional knowledge and assessing efficiency of specialized thesauruses of ontologies.

Materials and Methods

In developing professional ontologies, the major problem is about choosing and implementing a relevant form of knowledge representation. The professional area under study is noted for deep structuring and hierarchy of linguistic blocks and concepts. Besides, an instruction dispatch material was used as an initial expert linguistic corpus [8]. That is why, to solve the set problems, a factual form of representing professional ontologies was chosen.

Activation or actualization of a fact is treated as formal actualization of its components, attaching a sign of activity to them. In the functioning of a smart system, activation (fact actualization) is putting a code of a fact into a so called “working area” or a “notice board”.

The rule of collocating the structure of an elementary fact is the following. The elementary fact *f* is a triplet of atomic statements (lexemes of the specialized thesaurus) treated as an isolated directed graph in any operations within the operating formal system. While interpreting facts each of triplet elements is treated as a single linguistic constant or a value of a linguistic variable.

We use formal models of atomic statements making an active thesaurus without any relations with each other as a basis for representing facts. To create facts, it is necessary to make triplets of related atomic statements and consider them atomic lexemes of fact triplets. Let us present the ontology of atomic lexemes in a general form as

$$O_{KB_S} =< \bigcup_{j=1}^{N}[S_{1j}^{c_0} \cup S_{2j}^{c_0}], \varnothing, \{F\} >, \quad (1)$$

where $S_{1j}^{c_0} \bigcup S_{2j}^{c_0}$ is a combination of sets of atomic interpreted and interpreting lexemes related to the *j*-th context; $\varnothing$ is an empty set of links of atomic lexemes in the ontology model; *F* is a set of functions of ontology interpretation.

In expression (1), the sets $S_{1j}^{c_0}$ and $S_{2j}^{c_0}$ are classes of statements in the terminal alphabet A_t

$$\Sigma = A_t = \{\varepsilon\} \cup A_l \cup A_d \cup A_s \cup A_p \cup A_{sl} \cup A_{ab}, \quad (2)$$

where A_l is symbols of general vocabulary; A_d is figures; A_s is a set of specialized symbols of general vocabulary; A_p is a set of symbols of specialized professional vocabulary, specialized signs; A_{sl} is a set of symbols of specialized professional terms and slang; A_{ab} is a set of symbols of abbreviations

ISSN 2414-9055

of the professional area.

Let us build a fact ontology. We define sets of concepts for the elementary fact model. It is required to use only the atomic statements based on the alphabet (2). For instance, for the c_i-*th* context we use:

$$S^{c_j} = \left\{ s_k^{c_j} \mid k = 1, n_s \right\},$$

where n_s is the number of atomic statements of the c_i-*th* context.

Согласно принятой концепции триплета, графическая модель элементарного факта f_i для контекста c_j может быть представлена как орграф (далее – граф) следующего вида (Fig. 1). According to the established triplet concept, the graphic model of the elementary fact f_i for the context c_i can be presented as a directed graph (hereinafter referred as a graph) which looks as follows (Fig.1)

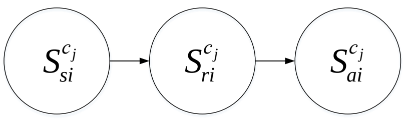


Figure 1. The graphic model of the fact f_i

The figure specifies S_{si} as the atomic statement – the relation source, S_{ri} as the atomic statement – the relation, S_{ai} as the atomic statement - the relation receiver.

This graphic interpretation of a fact can be presented by the following formal tuple

$$f_i^{c_j} = < S_{si}^{c_j}, S_{ri}^{c_j}, S_{ai}^{c_j} >, \qquad (3)$$

or in the theoretical-multiple interpretation

$$f_i^{cj} = \left\{ s_{si}^{c_j}, s_{ri}^{c_j}, s_{ai}^{c_j} \right\}. \qquad (4)$$

The subset of facts F^{cj} related to the context c_j, of the general fact set F will be presented as follows

$$f_i^{cj} \in F^{cj},$$

$$F^{cj} \subseteq F.$$

with

$$f_i^{cj} = \left\{ s_{si}^{c_j}, s_{ri}^{c_j}, s_{ai}^{c_j} \right\},$$

where $s_{si}^{c_j} \in S^0, s_{ri}^{c_j} \in S^0, s_{ai}^{c_j} \in S^0$.

On the basis of the established theoretical and multiple model of an elementary fact, we formalize its graph model [16]. The elementary fact f_i in its general form is presented by the graph

$$G_i^f = \left\{ V(G)_i^f, A(G)_i^f \right\}, \qquad (5)$$

where $V(G)_i^f = f_i^{cj}$ is a set of atomic statements forming a fact, $f_i^{cj} = \left\{ s_{si}^{c_j}, s_{ri}^{c_j}, s_{ai}^{c_j} \right\}$, $\left| V(G)_i^f \right| = 3$; $A(G)_i^f$ is a set of arcs (links), $\left| A(G)_i^f \right| = 2$.

