Formation of Search Queries Based on Thesaurus of Narrowly Specialized Subject Areas

H. Matsiuk[1], *A. Rzheuskyi*[2], *N. Kunanets*[2], *N. Veretennikova*[2]

[1] *Department of Social Communication and Information Activities*
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
galuna.matsiuk@gmail.com
[2] *Information Systems and Networks Department*
Lviv Polytechnic National University
Lviv, Ukraine
antonii.v.rzheuskyi@lpnu.ua, nek.lviv@gmail.com, nataver19@gmail.com

Received May 08.2019: accepted: June 15.2019

Abstract. The role of branch information retrieval thesaurus for improving and increasing the effectiveness of information support of scientific research, carried out using the technology of selective dissemination of information, is analyzed in the paper. The selective dissemination system provides both individual and group information needs, as subscribers can be individual users as well as a group of users, namely scientific teams and virtual scientific teams whose members can be geographically distributed. Within the scientific team, subgroups can be formed engaged in research of different directions of their subject area. The information retrieval thesaurus is considered as a language model of a particular subject area. The development of information technologies creates the possibility of information modeling of a narrowly specialized subject area and data presentation in the form of thesaurus, which is a hypertext that reflects the hierarchically organized semantic structure of a specific subject area.

Key words: search queries, thesaurus, narrowly specialized subject area, information support, scientific research, selective dissemination of information.

INTRODUCTION

The rapid growth of the extent of information flows is one of the features of the information society formation at the present stage of its development. It is not possible to interrupt the growth and speed of changes in information flows. The emergence of new research results generates an increase in the arrays of published scientific materials. The orientation problem and search in a huge array of under ordered materials often results in information duplication, which is accompanied by the waste of time, material resources and intellectual resources when performing a search. Therefore, there is a need to improve search processes, processing and providing relevant information. At the same time the knowledge presentation in a formalized form does not only greatly accelerate the processes of working with information, but also increases search productivity. The processes of thesaurus compiling are considered in the works of Morkovkin V. [1], Horodysheva A. [2], Gladun A. [3]

The **aim** of the paper is to reveal the peculiarities of the thesaurus formation as a means of terminology systematization of highly specialized subject areas in order to increase the system efficiency of selective information dissemination.

MAIN PART

Nowadays, information support of scientific researches relies on libraries [4, 5] as social and communication institutions. Generating own databases, finding relevant information, operating abstract branch databases, and selecting true information by library information workers [6] are included in the system of selective dissemination of information (SDI). However, even librarians are not able to cover the growing amount of information resources and the dynamics of scientific literature publication [7, 8]. The SDI system needs to optimize the search in modern information retrieval systems (Fig. 1). The effectiveness of the SDI system depends on the formation of profiles of individual or group users to a large extent. A terminology dictionary (thesaurus) of the database should be created at the preparatory stage and during the whole operation of the SDI system to look for information with the exact spelling of the search term and the registered links between them. Also, a hyperlink system can be automated, where all terms are connected with the corresponding sources by active links. SDI is a service of instant notification for scientists, which includes the selection of relevant information designed to meet the specific information needs of a user or a group of users and the direct showing of full-text documents or metadata about them. SDI staff provides information needs for both individual users and groups of users such as research teams and virtual scientific teams whose members can be geographically distributed and conduct interdisciplinary research. In recent years, the preparation of requests for state budget research works has been actively pursued in Ukraine. Such scientific teams can combine geographically or institutionally distributed researchers, representatives of different fields of knowledge that provide a comprehensive systematic study of the problem.

The main means of increasing the search completeness and accuracy is effectively developed linguistic support. Linguistic means is an interface between the natural language and the formal search algorithms of information retrieval systems. Linguistic support is formed from a number of elements. Firstly, this is an artificial language of data representation in information retrieval systems that defines the architecture, syntax and semantics of information representation in databases of information retrieval systems and information retrieval language, that is a language which a user uses to appeal to the information system for obtaining information that they are interested in [9]. Secondly, it is linguistic support that facilitates the effective implementation of processes such as document and request indexing, which contributes to the effective formation of thematic searches in databases, as well as intersystem information interaction. The development of information technology generates a possibility of information modeling of a highly specialized domain and presents the received data in the form of a thesaurus, which is a hypertext reflecting a hierarchically organized semantic structure of a certain subject domain. The concept of thesaurus is now actively used in the fields of artificial intelligence, information technology and library science. More and more intellectual tasks related to the information processing, especially document indexing, information retrieval and automatic analytical and synthetic document processing are solved using thesauri. If a mediator (thesaurus) is placed between the document array and the library information worker, it significantly reduces the volume of the obtained results and increases the level of their relevancy. We can present the following definition of a thesaurus. Thesaurus of a subject area is a knowledge system presented in the form of a set of key terms or descriptors of this domain, which are interconnected by certain semantic relations, which represent the basic relations of the concepts of described knowledge domain. The main purpose of the thesaurus is to increase the effectiveness of finding the necessary information. Thus, a thesaurus is a way of systematizing the knowledge of a certain subject area and it is an effective information retrieval tool (Fig. 2).

