UDC 004.91

# MULTI-AGENT LATENT SEMANTIC INTERNET TECHNOLOGY FOR THE FORMATION OF A SUBJECT-ORIENTED KNOWLEDGE MODEL

**Stenin A. A.** – Dr. Sc., Professor of the Department of technical Cybernetics, Kiev Polytechnic Institute. Igor Sikorsky, Kiev, Ukraine.

**Pasko V. P.** – PhD, Associate Professor, Department of technical Cybernetics, Kyiv Polytechnic Institute. Igor Sikorsky, Kiev, Ukraine.

**Soldatova M. A.** – PhD, Senior lecturer of the Department of technical Cybernetics, Kiev Polytechnic Institute. Igor Sikorsky, Kiev, Ukraine.

**Drozdovich I. G.** – PhD, Senior Scientific Associate at the Institute of telecommunications and global information space of the National Academy of Sciences of Ukraine.

## ABSTRACT

**Context.** The article proposes a latent-semantic technology for extracting information from Internet resources, which allows processing information in natural language, as well as a multi-agent search algorithm based on it. The relevance of this approach to the search for subject-oriented information determined by the fact that currently a direct lexical comparison of queries with document indexes does not fully satisfy the developer. The object of the study is a multi-agent latent-semantic algorithm for searching for subject-oriented information.

**Objective.** The work is to increase the efficiency of forming a knowledge model that is adequate for this subject area.

**Method**. A latent semantic technology based on the weighted descriptor method developed by the authors is proposed. The main difference from the existing methods is that the analysis of words occurring in the text both in frequency and taking into account semantics carried out by selecting the appropriate descriptors, which improves the quality of the information found.

**Results.** The developed latent-semantic technology of information search tested in the task of constructing a knowledge model of automated decision support systems for operational and dispatching control of urban engineering networks. The conducted modeling of the search for subject-oriented information in this subject area showed the effectiveness of the developed approach.

**Conclusions**. Improving the efficiency of search and semantic content of subject-oriented information of the knowledge model of this subject area achieved by using the weighted descriptor method based on Zipf's laws in this technology. The prospects for further research are to build evolutionary models of knowledge and improve the quality of updated information.

**KEYWORDS**: Internet resources, information search, Zipf's laws, Grebner bases, intelligent agents, weighted descriptors, latent semantic analysis, multi-agent automatic search procedure.

## ABBREVIATIONS

DSS – decision support systems;
IRS – information retrieval systems;
LSA – latent semantic analysis;
UEN – urban engineering networks.

## NOMENCLATURE

$f_i$ is a frequency of occurrence of the *i*-th linquistic variable;

$k_i$ is a number of documents with the *i*-th linquistic variable;

$N$ is a significant set of documents;

$N_0$ is a total number of documents under consideration;

*f* is a frequency of occurrences of words in texts;

*k* is a rank of a frequency;

$w_{ij}$ is a frequency significance coefficient;

$\alpha_i$ is a semantic significance coefficient;

$a_{ij}$ is a frequency of appearance of the *i*-th descriptor in the *j*-th document $i = 1,\ldots, m$ , $j = 1,\ldots, N_0$ ;

$a_{ik}$ is a partial private evaluation of the *i*-th descriptor significance;

$k$ is a number of expert;

$S_q$ is *q*-th current situation;

$x_i$ is *i*-th linquistic variable (descriptor);

$M(x_i)$ is a function of belonging of *i*-th descriptor;

$T_q^i$ is a basis, which containing *i*-th descriptor of *q*-th situation.

## INTRODUCTION

With the development of Internet technologies, a new giant source of information resources has appeared. Thanks to the widespread development and application of computer technology, we can get information in electronic form in all areas of human activity, such as science, production, commerce, literature, entertainment, etc. [1]. The Internet is compatible with various electronic networks and databases and allows easy access to almost any kind of information. However, the development of the Internet as an information repository took place without taking into account the need to search for the documents. As a result, on the Internet, in contrast to traditional IRS, where the document storage system is focused on active search [2–4], the Internet document storage system is not a given a priori with respect to the task of information retrieval, i.e. poorly structured. The Internet is a decentralized document repository with no of the single management and organization and of development. The Internet is heterogeneous, as not only different platforms uses, but also different standards of information presentation. The Internet brings together both modern


DOI 10.15588/1607-3274-2021-3-14

and legacy systems. Piece of information is stored in a form other than text (multimedia)

It follows that the task of extracting information from the Internet is complex since it is necessary to extract not only the type of data scheme, but also the semantic information associated with it. In addition, given the need to search for the specific information of the selected subject area, there is f need process a huge quantity of documents. Thus, it is also important to automate the search process for the effective selection of the most informative content. The process of search in an IRS can represented in the form of the scheme shown in Fig. 1 [4].