Besides, for the fact, let us introduce private interpretation of following tops depending on the task conditions, which in a general case can be arbitrary:

$$I^f = (P_s^f, P_e^f), \qquad (6)$$

where P_s^f is the incidentor of starting statements of fact links $P_s^f(s_{si}^{c_j}, s_{ri}^{c_j}) = s_{si}^{c_j}$ и $P_s^f(s_{ri}^{c_j}, s_{ai}^{c_j}) = s_{ri}^{c_j}$;

P_e^f is the incidentor of final statements of fact links

$P_e^f(s_{si}^{c_j}, s_{ri}^{c_j}) = s_{ri}^{c_j}$ и $P_e^f(s_{ri}^{c_j}, s_{ai}^{c_j}) = s_{ai}^{c_j}$.

Now, we can provide a general formal model of the elementary fact graph

$$\begin{aligned} G_i^f &= \{V(G)_i^f, A(G)_i^f, I^f\}, \\ G_i^f &= \{V(G)_i^f, A(G)_i^f, P_s^f, P_e^f\}. \end{aligned} \qquad (7)$$

We develop the formal language of the knowledge base using facts. As interpretation of fact semantics depends on the context, groups (classes) of facts should be considered while forming the formal language. We determine the set of contexts to interpret facts

$$C^f = \left\{ c_i \mid i = 1, n_c \right\}, \qquad (8)$$

where n_c is the number of contexts (subject areas).

The set of classes of contexts is

$$G^{fc_i} = \left\{ g_m^{fc_i} \mid m = 1, n_g \right\}, \qquad (9)$$

where n_g is the number of classes of contexts.

The subset of elementary facts of the knowledge base of the smart system related to the c_i-*th* context is:

$$F^{c_i} = \{ f_k^{c_i} \mid k = 1, n_f \}, F^{c_i} \subset F, \qquad (10)$$

where n_f is the number of elementary facts related to the c_i-*th* context.

Considering classification of facts by subsets

F^{c_i} according to features (contexts) $g_m^{fc_i}$ it can be written as

$$F^{c_i} = \left\{ \left\{ F^{c_i}_{1g^{c_i}_m} \right\}, \left\{ F^{c_i}_{2g^{c_i}_m} \right\}, \dots, \left\{ F^{c_i}_{mg^{c_i}_m} \right\}, \dots, \left\{ F^{c_i}_{n_g g^{c_i}_m} \right\} \right\}, \quad (11)$$

where $\left\{ F^{c_i}_{mg^{c_i}_m} \right\}$ is the class of facts corresponding to the classifying feature $g_m^{fc_i}$.

If we consider the fact that all elementary facts have no duplicates, the properties of elementary facts of the knowledge base should be the following:

$$F^{c_i} = F^{c_i}_{1g^{c_i}_m} \cup F^{c_i}_{2g^{c_i}_m} \dots \cup F^{c_i}_{mg^{c_i}_m} \dots \cup F^{c_i}_{n_g g^{c_i}_m} = \bigcup_{m=1}^{n_g} F^{c_i}_{mg^{c_i}_m}, \quad (12)$$

$$F^{c_i}_{1g^{c_i}_m} \cap F^{c_i}_{2g^{c_i}_m} \dots \cap F^{c_i}_{mg^{c_i}_m} \dots \cap F^{c_i}_{n_g g^{c_i}_m} = \bigcap_{m=1}^{n_g} F^{c_i}_{mg^{c_i}_m} = \varnothing, \quad (13)$$

$$\forall F^{c_i}_{mg^{c_i}_m} \left(F^{c_i}_{mg^{c_i}_m} \subseteq F^{c_i} \right), \quad (14)$$

$$\forall F^{c_i}_{pg^{c_i}_m} \forall F^{c_i}_{qg^{c_i}_m} \left(F^{c_i}_{pg^{c_i}_m} \neq F^{c_i}_{qg^{c_i}_m} \right). \quad (15)$$

The generalized syntactic interpretation of the elementary fact is given below:

<fact> ::= < statement > < statement > < statement >;

< statement > ::= <lexeme > | < statement > < lexeme >.

We determine the formal language of the model for representing elementary facts of the professional area. We consider the fact that the formal language of facts corresponds to a certain subject area, i.e. they belong to some of fact classes G^{c_i} in the current context

$$\forall F^{c_i}_k, k = 1, n_f \left(\bigvee_{m=1}^{n_g} f^{c_i}_k \in F^{c_i}_{m\, g^{c_i}_m} \right). \quad (16)$$

The language of elementary facts for the subject area $c_i \in C$ and some m-th class $g^{c_i}_m$ is determined in the following way

$$L(G)^{fc_i}_{g^{c_i}_m} = \langle \Sigma^f, N^f, P^f, S^f \rangle, \quad (17)$$

where G is formal grammar based on facts; Σ^f is the basic final terminal alphabet of facts; N^f is the auxiliary final non-terminal alphabet; P^f is rules of substitution (production) of formal grammar based on facts: $\exists a, \exists b, (a,b) \in P : a \to b$; S^f is the starting non-terminal symbol of grammar G based on facts; $N^f \cap \Sigma^f = \varnothing$ and $P \subset ((N^f \cup \Sigma^f)^+ \times (N^f \cup \Sigma^f)^*)$

We determine rules of formal grammar P based on facts for the language $L(G)^{fc_i}_{g^{c_i}_m}$

$f \to$ <lexeme>< lexeme >< lexeme >
< lexeme > $\to$ < lexeme >|<statement>.