Knowledge graphs help create search patterns to meet effectively the information needs of SDI subscribers. Obtaining the request, links to related concepts are built up to each keyword, which ensures unambiguous interpretation of each concept and achievement of clarity in determining the subject area of the information request. This approach improves the process of reaching the relevance of search results. So, formed knowledge graphs are constantly replenished with new links, concepts and can be reused in the search in case of obtaining similar requests of SDI subscriptions or extension of the search area. Thesauri are used as tools of terminological control in the process of analyzing and indexing documents and information requests, as well as in the process of automated information retrieval. The functional role of the thesaurus in the information retrieval system brings forward high requirements for the quality of the thesaurus preparation, and the search effectiveness largely depends on its perfection degree. The thesaurus is a kind of general or special dictionary, which has difference in the technique of designing and constructing its vocabulary articles that allows to display synonymous, antonymic, paronymic and other semantic relations between lexemes. Such a comprehensive reflection of a terminology domain system makes the thesaurus a very effective and equivalent assistant for the description of a particular subject area. That is why the dictionaries of a thesaurus type are becoming more widespread. A characteristic feature of the thesaurus is that the articles are constructed to reflect semantic shades, close to the meaning of words. While the traditional interpretative dictionary gives only the word definition, the thesaurus articles reveal the word meaning both through the definition and through the comparison with other words and conceptual groups, the explication of the semantic connections of one term with others. The main difference between the thesaurus and the traditional dictionary is that it provides not only a compilation of articles, but also a system of term-descriptors. Keywords are equivalent and close in value, and the information processing and searching is carried out with their help. They are grouped into a class of conditional equivalence. Each class is a vocabulary unit that is presented in the form of a separate word, phrase or code. This vocabulary unit is a descriptor. Classificated content descriptors form a thesaurus. The process of constructing a thesaurus has the following steps:

- the preliminary selection of lexical units (composition of keyword lists, dictionaries);
- the construction of conditional equivalence classes, that is the transformation of lexical units into a given standard form;
- the establishment of existing semantic relationships.

The methodology development, the lexical material selection and further work on the thesaurus completion involve a considerable amount of research work [10]. In the process of this work, the domain terminology analysis is conducted in order to identify the term-pretenders for inclusion in the thesaurus and their individual meanings. The pair of concept – term is determinant in defining the basic principles of ordering and standardization of branch terminology, and system of concepts – system of terms are important at a higher level.