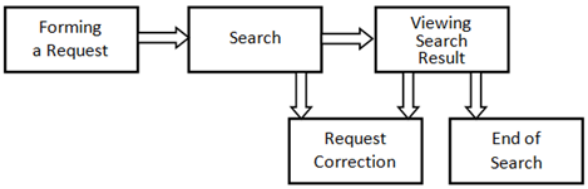


Figure 1 – Diagram of the search process in Information Retrieval Systems

**The object of research** is a multi-agent procedure for searching the Internet for the necessary information, taking into account its semantic value for a given subject area.

**The subject of research** is the latent-semantic method of weighted descriptors, which allows extracting the most significant documents in terms of meaning and meaning that are very close to this subject area.

**The aim of the research** is to form an effective query and automated procedure for analyzing information extracted from the Internet for a given subject area.

## 1 PROBLEM STATEMENT

Taking into account the above, we can say that the main task of the development of IRS on the Internet is the development of methods and tools for semantic analysis of the text in a natural language using a multi-agent approach to reduce the time of searching for the necessary information and increase its semantic value [5].

The formalized task of searching for innovative subject-oriented document can presented in the form of matrix "descriptors-documents"

$$A = \{T, W, R, \Lambda\}. \quad (1)$$

Necessary:

1. Define a set of descriptors $W=\{wi\}$, that reflect the semantic content of the terms of reference and the current situation Si at the time of innovative design of the technical object

$$S_q = \left\{ M_{M_{Sq}(xi)}(T_q^i) / T_q^i \right\}, q = 1,...Q; i = 1,..., N_q; x_i \in X.$$

2. Based on (1) and (2), determine the initial matrix "descriptors -documents" $A$ using the relevance criteria and expert assessments, where m is the number of descriptors, $N_0$ is the number of documents found.

3. Perform LSA k-approximation of matrix A in order to determine the most informative documents, i.e. find the matrix

$$\widetilde{A} \approx USV^T. \quad (2)$$

## 2 REVIEW OF THE LITERATURE

Currently, in the sense of automation of the IRS, actively working on the development of algorithms, which automatically generate intermediary programs (intelligent agents) in the search for information from the Internet.

Achieving full automation in this matter is unlikely, and we can only talk about the creation of automated methods and systems for extracting information from the Internet. Actual research in the field of work with poorly structured information on the basis of "intelligent agents" led to the emergence of a large quantity of alternative tools for their creation [4, 8].

The main approaches to solving the problem of data extraction from the Internet are borrowed from such areas as data processing in a natural language, machine learning, ontology, etc. In particular, the main task of extracting data from the Web is to obtain certain pieces of information (fields) from HTML documents [6, 7].

This task is close to the task of automatic clustering and is to find the decomposition of HTML documents into classes that contain documents with a similar structure. The task of displaying applied objects in points of multidimensional space is to determine the basis of features that form a multidimensional space, and the method of decomposition of the document on this basis, that is, the calculation of the coordinates.

Different approaches were used to determine the coordinates of the document in the space of basic features. In particular, the authors of [7] propose to use the approach popular in calculating the weights of terms in the IRS, using a vector model of presentation of documents. The coordinates of the document are determined by the formula:

$$w_i = f_i / \log(N / k_i). \quad (3)$$

Often, an entropic measure uses to assess the quality of clustering. However, this approach, which determines the significance of the term only in frequency does not guarantee the significance of the document in meaning.

To overcome this contradiction, the W3C Interest Group Note consortium develops Web semantic [8, 9]. According to the idea of its creators, the implementation of this paradigm on the Internet will allow information systems to understand the content of information to some extent and act as intellectual intermediaries capable of manipulating it on the instructions of a person [10, 11].

In this sense, the use of George Zipf's laws is relevant [12]. Zipf found that if you multiply the probability of


DOI 10.15588/1607-3274-2021-3-14

finding a word in the text by the rank of frequency, the resulting value is approximately constant for all texts in one language (for English texts $C \approx 1$, for Russian texts $C \approx 0.06 - 0.07$:

$$C = (fxk) / N \approx \text{const}, \quad (4)$$

frequency of occurrence of the word in the text has the form of a hyperbola (Fig. 2). The rank is permanent for texts of the same language.

