Model and technology of lingware support in tasks of the nuclear knowledge management system

A S Gavrilkina, O L Golitsyna and V A Fedorova

National Research Nuclear University MEPhI, 115409, Russia, Moscow, Kashirskoe shosse, 31

E-mail: asgavrilkina@yandex.ru

Abstract. Model and technology of lingware application in the tasks of the knowledge management are presented. The used model of thesauri paradigmatic relationships allows measuring semantic similarity of documents. An algorithm for compiling a semantic intersection of conceptual patterns based on the coinciding maximum principle is described. The semantic neighborhood construction in accordance with the value of semantic similarity measure and descriptors ranking allows using in the expansion of the search query the descriptors, whose semantics are closer due to the links, which reflect the common semantics of genus-species and associative relationships. An algorithm for thesauri used in automatic classification tasks is considered. The technology of conceptual hierarchical dynamic formation of structures is proposed. The proposed method of constructing a terminological network structure with hierarchical relationships for the collection of texts of a particular subject area is based on the sequential formation of lexicographical neighborhoods of "nuclear" terms, which are used as high-frequency nouns. To build up a conceptual structure a documental databases are used.

1. Introduction

Considering the knowledge management as a set of mechanisms for a selective and varying knowledge elements used in different times and in different circumstances [1], it is necessary to note that for the solution of this problem not only the knowledge but also the means (primarily linguistic) used for knowledge representation and transmission must be preserved. Accordingly, the technology and the means to monitor changes and adaptation means of knowledge representation should be done as a part of knowledge management systems (KMS). It will eventually provide:

- necessary and sufficient specification to a potential user so he/she can accurately find and use an object in his/her practical activity (in other circumstances different to it was created and described);
- continuous update, i.e. it must allow modifications that reflect the refinement and expansion of knowledge;
- identifiability within commonly accepted international, national, and sectoral classification systems, i.e. the attributes that describe an object of knowledge must conform to systematization practices of the professional community (for example, in the case of knowledge in the nuclear field, these are the recommendations that are issued by the IAEA unit for the management of nuclear knowledge).

Knowledge preservation and transfer are inevitably associated with their physical media representation as well as with the identification at the level of the whole sample and at the level of a separate knowledge fragment. For this purpose the following components are used in the information practice:

- conceptual and terminological system (glossaries, thesauri, ontologies) — the instruments of learning and fixing knowledge on the media;
- classification or categorization scheme providing atomic uniformity of the subject area based on goal, organizational or methodological concepts.

These components traditionally are a part of the information systems lingware (and therefore the KMS) and are used in such tasks as automatic indexing, automatic classification, query expansion, etc.

In practice of information activities thesauri are the most widely used. The thesauri's main purpose is to reflect the immanent relationships (paradigmatic - at the linguistic level), i.e. permanent, stable for the subject area links between objects.

However, the objects in a subject area also have the so-called situational relationships, which are perceived as "temporary" and are connected with the certain situation. If the immanent relationships are inherent in the subject area in general, the situational relationships link the objects in a subject area within its fragment (corresponding to a certain process), which in addition is time-varying.

Any object exists at least in two states – as "integrated" into the current conceptual view of the subject area in general and as participant represented in a real situation. This makes it necessary to support not only the strict conceptual structures (which include thesauri), but also to have generating means of operative (situational) mobile terminological networks, for example, which enable to identify the property of novelty on the basis of the analysis of commonalities and differences between the current terminology situation and the "established" terminology of the subject area.

Let us consider some of the models and the technology of knowledge management tasks using both thesauri and rubricators (subject headings indexes) and mobile terminological structures, being built "on the fly" on base of information files.

2. Model of conceptual images comparison based on semantic similarity

2.1. Semantic similarity measure

Semantic similarity measure is a numerical value, which represents the measure of similarity between two objects. It is often used in fuzzy matching algorithm both in information retrieval and analytic problem. Semantic similarity is mentioned when objects are presented as descriptions, built by using denoting subjects (classes of subjects) terms, which are the stable concepts signs in the subject area [2].

The formalization and comparative assessment of semantic similarity make the subject of many studies. Note can be given to papers that contain quite extensive overviews on semantic similarity [3, 4].

The measure analytical expression depends on the used features of semantic similarity (for specialized conceptual and symbolic resources such features might be related to the types of relationships between concepts).

The conceptual network in information retrieval thesaurus is formed mainly by hierarchical and associative relationships.