We extend the language of facts to all classes of contexts of factual knowledge bases.

$$L(G)^f = \langle \Sigma^f, N^f, P^f, S^f \rangle, \quad (18)$$

where $\Sigma^f = F = \{f_k \mid k = 1, n_f\}$ are all the facts of the knowledge base level; $F = \left\{ \left\{ F_{1g^{c_0}_m} \right\}, \left\{ F_{2g^{c_0}_m} \right\}, \dots, \left\{ F_{mg^{c_0}_m} \right\}, \dots, \left\{ F_{n_g g^{c_0}_m} \right\} \right\}$; N^f = {fact, class_of facts, layer_of facts}; S^f = <layer_of facts>.

We extend the rules of formal grammar to the whole level of elementary facts considering the fact that there are three non-terminals in the developed formal language $L(G)^f$

$S^f \to$ <fact>;
$S^f \to$ < fact ><class_of facts>;
<class_of facts > $\to \forall g_m (\{F_{mg^{c_0}_m}\} \subseteq F)$;
< fact > $\to \forall f_k (f_k \in F)$.

The structural-linguistic model of the ontology is provided for the level of the knowledge base of elementary facts KB_F. We use the generalized formula of the ontology:

$$O_{KB_F} = \langle X^f, R^f, F^f \rangle. \quad (19)$$

For the facts related to the arbitrary context, we have

$$X^f = F = \{f_k \mid k = 1, n_f\} = \{\{s_{sk}, s_{rk}, s_{ak}\} \mid k = 1, n_f\}.$$

i.e.

$$X^f = \{f_k \mid k = 1, n_f\} = \{\{s_{sk}, s_{rk}, s_{ak}\} \mid k = 1, n_f\}.$$

As for the level of the knowledge base under study there are no links between the facts, i.e. they are isolated, $R = \varnothing$.

To determine the set of functions of interpretation F^f let us assume that part of the facts can be used to interpret other facts of the current level. In this case, facts can be divided into situational groups (classes) - the subset of interpreted facts (with the index 1) and the subset of interpreting facts (with the index 2) as is given in (20):

$$F = \{\{F_1\}, \{F_2\}\}, \quad (20)$$

ISSN 2414-9055

where $F_1 \cup F_2 = F$ is the whole set of facts, $F_1 \cap F_2^0 = \varnothing$.

Then

$$\exists(f_{1i} \in F_1), \exists(f_{21}, f_{22}, ..., f_{2k} \in F_2) \\ (f_{1i}^{c_0} = f^f(f_{21}^{c_0}, f_{22}^{c_0}, ..., f_{2k}^{c_0}), f^f \in F^f) . \quad (21)$$

The interpretation function will look like

$$f_j : Op(\{(f_{2j}, I_j)\}) \rightarrow (f_{1j}, I_j), \quad (22)$$

where *Op* is the operation of fact aggregation like enumeration with regulation, choice of the most probable fact, choice of the most significant fact, choice of the most recent fact, logical linking, logical implication, etc.

On the basis of the developed theoretical and multiple models, there is a formal model of unified ontology of facts to obtain hierarchy of professional ontologies of the active vocabulary type along with the generalized structure:

$$O_{KB_F} =< \bigcup_{j=1}^{N} [F_{1j} \cup F_{2j}], \varnothing, \{F^f\} > . \quad (23)$$

Practical application of the developed mathematical models of fact representation and professional ontology models is demonstrated by sampling of characteristics of electric equipment. We introduce sets of statements conditionally related to the same context c^0:

$s_1^{c_0}$ = «line»; $s_2^{c_0}$ = «belongs to nominal voltage class»; $s_3^{c_0}$ = «110kV»; $s_4^{c_0}$ = «nominal voltage class»; $s_5^{c_0}$ = «is of value»; $s_6^{c_0}$ = «6-10kV»; $s_7^{c_0}$ = «35kV»; $s_8^{c_0}$ = «transformer»; $s_9^{c_0}$ = «the number of coils is»; $s_{10}^{c_0}$ = «2»; $s_{11}^{c_0}$ = «transformer coil»; $s_{12}^{c_0}$ = «is made of»; $s_{13}^{c_0}$ = «copper»; $s_{14}^{c_0}$ = «the cooling type is»; $s_{15}^{c_0}$ = «oil»; $s_{16}^{c_0}$ = «air».

We introduce sets of facts indicating the relevance index: $F^{c_0} = \{[\{F_1^{c_0}\}, \{F_2^{c_0}\}]\}$. We build the set of facts

$$F = \{f_1, f_2, ..., f_8,\},$$

where $f_1 = (s_1^{c_0}, s_2^{c_0}, s_3^{c_0}, 1)$; $f_2 = (s_4^{c_0}, s_5^{c_0}, s_3^{c_0}, 1)$; $f_3 = (s_4^{c_0}, s_5^{c_0}, s_6^{c_0}, 1)$; $f_4 = (s_4^{c_0}, s_5^{c_0}, s_7^{c_0}, 1)$; $f_5 = (s_8^{c_0}, s_9^{c_0}, s_{10}^{c_0}, 2)$; $f_6 = (s_{11}^{c_0}, s_{12}^{c_0}, s_{13}^{c_0}, 2)$; $f_7 = (s_{11}^{c_0}, s_{14}^{c_0}, s_{15}^{c_0}, 2)$; $f_8 = (s_{11}^{c_0}, s_{14}^{c_0}, s_{16}^{c_0}, 2)$; $F_1 = \{f_1, f_5\}$, $F_2 = \{f_2, f_3, f_4, f_6, f_7, f_8\}$.