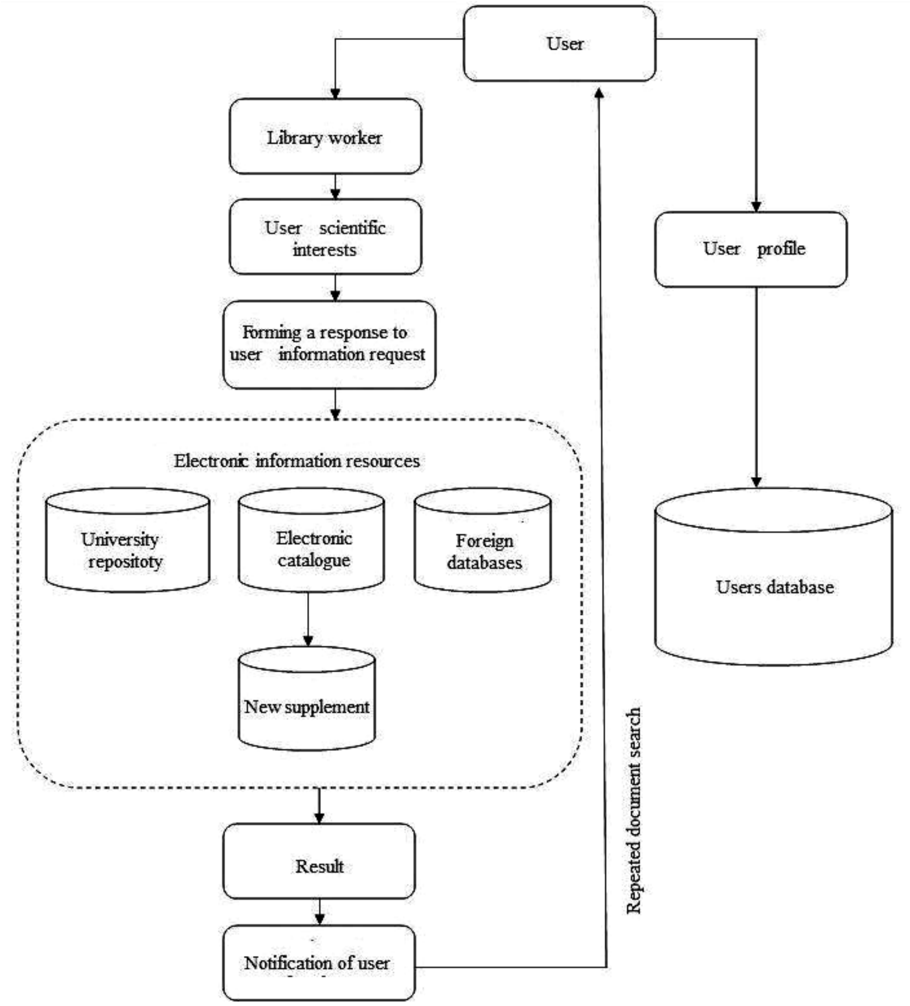

Fig. 1. Realization of SDI process

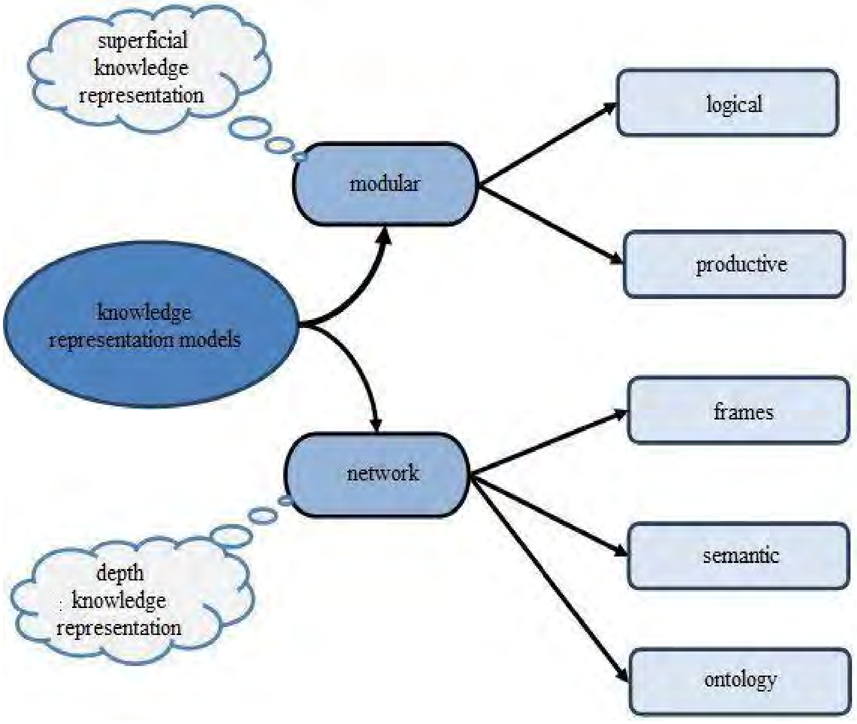
superficial knowledge representation
modular
logical
productive
knowledge representation models
network
frames
semantic
ontology
depth knowledge representation

Fig. 2. Classification of Knowledge Representation Models

FEATURES OF THESAURUS COMPILING

A thesaurus compiling is a multioperational complex process. Each stage of work involves the analysis of several options. It is necessary to ensure certain conditions in order to obtain a new intellectual product, such as a clear understanding of the purpose and the task statement; adequacy of the information base; full description of the subject area objects; clear collaboration of a group of researchers. We will consider the process of constructing a dictionary on the example of the subject area in smart city (Fig. 3).