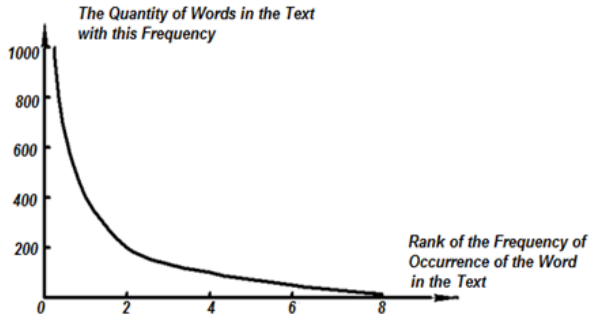


Figure 2 – Zipf curve

The discovery of Zipf's laws was the beginning of infometrii, the science of measuring the quantitative characteristics of information. Explanation of Zipf's laws based on correlation properties of additive Markov chains with step memory function [13]. Zipf's laws are universal. They, in particular, meet the characteristics of the popularity of Internet sites.

## 3 MATERIAL AND METHODS

As mentioned above, the semantic approach is currently one of the main ways to improve the IRS, as direct lexical comparison of requests with document indexes does not fully satisfy the developer. This is due to found documents have either polysemy (i.e, lot of extra words) or synonymy (i.e, not all meaningful words extracted). Therefore, within the framework of the semantic approach, a latent-semantic method of weighted descriptors, which allows extracting the most significant in meaning and significance documents that are very close to the subject area, is proposed.

The method based on the idea of Groebner basis [14], which statistically constructed conceptual descriptors. This method assumes that the conceptual descriptors in the sentences have an underlying, "latent" meaning, which obscured by the use of different words. "Ideal" in determining the basis of Groebner consider the terms of reference for the implementation of some innovative project. To obtain meaningful conceptual descriptors, we will use Zipf's laws described above.

Zipf curve analysis (Fig. 2) shows that the most significant words, and, consequently, the meaningful conceptual descriptors constructed on them, lie in the middle part of the Zipf curve. In the same time, in the Russian language the most common prepositions, pronouns, etc., and in English – service words and others. They answer by the left part of the diagram. The right part of the Zipf curve corresponds to words that have no decisive semantic meaning and are not of interest in the formation of conceptual descriptors.

Therefore, the success of the IRS depends on how the range of the most significant words and conceptual descriptors constructed on their basis will be determined. In many ways, the definition of the range large ($N_0$>>0) depends on the correct compilation of the special dictionaries – the thesaurus of a subject area and the "stop-dictionary". The thesaurus of a subject area gives you the opportunity to correctly determine the set of concept descriptors from the technical specifications and the most important concepts in this subject area. Stop dictionary cuts off "interference" in the form of "extra" words, i.e. for the Russian language – it particles, prepositions, pronouns, etc.

The choice of the quantity of descriptors determined by the quantity of constructed Groebner basis, the totality of which fully reflects the "ideal", i.e. an innovative project reflected in the terms of reference.

Let us initially construct and select n descriptors. Then on their request to the Internet, we will get a rectangular matrix "descriptors-documents" $A_1 = \{a_{ij}\}$ with dimension $mxN_0$.

Usually, the Internet search for the selected descriptors the initial quantity of documents turns out to be quitelaIn In this case, the work with large-dimensional matrices within the LSA is significantly complicated. Therefore, in this paper proposed to determine the most significant set of documents from the set in two stages.

At the first stage, we rank the set of $N_0$ found documents by the frequency of using basic descriptors in them and by their semantic significance. To do this, we construct a criterion of the relevance of the following form:

$$R_j = \sum_{i=1}^{n} \alpha_i w_{ij}, j = \overline{1, N}. \quad (5)$$

As a frequency significance coefficient can use the generally accepted formula (3), in which we understand the frequency of appearance of the descriptor in the document.

The semantic coefficient of significance determined by the method of expert evaluations according to the following formula:

$$\alpha_i = \sum_{k=1}^{K} \alpha_{ik} / \sum_{i=1}^{m} \sum_{k=1}^{K} \alpha_{ik}. \quad (6)$$

Private estimates are determined by their ranking by a natural quantity series, in this case, from 1 to m. A larger quantity corresponds to a large assessment of the significance of the descriptor.


DOI 10.15588/1607-3274-2021-3-14

As a result, we form the set of $N$ significant documents and the initial matrix $A(m\ x\ n)$ for LSA, where $n=N << N_0$. As a rule, the sample set $N$ depends on the concrete problem and chosen empirically. In this case, we accept $a_{ij}=1$, if the $i$-th descriptor is in the $j$-th document, and $a_{ij}=0$, otherwise.

At the second stage, we use singular value decomposition of matrix A, which is described below.

The LSA method based on the principles of factor analysis, in particular, the identification of latent relationships of the studied phenomena or objects [15]. LSA can compared with a simple view of a neural network consisting of three layers: the first layer contains a set of words, the second – a set of documents, and the third, the middle (hidden layer), is a set of nodes with different weight coefficients connecting the first and second layers.

The basic idea of $k$-approximation of the latent-semantic approach to the matrix $A$ is to replace a matrix $\tilde{A}$ containing only $k$ – the first linearly-independents components of the matrix, and reflects the basic structure of the various dependencies presented in the $A$.