In this case, there are the following interpretation functions:

$$f_1^f : Op\left(\left\{(f_2^{c_0}, 1), (f_3^{c_0}, 1), (f_4^{c_0}, 1)\right\}\right) \rightarrow (f_1^{c_0}, 1);$$

$$f_2^f : Op\left(\left\{(f_6^{c_0}, 2), (f_7^{c_0}, 2), (f_8^{c_0}, 2)\right\}\right) \rightarrow (f_5^{c_0}, 2).$$

Thus, the structural-linguistic model of the unified professional ontology of elementary facts is built. The applied mathematical apparatus is invariant as to the professional areas and allows building and controlling factual knowledge bases.

Efficiency of the obtained factual ontology model is verified by building its thesaurus. The limited linguistic corpus concerning accident prevention and elimination in the electrical part of power stations and power grids is chosen for sampling of lexical blocks of factual collocations [8]. The facts are based on fixed linguistic forms of the professional area. The obtained collocations enable a general thesaurus of the lexical corpus. In its turn, the general corpus provides the basis for specialized terms and slang. The volume of the factual knowledge base is 63 facts. Basic characteristics of the professional lexical corpus and thesauruses are provided in Table 1. Several initial and final lines are given to save space.

Table 1. Basic characteristics of the professional lexical factual corpus and the fact thesaurus

N	V_F	V_F^+	V_{FT}	V_{FT}^+	V_{FTA}	V_{FTA}^+	N_{FW}	N_{FW}^+	N_{FWA}	N_{FWA}^+	N_{FWTA}	N_{FWTA}^+	D_p
1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	7	37	37	37	26	37	4	4	2	2	2	2	50.00
2	28	65	15	52	0	37	3	7	1	3	0	2	42.86
3	31	96	24	76	13	50	5	12	2	5	2	4	41.67
4	23	119	5	81	5	55	4	16	2	7	1	5	43.75
5	22	141	0	81	0	55	4	20	2	9	0	5	45.00
6	49	190	33	114	33	88	6	26	4	13	2	7	50.00
7	54	244	13	127	0	88	8	34	6	19	0	7	55.88

ISSN 2414-9055

8	32	276	0	127	0	88	6	40	3	22	0	7	55.00
9	68	344	25	152	25	113	10	50	6	28	3	10	56.00
10	52	396	16	168	28	141	6	56	5	33	2	12	58.93
…	…	…	…	…	…	…	…	…	…	…	…	…	…
54	111	2564	0	733	0	441	17	356	7	196	0	45	55.06
55	46	2610	19	752	7	448	6	362	3	199	1	46	54.97
56	45	2655	14	766	0	448	7	369	4	203	0	46	55.01
57	76	2731	31	797	0	448	13	382	5	208	0	46	54.45
58	39	2770	26	823	15	463	5	387	4	212	1	47	54.78
59	57	2827	13	836	0	463	7	394	5	217	0	47	55.08
60	54	2881	11	847	11	474	8	402	6	223	1	48	55.47
61	37	2918	21	868	0	474	5	407	4	227	0	48	55.77
62	31	2949	0	868	0	474	4	411	3	230	0	48	55.96
63	34	2983	0	868	0	474	4	415	3	233	0	48	56.14

Table 1 contains the following conventional signs: N is the fact number in the knowledge base; V_F is the fact volume, symbol; V_F^+ is the volume of facts progressively, symbol; V_{FT} is the specific volume of the general thesaurus per fact, symbol; V_{FT}^+ is the volume of the general fact thesaurus progressively, symbol; V_{FTA} is the specific volume of the slang and abbreviation thesaurus per fact, symbol; V_{FTA}^+ is the volume of the slang and abbreviation thesaurus progressively, symbol; N_{FW} is the specific number of lexemes per fact, words; N_{FW}^+ is the number of lexemes progressively, words; N_{FWA} is the specific number of slang and abbreviation lexemes progressively, words; N_{FWA}^+ is the number of slang and abbreviation lexemes progressively, words; N_{FWTA} is the number of slang and abbreviation words per fact, words ; N_{FWTA}^+ is the number of slang and abbreviation lexemes progressively, words; D_p is the degree of proficiency of the fact base, %.

Table 1 shows that sampling can be performed both in lexemes and symbols or codes. Fig. 1 contains lexeme sampling of the linguistic corpus of accident elimination in the power system ([8], 12).