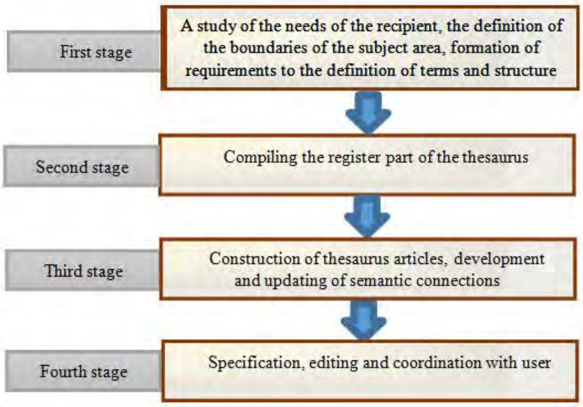
First stage
A study of the needs of the recipient, the definition of the boundaries of the subject area, formation of requirements to the definition of terms and structure
Second stage
Compiling the register part of the thesaurus
Third stage
Construction of thesaurus articles, development and updating of semantic connections
Fourth stage
Specification, editing and coordination with user

Fig. 3. Compiling the Thesaurus of Smart City Subject Area

The research work on thesaurus compiling begins with the selection of lexical material. At this stage certain difficulties relate to the need to differentiate the terminology and commonly used vocabulary. As the terminological material extracted from texts is accumulated, the main terminological fields of the studied domain are determined. In modern science, a lot of efforts have been made to find objective criteria that will help to identify unambiguously the terminology units expressing the concepts and relationships with one or another knowledge domain, and thus determine the scope and limits of terminology. Researchers solve the problem of identifying terms in one or another subsystem in different ways, using the various definitions of a term, offering their methods for selecting terms, criteria for their selection in the case when the concept system has not been developed yet and there is no clear boundary between the term and not the term [11]. Constructing an information retrieval thesaurus, the first task is to select terms for inclusion in it. There are several possible sources for selecting terms when creating a thesaurus. First of all, the thesauri of close subject areas have to be studied that may contain a significant number of terms, which could enter the structure of the new thesaurus. The terms - candidates for inclusion in the thesaurus can be suggested by the experts of the subject area. Another source for terminological units is the scientific texts. The terms included in the thesaurus have to meet the following requirements:

- thesaurus terms should share the concepts that are present in scientific texts and should be selected for reasons of effectiveness of their use in the document search;
- an important factor for the term inclusion is the frequency of its use in the texts;
- an inclusion of new terms in the thesaurus should consider already existing terms, as it is necessary to check whether the term – candidate is a separate concept, which does not match the existing terms of the thesaurus; also, it should be avoided the term inclusion whose meanings intersect with the meanings of existing terms in the thesaurus.

The selection of terms was carried out based on the following criteria such as:

- Informativeness. It is a possibility of a candidate-term to designate a linguistic phenomenon and at the same time to be the basis for an information request;
- Frequency. The terms are selected that are the most used in the subject area;
- Topic relevance. The selected terms may reflect the problem;
- Actuality. The preference was given to terms that reflect contemporary developments in the field of science or are of great importance for its understanding;
- Practical significance. Terms are chosen and their understanding has a significant effect on text clarification.

The main function of the thesaurus is to facilitate the search and selection of the necessary information, but at the same time the structure and principles of an organization, which ensure the sufficient term filling and the depth of semantic interrelation disclosure, allow the thesaurus to simulate the terminological system of the subject domain most fully and systematically (Fig. 4). The practical value of the subject area thesaurus is conditioned by a possibility of using it simultaneously for the analysis and design of the lexical system, the classification and storage of terminological data, the information processing in search engines. The thesaurus represents the term system of the relative branch. You can not only search with its help, but also study certain concepts, get the relation system of the term with other units, to find out its role in the knowledge system of a particular industry. The thesaurus affects the language competence of the specialists, determines the updating of the terminology of a certain knowledge domain and makes available scientific information sources, helping multilingual speakers understand each other and contribute to study the languages presented by the thesaurus.

CONCLUSIONS

Thus, mathematical models, methods and algorithms of distance learning processes are proposed and explored as well as the creation of an integrated network-oriented information and learning environment is based on it, and adaptive control of remote network-oriented learning process is considered. A network-oriented distance learning process can be considered as a process of managing a complex system in which the object of management is distance learning. The advantages of the proposed and developed algorithms of presenting educational material using the apparatus of multivalued logic are:

- Minimal use of expert assessments.
- Using a relational model for data storage can reduce the system complexity through the use of relational algebra operators.
- Setting up learning and testing process on the capabilities of a particular student.