The additive measure is proposed to use in [5] to calculate the semantic similarity of thesaurus descriptors:

$$S(d_1, d_2) = \alpha \cdot S_H(d_1, d_2) + \beta \cdot S_A(d_1, d_2) \quad (1)$$

where $S_H(d_1, d_2) = \frac{|UC(d_1, H^{d_1}) \cap UC(d_2, H^{d_2})|}{|UC(d_1, H^{d_1}) \cup UC(d_2, H^{d_2})|}$ is the contribution of hierarchical relationships; $S_A(d_1, d_2) = \frac{|UC(d_1, A^{d_1}) \cap UC(d_2, A^{d_2})|}{|UC(d_1, A^{d_1}) \cup UC(d_2, A^{d_2})|}$ is the contribution of associative relationships.

Multiplicity $UC(d_i, H^{d_i})$ contains descriptor d_i, and all its parents (based on transitive hierarchical relationships) in a hierarchical chain – H^{d_i}:

$$UC(d_i, H^{d_i}) = \{d_j \in H^{d_i} | \exists m = (d_i, x_{i_1}, x_{i_2} \dots, x_{i_k}, d_j) \cup (d_i = d_j)\} \quad (2)$$

where *m* is path, connecting d_i and d_j through broader terms descriptors $x_{i_1}, x_{i_2}, \dots, x_{i_k}$.

The association has the property of symmetry, however, the specific semantics of symmetry cannot be determined since the actual type of communication is not specified. For the construction of the multiplicity $UC(d_i, A^{d_i})$ [3] proposes this concept and rules of "semantic double" construction

$$UC(d_i, A^{d_i}) = \{d_j \in A^{d_i} | \exists a = (d_i, \widetilde{d_j}) \cup d_j \cup (d_i = d_j)\} \quad (3)$$

where *a* is associative relationship, connecting d_i and $\widetilde{d_j}$ ($\widetilde{d_j}$ – «semantic double» of descriptor d_j.

To determine the values of the coefficients α and β there are «vertical» and «horizontal» development of specific information retrieval thesaurus:

$$\alpha = \frac{S_H}{S_H + S_A} \quad (4)$$

$$\beta = \frac{S_A}{S_H + S_A} \quad (5)$$

where $S_H = \max_{d_1, d_2 \in T} S_H(d_1, d_2)$, $S_A = \max_{d_1, d_2 \in T} S_A(d_1, d_2)$, *T* – set of thesaurus descriptors.

2.2. Algorithm for constructing the semantic intersection of conceptual images

In [5] one also uses an algorithm to measure the semantic similarity of fuzzy matching descriptions of two objects, based on the coinciding maximum principle.

Descriptions of objects represented by their linear conceptual images - a thesaurus descriptors sets - $D_1 = (d_1^1, d_2^1, \dots, d_n^1)$ и $D_2 = (d_1^2, d_2^2, \dots, d_m^2)$. For all descriptors pairs in the thesaurus a semantic similarity matrix may be calculated measuring $n \times m$: $W = (w_{ij})$, where $w_{ij} = S(a_i, b_j), i = \overline{1..m}, j = \overline{1..n}$.

In contrast to the different approaches, the sufficiency of the value of semantic similarity to include a pair of descriptors in the semantic intersection can be determined not by defined fixed threshold, and based on the local context, formed descriptors of each image. A set of semantic intersection is formed from the descriptor for which the matching condition highs:

$$\max_{j=1,n}(w_{ij}) = \max_{i=1,m}(w_{ij}) \quad (6)$$

This algorithm can solve the problem of establishing a semantic similarity measure threshold to include a pair of descriptors in the fuzzy intersection.

2.3. Semantic concepts neighborhood formation

The concept, which is included as a descriptor in the corresponding thesaurus, is generally accompanied by a descriptor section showing there is a direct connection (hierarchical and associative) with other descriptors. However, using this connection it is necessary to consider that the thesaurus's hierarchical relations link concepts and do not conform to the objects or subjects, so disturbed the correct structure of the tree - the same term can have several "parents" - the broader terms at the previous level. Meanwhile associative relationships do not have the property of transitivity. Therefore, an automatic query expansion built on an automatic thesaurus using direct links, in practice do not lead to constructive results.

The semantic neighborhood construction in accordance with the value of semantic similarity measure and descriptors ranking allows using in the expansion of the search query not just the

descriptors, combined descriptor entry, and those whose semantics are closer due to the links which reflect the common semantics of genus-species and associative relationships (for example more than one of the generic descriptor, a different set of subordinate descriptors for ones concepts). Figure 1 shows a descriptor entry NEUTRONS from the INIS thesaurus [6] and semantic fragment of the neighborhood (the terms are arranged in descending order of measure), showing that the descriptors in a descriptor entry of the same level (narrow terms) have different values of semantic similarity measure.