Объем - Volume of fact base and thesaurus, words

объемы выборки лексем - Volumes of lexeme sampling of fact base and professional thesaurus

Выборка - Sampling of fact base

колич. лексем - number of lexemes per fact

колич. сленга - number of slang per fact

Figure 1. Dynamics of the lexeme sampling of the knowledge base progressively

Fig. 1 demonstrates that the growing volume of lexemes of the factual knowledge base exceeds that of the thesaurus. It is mostly explained by specific features of the problem area and vast application of the professional slang.

In building automated systems of processing texts and knowledge bases of smart systems, samplings in some codes, symbols and bytes are of particular interest. Fig.2 reveals results of absolute sampling in symbols (letters) progressively in knowledge base facts of the general and specialized professional thesauruses.

ISSN 2414-9055

Объем - Volume of fact base and thesaurus, symbol

Абсол. объемы - Absolute volumes of sampling of fact base and thesaurus

Выборка - Sampling of fact base

исходная выборка - initial sampling progressively

общ. - general thesaurus progressively

проф. - professional thesaurus progressively

Figure 2. Dynamics of the absolute sampling of knowledge base facts and thesauruses progressively

Fig. 2 shows that the specific professional thesaurus of the slang and abbreviations increases less intensively which indicates its increased efficiency and relevance.

On the basis of the obtained data presented in Table 1, a factor assessing expertise of the factual knowledge base is suggested:

$$D_p = \frac{N^+_{FWTA}}{N^+_{FW}} 100\% . \quad (24)$$

Особенностью этого показателя является учет структуры фактов и специфики семантической нагрузки каждого из трех концептов факта. This factor is noted for consideration of the fact structure and peculiarity of semantic loads of each of the three fact concepts. Fig. 3 depicts dynamics of changes in the expertise factor of the factual knowledge base.

Степень - Expertise rate of fact base
доля - percentage of professional slang, %
Выборка - Sampling of fact base
степень проф. - expertise rate

Figure 3. The expertise factor of the factual knowledge base progressively

As is seen, the factor is quite fast to stabilize versus the average value. The factor characterizes availability of specific professional vocabulary and slang in structured forms of representing knowledge. Interpretation of the chart in Fig. 3 indicates that in the specific vocabulary of instruction materials on accident elimination and prevention in power systems structured as facts of knowledge representation, the percentage of the specific slang and abbreviations is above 55%.

To provide numerical assessment of efficiency of professional thesauruses as facts, experiment and measurement data of knowledge bases are given in Table 2.

There are several variants of efficiency indices of thesauruses.

The first factor assesses absolute efficiency expressed in symbols (or bytes) as redundancy eliminated by applying either the general or the specialized thesaurus:

$$\begin{aligned} E^{abs}_{FT} &= \Delta V_{F-FT} = V^+_F - V^+_{FT}, \\ E^{abs}_{FTA} &= \Delta V_{F-FTA} = V^+_F - V^+_{FTA} \end{aligned} . \quad (25)$$

Fig. 4 illustrates absolute efficiency of thesauruses.

ISSN 2414-9055

Абс. эфф. тез. - Absolute efficiency of thesauruses
Абс. эфф. Символ - Absolute efficiency, symbol
Выборка - Sampling of fact base
Абс. эфф. общ. - Absolute efficiency of general thesaurus
Абс. эфф. проф - Absolute efficiency of professional thesaurus

Figure 4. Dynamics of absolute efficiency indices of the general and specialized thesauruses

Absolute values of thesaurus efficiency indicate stable growth of redundancy of the initial linguistic corpus as compared to vocabulary. However, thesauruses also increase in volumes. That is why, the ratio of absolute efficiency and the volume of the fact base progressively is of particular interest. Thus, another efficiency index, the relative efficiency index, is used in the research.

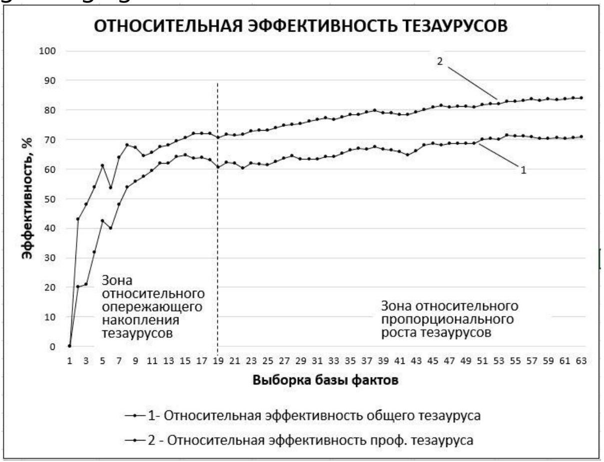
$$E_{FT}^{rel} = \frac{E_{FT}^{abs}}{V_F^+} * 100\% = \frac{V_F^+ - V_{FT}^+}{V_F^+} * 100\%, \quad E_{FTA}^{rel} = \frac{E_{FTA}^{abs}}{V_F^+} * 100\% = \frac{V_F^+ - V_{FTA}^+}{V_F^+} * 100\% . \quad (26)$$