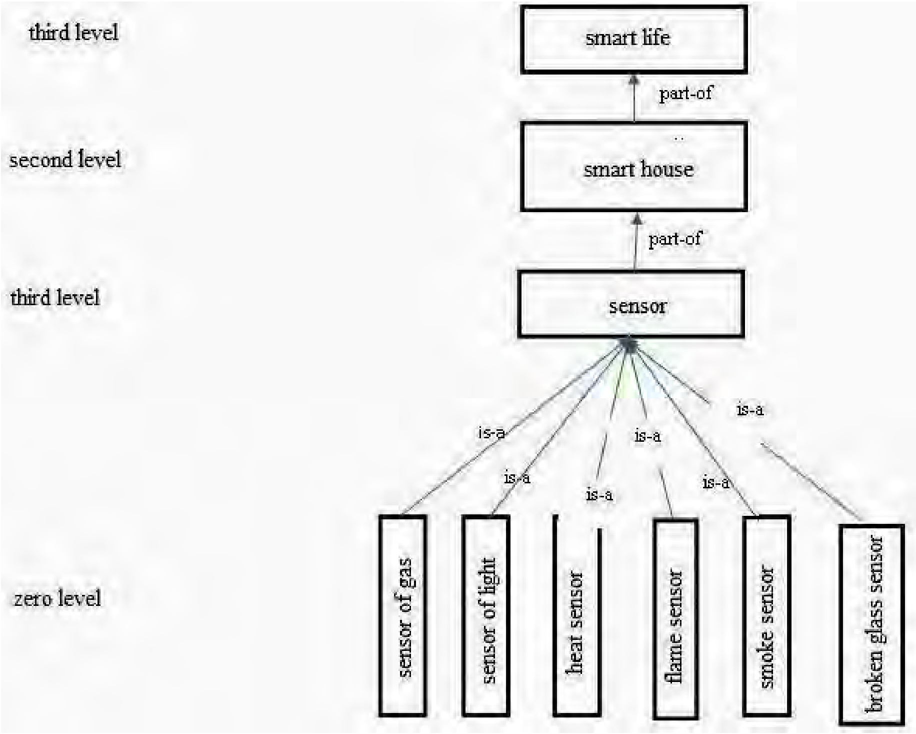

Fig. 4. Example of semantic network

REFERENCES

1. **Morkovkin, V. V. 1970.** Ideograficheskie slovari [Ideographic dictionaries], Moscow, Russia: Izd-vo MHU. [In Russian].
2. **Gorodishcheva A. N., Gorodishchev A. V. 2011**. Standards, specifications and features of the logicallinguistic terminology management systems, communications”. Science Prospects, 5 (20), pp. 62–65.
3. **Hladun, A. Ya., Rohushyna, Yu. V. 2008.** Osnovy metodolohii formuvannia tezaurusiv z vykorystanniam ontolohichnoho ta mereolohichnoho analizu [Fundamentals of the thesaurus creation methodology using ontological and mereological analysis]. Shtuchnyi Intelekt, 5, 112–124. [In Ukrainian].
4. **Rzheuskyi A., Kunanets N., and Kut V. 2017.** The analysis of the United States of America universities library information services with benchmarking and pairwise comparisons methods”, in: 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), vol. 1, pp. 417–420. Lviv, Ukraine.
5. **Rzheuskyi A., Kunanets N., and Kut V. 2018.** Methodology of research the library information services: the case of USA university libraries. Advances in Intelligent Systems and Computing, vol. 689, 2018, pp. 450–460.
6. **Bomba A., Nazaruk M., Pasichnyk V., Veretennikova N., and Kunanets N.** “Information technologies of modeling processes for preparation of professionals in smart cities”. Advances in Intelligent Systems and Computing book series, vol. 754, 2018, pp. 702–712.
7. **Veretennikova N., Kunanets N. 2018.** “Recommendation systems as an information and technology tool for virtual research teams”. Advances in Intelligent Systems and Computing, vol. 689, pp. 577–587.
8. **Rzheuskyi A., Veretennikova N., Kunanets N, and Kut V.** “The information support of virtual research teams by means of cloud managers. International Journal of Intelligent Systems and Applications (IJISA), 10(2), 2018, pp. 37–46.
9. **Olifer, V., Olifer, N. (2009). Osnovy kompiuternykh setei** [Basics of computer networks]. St. Petersburg, Russia. [In Russian].
10. **Kaminskyi R., Kunanets N., Rzheuskyi A. 2018.** Mathematical support for statistical research based on informational technologies. CEUR Workshop Proceedings, vol. 2105, pp. 449–452.
11. **Shunevich, B. I. 2008.** Suchasni sposoby vidboru terminiv ta ukladannia perekladnykh slovnykiv novykh terminosystem [Modern ways of selecting terms and making translated dictionaries of new terminology systems]. Visnyk Zhytomyrskoho Derzhavnoho Universytetu Imeni Ivana Franka, 38, 90–93. [In Ukrainian].