0,56-THERMAL NEUTRONS	**NT (narrow terms):**
0,54-COLD NEUTRONS	*ANTINEUTRON*
0,54-RESONANCE NEUTRONS	BETA-DELAYED NEUTRONS
0,54-PILE NEUTRONS	FAST NEUTRONS
0,54-INTERMEDIATE NEUTRONS	COSMIC NEUTRONS
0,54-POLYNEUTRONS	SLOW NEUTRONS
0,54-FISSION NEUTRONS	EPITHERMAL NEUTRONS
0,54-EPITHERMAL NEUTRONS	FISSION NEUTRONS
0,54-SLOW NEUTRONS	POLYNEUTRONS
0,54-FAST NEUTRONS	INTERMEDIATE NEUTRONS
0,54-BETA-DELAYED NEUTRONS	RESONANCE NEUTRONS
0,53-NUCLEONS	PILE NEUTRONS
0,48-PHOTONEUTRON	SOLAR NEUTRONS
0,48-*ULTRACOLD NEUTRONS*	THERMAL NEUTRONS
0,48-TRINEUTRONS	COLD NEUTRONS
0,48-TETRANEUTRONS	PHOTONEUTRON
0,48-PROMPT NEUTRONS	**BT (broader terms):**
0,48-DELAYED NEUTRONS	NUCLEONS
0,45-PHOTONUCLEONS	**RT (related terms):**
0,45-PROTONS	*NEUTRON FLUX*
0,42-DINEUTRONS	*NEUTRON DENSITY*
0,42-BARYONS	*NEUTRON BEAMS*
0,4-PROMPT PROTONS	*CINDA*
0,4-TRAPPED PROTONS	*NEUTRON SPECTRA*
0,4-RETARDED PROTONS	*NEUTRON TEMPERATURE*
0,34-SOLAR NEUTRONS	*NEUTRON SEPARATION ENERGY*
0,34-COSMIC NEUTRONS	

Figure 1. Descriptor entry NEUTRONS and fragment of semantic descriptor neighborhood NEUTRONS (denoted as mismatched descriptions in italics).

3. Technology of lingware support

3.1. The technology of joint use of thesauri and rubricator in a task of automatic classification

Automating the process of the input information flow distribution by a predetermined classes (for example, by subject classifier categories) is based on the formation of a preliminary descriptive specification of classes (categories) on the one hand, and the input documents on the other hand [7].

Method of classification using thesaurus involves the construction of descriptive specification headings on the basis of the thesaurus descriptors. In this case the intersection of a row and column matrix of “descriptor-category” contains the value which characterizes the semantic proximity of the concrete descriptor to the category.

To construct this matrix it is necessary to solve two problems:

- to distribute preliminary thesaurus descriptions according to categories;
- to calculate semantic similarity measure of "descriptor-category".

To solve the first problem an abstracts database can be used, in which each abstract is provided with a search pattern formed on the basis of expert estimation. Such a search pattern is usually

represented as a set of descriptors and subject indexes. In this case technology of comparison of the descriptor with the category includes the following steps:

1. Building a list of descriptors belonging to the category.
2. Identification of descriptors relating to more than one category.
3. Calculation of value of descriptor-category correspondence for descriptors identified at step 2 (for example, as such measures coefficient of linear correlation can be used).
4. Descriptors distribution by categories based on the principle of correlation coefficient maximum.

However, this approach does not guarantee all thesaurus descriptors distribution by category. Algorithm application for classification descriptors of INIS thesaurus in accordance with the INIS categories [8] using the abstracts database SARI showed that near 50% of thesaurus descriptors could not be distributed by rubrics (since these descriptors do not exist in the database).

To ensure classification completeness this primary distribution can be further supplemented by the thesaurus relations and semantic similarity measures. A descriptor, to which a category is not assigned yet, is regarded as the node of a permissible conceptual path. Then previous and subsequent nodes are determined which correspond to descriptors that are already matched to the subject headings and the semantic similarity measure of these descriptors with the specified descriptor is calculated. At the last step the category is assigned to that descriptor for which the value of the measure is maximum.

3.2. Technology of the terminology systems actualization

The above-mentioned strictness of conceptual terminological structures such as thesauri does not allow to effectively track information flow changes and the formation of certain language using directions. To construct and promptly update a subject area concepts and terminology systems, an interactive and iterative technology based on the methods and means of information retrieval in the documentary resources of scientific, technological, and educational information can be used. The technology involves the following steps [9]:

- development of an information concept of a subject area at the level of documents that are selected from the document sources, which adequately and objectively reflect the state of the art in the subject area;
- the design of a terminology base in the form of glossaries, i.e., the lists of terms that express the basic meanings of the selected relevant documents;
- construction of concepts and terminology networks (thesauri and ontologies) for the subject area.