Table 2. Calculation data and relative efficiency indices of the fact thesaurus

N	V_F^+	V_{FT}^+	V_{FTA}^+	ΔV_{F-FT}	ΔV_{F-FTA}	$P_{FT/F}^+$	$P_{FTA/F}^+$	E_{FT}^{rel}	E_{FTA}^{rel}
1	2	3	4	5	6	7	8	9	10
1	37	37	37	0	0	100	100	0	0
2	65	52	37	13	28	80	57	20	43
3	96	76	50	20	46	79	52	21	48
4	119	81	55	38	64	68	46	32	54
5	141	81	55	60	86	57	39	43	61
6	190	114	88	76	102	60	46	40	54
7	244	127	88	117	156	52	36	48	64
8	276	127	88	149	188	46	32	54	68
9	344	152	113	192	231	44	33	56	67
10	396	168	141	228	255	42	36	58	64
...	...	...	...	...	...	...	...	...	...
54	2564	733	441	1831	2123	29	17	71	83
55	2610	752	448	1858	2162	29	17	71	83
56	2655	766	448	1889	2207	29	17	71	83
57	2731	797	448	1934	2283	29	16	71	84
58	2770	823	463	1947	2307	30	17	70	83
59	2827	836	463	1991	2364	30	16	70	84
60	2881	847	474	2034	2407	29	16	71	84
61	2918	868	474	2050	2444	30	16	70	84
62	2949	868	474	2081	2475	29	16	71	84
63	2983	868	474	2115	2509	29	16	71	84

Table 2 contains the following conventional signs: ΔV_{F-FT} is the difference of fact volumes progressively and the general thesaurus of facts progressively; ΔV_{F-FTA} is the difference of fact volumes progressively and the slang and abbreviation thesaurus progressively; $P_{FT/F}^+$ is the fraction of the general thesaurus of facts progressively in the fact base progressively; $P_{FTA/F}^+$ is the fraction of the slang and abbreviation thesaurus progressively in the fact base progressively; E_{FT}^{rel} is relative efficiency of the general fact thesaurus progressively; E_{FTA}^{rel} is relative efficiency of the slang and abbreviation thesaurus progressively.

Fig. 5 illustrates dynamics of relative efficiency of the thesauruses. The diagrams have two distinct zones - that of relative advance in thesaurus accumulation and that of relative proportionate growth of thesauruses. Efficiency of the specific slang and abbreviation thesaurus is higher than that of the general one. The zone of relative advance in thesaurus accumulation is characterized by accumulation of facts in a single terminological and semantic block while the thesaurus is formed rapidly and almost does not grow. The zone of relative proportionate growth of thesauruses indicates the increased volume of the linguistic corpus when new terminological and semantic blocks accumulate. In this case, the thesaurus starts growing again.



Относ. эфф.- Relative efficiency of thesauruses
Эффективность - Efficiency, %
зона относ. опереж. - Zone of relative advance in thesaurus accumulation
зона относ.пропорц.- Zone of relative proportionate growth of thesauruses
Выборка - Sampling of fact base
Относ. эф. общ.- Relative efficiency of general thesaurus
Относ. эф. проф. - Relative efficiency of professional thesaurus

Figure 5. Dynamics of growing relative efficiency indices of the general and specialized thesauruses

Relative efficiency indices of professional thesauruses are informative and adequately reflect the semantic and numerical character of the fact base of the linguistic corpus of accident elimination in the power system. Yet, these indices change when knowledge bases accumulate and comprise several specific areas. There is a necessity for a stable integral efficiency index of thesauruses. That is why, the research suggests the third index, the integral factor of thesaurus efficiency. It is calculated by linear approximation of absolute efficiency indices. Data on calculating the integral factor of efficiency are presented in Table 3.

Table 3. Calculation data of the integral efficiency factor of the fact thesaurus

V_{KB}	V_F^+	V_{FT}^+	V_{FTA}^+	N^2	NV_F^+	NV_{FT}^+	NV_{FTA}^+	$a_1V_{KB}+b_1$	$a_2V_{KB}+b_2$	$a_3V_{KB}+b_3$
1	2	3	4	5	6	7	8	9	10	11
37	37	37	37	1369	1369	1369	1369	37	83.4130	93.2097
65	65	52	37	4225	4225	3380	2405	65	91.1166	97.2376
96	96	76	50	9216	9216	7296	4800	96	99.6455	101.6970
119	119	81	55	14161	14161	9639	6545	119	105.9734	105.0056
141	141	81	55	19881	19881	11421	7755	141	112.0262	108.1703
190	190	114	88	36100	36100	21660	16720	190	125.5075	115.2191
244	244	127	88	59536	59536	30988	21472	244	140.3643	122.9871
276	276	127	88	76176	76176	35052	24288	276	149.1684	127.5903
344	344	152	113	118336	118336	52288	38872	344	167.8770	137.3723
396	396	168	141	156816	156816	66528	55836	396	182.1836	144.8525
…	…	…	…	…	…	…	…	…	…	…
2564	2564	733	441	6574096	6574096	1879412	1130724	2564	778.6591	456.7232
2610	2610	752	448	6812100	6812100	1962720	1169280	2610	791.3150	463.3404
2655	2655	766	448	7049025	7049025	2033730	1189440	2655	803.6957	469.8137
2731	2731	797	448	7458361	7458361	2176607	1223488	2731	824.6054	480.7465
2770	2770	823	463	7672900	7672900	2279710	1282510	2770	835.3353	486.3567
2827	2827	836	463	7991929	7991929	2363372	1308901	2827	851.0176	494.5563
2881	2881	847	474	8300161	8300161	2440207	1365594	2881	865.8744	502.3243
2918	2918	868	474	8514724	8514724	2532824	1383132	2918	876.0541	507.6468
2949	2949	868	474	8696601	8696601	2559732	1397826	2949	884.5831	512.1062
2983	2983	868	474	8898289	8898289	2589244	1413942	2983	893.9374	516.9971
Total	87139	87139	28588	18072	169539175	169539175	53026299			

ISSN 2414-9055

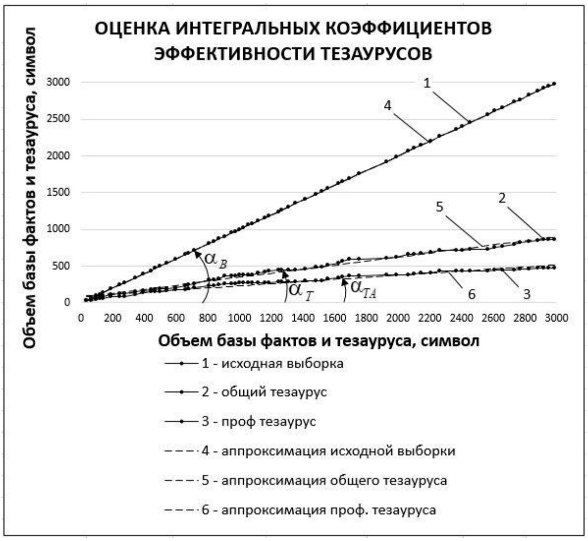
Table 4 reveals values of approximation parameters.

Table 4. Values of approximation parameters

a_1	b_1	a_2	b_2	a_3	b_3
1	2	3	4	5	6
1.0000	0.0000	0.2751	73.2334	0.1439	87.8873

In Fig. 6, there are graphical results of approximation and explanations for calculations. The physical sense of the integral factor of thesaurus efficiency implies the ratio of the slope angle of the approximated line of the thesaurus growth rate to that of the initial base. As sampling volumes are given axially in the same units and scales, one can assume that the slope angles of approximated lines will remain stable.

Formulae (26) and (27) are used to calculate integral efficiency factors of the general and professional thesauruses.



оценка интер. - Assessment of integral efficiency factors of thesauruses
объем базы фактов - Volume of the fact base and the thesaurus, symbol
1 - initial sampling
2 - general thesaurus
3 - professional thesaurus
4- approximation of initial sampling
5 - approximation of general thesaurus
6- approximation of professional thesaurus

Figure 6. Illustration of calculation of integral factors of thesaurus efficiency

$$K_{ET} = (1 - \frac{\alpha_T}{\alpha_B}) * 100\% = (1 - \frac{arctg(a_T)}{arctg(a_B)}) * 100\% \quad , \quad (26)$$

$$K_{ETA} = (1 - \frac{\alpha_{TA}}{\alpha_B}) * 100\% = (1 - \frac{arctg(a_{TA})}{arctg(a_B)}) * 100\% \quad , \quad (27)$$

where K_{ET} is the efficiency factor of the general thesaurus; K_{ETA} is the efficiency factor of the slang and abbreviation thesaurus; α_T is the slope angle of the approximated line for the general thesaurus; α_{TA} is the slope angle of the approximated line for the slang and abbreviation thesaurus; α_B is the slope angle of the approximated line for the fact thesaurus; a_T is the factor under V_{KB} in the line equation for the general thesaurus; a_{TA} is the factor under V_{KB} in the line equation for the slang and abbreviation thesaurus; a_B is the factor under V_{KB} in the line equation for the fact thesaurus.

Substitution of calculation values results in the following:

$$K_{ET} = \left(1 - \frac{arctg(0{,}2751)}{arctg(1)}\right) * 100\% = 65{,}82\% \quad ,$$

$$K_{ETA} = \left(1 - \frac{arctg(0{,}1439)}{arctg(1)}\right) * 100\% = 81{,}81\% \quad .$$

Thus, calculations result in the general thesaurus efficiency of 65.82%, that of the slang and abbreviation thesaurus making 81,81%. Consequently, efficiency of knowledge bases will be higher if the professional area of the DSS is more specialized. The professional area of accident and emergency control of the power systems is overspecialized which justifies the need for building its specialized thesauruses.

Results

1. The formal logical model of factual collocation ontology is developed.
2. The factual knowledge base of the linguistic corpus subset of accident elimination and prevention in the power system is built.
3. The general thesaurus of professional vocabulary of accident elimination and prevention in the power system is built;
4. The specialized thesaurus of professional terms and slang of accident elimination and prevention in the power system is built;

ISSN 2414-9055

5. The lexical sampling and professional thesauruses are statistically processed.
6. The absolute, relative and integral efficiency indices of the factual collocation thesaurus of the linguistic corpus of accident elimination in the power system are developed.
7. Practical significance, value and applicability of the developed models and assessment criteria for thesaurus efficiency are confirmed.