The proposed method of constructing a terminological network structure with hierarchical relationships for the collection of texts of a particular subject area is based on the sequential formation of lexicographical neighborhoods of "nuclear" terms, which are used as high-frequency nouns. Lexicographical neighborhood includes the phrase with a noun as the main word. Construction of phrases is made according to patterns [10, 11] with the following basic rules:

- adjective precedes the noun;
- in phrases with two or more nouns without prepositions, second and subsequent nouns should be in the genitive, dative or instrumental case (place names are not considered);
- nouns related by prepositions may be in any case.

Table 1 shows examples of word combinations of patterns corresponding the above rules by which combinations of operating documentation were extracted and assessed the correctness of their selection. The constructed combinations were sorted in descending order of frequency of occurrence and reviewed the first 300 for each pattern. The right column shows the number of incorrect phrases per considered 300.

Table 1. Number of incorrect phrases according to patterns.

Pattern	Number of incorrect phrases
adjective + noun	0 (0%)
noun + noun in the genitive, dative or instrumental case	30 (10%)
noun + prepositions + noun	85 (28.3%)
noun + adjective + noun in the genitive, dative or instrumental case	18 (6%)
adjective + noun + noun in the genitive, dative or instrumental case	33 (11%)

Figure 2 shows the lexicographical neighborhood of term NEUTRONS built for glossary.

NEUTRONS	NEUTRON ABSORPTION
FAST NEUTRONS	NEUTRON FLUX
NEUTRON INTERACTIONS	NEUTRON FLUX DENSITY
NEUTRON DIFFUSION	NEUTRON FLUX DENSITY MEASUREMENT
NEUTRON MODERATION	NEUTRON FLUX IN ACTIVE ZONE
NEUTRON CAPTURE	NEUTRON FLUX IN REACTOR
NEUTRON SOURCE	NEUTRON MULTIPLICATION
PROMPT NEUTRONS	NEUTRON REACTION
SLOW NEUTRONS	RESULT OF NEUTRON REACTION
EPITHERMAL NEUTRONS	REGISTRATION OF NEUTRONS
NEUTRON IRRADIATION	NEUTRON VELOCITY
NEUTRON REFLECTOR	THERMAL NEUTRONS
NEUTRON REFLECTION	THERMAL NEUTRON CAPTURE
NEUTRON TRANSFER	THERMAL-NEUTRON ABSORPTION
NEUTRON ABSORBER	REGISTRATION OF THERMAL NEUTRONS

Figure 2. Fragment of lexicographical neighborhood for descriptor NEUTRONS.

There is a fairly insignificant intersection with descriptors thesaurus entry. This lexicographical neighborhood allows associating terms in different aspects. Relationships exist not only genus-species, but also situational ones which characterize the use of the object in the present context.

4. Conclusion

The proposed models and technologies are focused on the independent lingware objects contour formation in the KMS framework. These objects support requires the development of models and technologies from the point of view of their subject area adequacy. It should be done in two directions: updating stable conceptual terminology structures for the subject area as a whole (with the aim of embedding knowledge in the current vision), and dynamic situational terminological structures formation (searching for and fixing similarities and differences related to the concrete knowledge objects).

The addressed models and algorithms have been tested on an array of full text dissertations conceptual search images.

References

[1] Golitsyna O L, Kupriyanov V M and Maksimov N V 2015 *Autom. Doc. Math. Linguist*. **42** (3) 150-161
[2] SINTOL 1968 *Sbornik perevodov po voprosam informacionnoj teorii i praktiki* (Moscow: VINITI) no 10 p 177
[3] Kryukov K V, Pankova L A, Pronina V A, Sukhoverov V S and Shipilina L V 2010 Measures of semantic similarity in ontology *Probl. Upr.* **5** 2-14
[4] Ngok N B and Tuzovskii A F 2013 The model of information retrieval based on semantic meta descriptions *Upr. Bol'shimi Sist.: Sb.Tr.* **41** 51-92

[5] Golitsyna O L, Maksimov N V and Fedorova V A 2016 *Autom. Doc. Math. Linguist*. **50** (4) 139-153
[6] INIS Thesaurus 2016 *IAEA-INIS-01 (2016/10)* (Vienna)
[7] Gulyaev O V and Loukachevitch N V 2013 *Novye informacionnye tehnologii v avtomatizirovannyh sistemah* **16** 238-244
[8] Subject Categories and Scope Description 2010 *INIS/ETDE Joint Reference Series* no 2 (Rev 1)
[9] Golitsyna O L, Maksimov N V, Strogonov V I and Tikhomirov G V 2011 *Sistemy Upr. Inf. Tekhnol*. **44** (1.1) 126-134
[10] Khokhlova M V 2012 *Proc. of the Annual Int. Conf. "Dialogue"* (Moscow: RSUH)
[11] Nayhanova L V 2005 *Internet Portals: Content and Tehnologies* **4** 452-479