Conclusions.

Practical usefulness of the results obtained involves the following. The thesaurus volume is much smaller than the initial linguistic corpus, so creation, representation, perception and interpretation of facts by means of a thesaurus becomes much simpler. Thesaurus application systematizes building and use of knowledge bases in smart systems. The developed approach allows building unified systems of decision support for various professional areas and reducing the time required to develop decision-support systems, thus making them less expensive. Applying the ontology model to the professional area is expedient due to the specific structure of professional vocabulary. The obtained dependencies show that the increased volume of professional vocabulary increases text redundancy faster than the thesaurus volume does which is explained by the increased efficiency of the latter. It is obvious that for general vocabulary, the efficiency of ontology models will be reduced because of the necessity to use a larger thesaurus. That is why, ontology models are reasonable to use in specific professional areas noted for a wide range of professional terms (professional slang and abbreviations). The obtained results of the research are going to facilitate creation of ontology models for various forms of knowledge representation.

References

1. Avariynost v energosisteme Ukrainyi za god vyrosla vdvoye (2015), Retrieved from: https://economics.unian.net/energetics/1073586-avariynost-v-energosisteme-ukrainyi-za-god-vyirosla-vdvoe.html – 30.04.2015.
2. Besanger Y., Eremia M., Voropai N. (2013), Major grid blackouts: Analysis, classification, and prevention //Handbook of Electrical Power System Dynamics: Modeling, Stability, and Control. New Jersey:Wiley – IEEE Press, p. 789 –863
3. **Smolovik S.V.** (2008), Rol "chelovecheskogo faktora" v razvitii krupnyh sistemnyh avariy, Elektroenergetika, Vol. 1, No 1, 16-19.
4. Bartolomey P.I., Berdin A.S., Begalova E.H., Kryuchkov P.A., (2000), Problemy informatsionnogo obespecheniya zadach ASDU energosistem, Ekaterinburg: izd-vo UGTU-UPI, 16-19.
5. Morkun V.S. (2005). Adaptive systems of optimal control over technological processes. Kriviy Rih: Mineral.
6. Morkun V., Tron V., Paraniuk D. (2015), Method of automatic interpretation of information about the geological structure in the process of exploratory wells drilling. Metallurgical and Mining Industry, No 3, 45-48.
7. Samoylov Yu.V. (2017), Obzor programmnyih prilozheniy dlya realizatsii ontologicheskogo podhoda k upravleniyu znaniyami. Sbornik statey VI Mezhdunarodnoy nauchno-prakticheskoy konferentsii «Innovatsionnyie nauchnyie issledovaniya» – Moscow: MTsNS «Nauka i prosveschenie», 82-86
8. Instruktsiya DS-8 po predotvrascheniyu i likvidatsii tehnologicheskih narusheniy v elektricheskoy chasti elektrostantsiy i elektricheskih setey regiona Dneprovskoy ES (2008), Ministerstvo topliva i energetiki Ukrainyi. Gosudarstvennoe predpriyatie Natsionalnaya energeticheskaya kompaniya «Ukrenergo», Zaporozhe: Dneprovskaya elektroenergeticheskaya sistema.
9. Bashlyikov A.A., Eremeev A.P. (2017), Osnovyi konstruirovaniya intellektualnyih sistem podderzhki prinyatiya resheniy v atomnoy energetike: uchebnik, Moscow: INFRA-M.
10. Glazunova A.M., Kolosok I.N. (2015), Reshenie zadach dispetcherskogo upravleniya intellektualnyimi elektroenergeticheskimi sistemami na baze metodov otsenivaniya sostoyaniya. Energetika Rossii v XXI veke. Innovatsionnoe razvitie i upravlenie, Irkutsk, 1-8.
11. Negnevitsky M., Tomin N., Panasetsky D., Kurbatsky V. (2013), Intelligent Approach for Preventing Large-Scale Emergencies in Electric Power Systems. IEEE International Conference on Electric Power Engineering PowerTech, Grenoble, France, 16-20 June, 1-6.
12. Grishanov S.A., Kanashevich N.A. (2012), Realizatsiya ekspertnoy sistemyi dlya diagnostiki generatorov teplovyih elektricheskih stantsiy. Sbornik nauchnyih trudov X Mezhdunarodnoy nauchno-tehnicheskoy konferentsii molodyih uchenyih i spetsialistov v gorode, Kremenchug: KNU, 305-306.
13. Barkalov S.A., Dushkin A.V., Kolodyazhnyiy, S.A. (2017), Vvedenie v sistemnoe proektirovanie intellektualnyih baz znaniy – M.: Goryachaya liniya-Telekom.
14. Antamoshin A.N., Bliznova O.V., Bolshakov A.A. (2016), Intellektualnyie sistemyi upravleniya organizatsionno-tehnicheskimi sistemami, Moscow: Goryachaya liniya - Telekom.
15. Lyubarskiy Yu.Ya. (1990), Intellektualnyie informatsionnyie sistemyi – Moscow:Nauka.
16. Kasyanov V.N., Evstigneev V.A. (2003), Grafyi v programmirovanii: obrabotka, vizualizatsiya i primenenie – SPb.: BHV-Peterburg.