More formally, according to the singular decomposition theorem [16], a rectangular real matrix can have decomposed into a product of three matrices:

$$A=USV^T. \quad (7)$$

In this transformation $U$ and $V$ specially constructed orthogonal matrices $U$ and $V$ and $S$ is the diagonal matrix whose diagonal values represent the singular values of the $A$ matrix. This decomposition has a remarkable feature. If matrix $S$ leave only k the largest singular values, in the matrices $U$ and $V$ leave only the columns corresponding to these values, then the product of the resulting *matrices S, U* and $V$ will be the best approximation of the original matrix $A$ to the matrix (2) with a k rank [17].

The well-known MATCAD software package [18,19] provides an **svd** (A) function that implements the singular value decomposition of matrix A.

## 4 EXPERIMENTS

As an example of the use of LSA, the terms of reference for the development of intelligent DSS software for operational dispatching control of urban engineering networks (UEN) are taken. The analysis of the semantic content of the text of the technical task made 13 descriptors ($m$ = 13) as Groebner bases. They are the following words and phrases which are most often found, bearing the main semantic significance of the essence of the technical task: intelligent DSS, UEN, dispatcher, operational management, emergencies, fuzzy logic, situational uncertainty, fuzzy situational network, probabilistic transitions, multi-agent technology, decision-making, etc. [20].

In the formation of descriptors, the operation stemming is used [21]. In addition, of the descriptors were deleted stop characters, i.e. all conjunctions, particles, prepositions, etc.

## 5 RESULTS

On the descriptors, a search of documents was carried out on the Internet, the quantity of which after ranking and truncation at the first stage was $m$=13. As a result, we got matrix ($m$ x $n$):

$$A:=\begin{pmatrix}1&0&0&1&0&1&0&1&0\\0&0&0&1&0&0&0&1&0\\0&0&0&1&0&0&0&1&0\\0&0&1&0&1&0&0&0&1\\0&0&1&0&1&0&0&0&1\\1&0&0&1&0&1&0&1&0\\1&0&0&0&0&0&0&1&0\\0&0&1&0&1&0&0&0&1\\0&1&0&0&0&0&1&0&0\\0&0&1&0&0&0&1&0&0\\0&1&0&0&0&1&0&0&0\\0&1&0&0&0&0&1&0&0\\0&0&1&0&1&0&0&0&0\end{pmatrix}.$$

Next, we perform a singular value decomposition of the resulting matrix in MATHCAD, i.e. we present A as matrix "descriptors-documents" with the diagonal elements of the S matrix (singular values) ordered in descending order:

$$M:=svd2(A)$$

$$M=\begin{pmatrix}\{9,1\}\\\{13,9\}\\\{9,9\}\end{pmatrix}.$$

$$V:=M_3=\begin{pmatrix}-0.426&-0.046&-6.326\times10^{-3}&-0.535&-4.818\times10^{-3}&-0.372&-0.011&-0.625&-3.862\times10^{-3}\\4.892\times10^{-3}&-0.022&-0.653&6.277\times10^{-3}&-0.585&1.894\times10^{-3}&-0.089&7.407\times10^{-3}&-0.471\\-0.028&0.701&0.037&-0.064&-0.1&0.162&0.671&-0.086&-0.088\\0.217&0.244&-0.148&-0.277&0.115&0.684&-0.436&-0.327&0.13\\0.692&-0.325&0.224&-0.489&-0.123&-0.031&0.271&-0.016&-0.194\\-0.273&-0.336&0.438&0.293&-0.134&0.453&0.115&-0.312&-0.448\\0.031&-0.298&-0.141&0.17&-0.444&0.15&0.296&-0.239&0.705\\0.303&-0.119&-0.404&0.391&0.528&-0.122&0.228&-0.464&-0.133\\0.354&0.354&0.354&0.354&-0.354&-0.354&-0.354&-0.354&0\end{pmatrix}$$


DOI 10.15588/1607-3274-2021-3-14

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | –0.574 | $6.205\ 10^{-3}$ | $-6.205\ 10^{-3}$ | 0.199 | 0.13 | 0.163 | 0.157 | 0.251 | –0.539 |
| 2 | –0.34 | $4.148\ 10^{-3}$ | –0.066 | –0.405 | –0.423 | –0.019 | –0.096 | –0.166 | –0.402 |
| 3 | –0.34 | $4.148\ 10^{-3}$ | –0.066 | –0.405 | –0.423 | –0.096 | –0.096 | –0.166 | 0.402 |
| 4 | $-4.395\ 10^{-3}$ | –0.518 | –0.067 | 0.065 | –0.077 | –0.147 | 0.168 | –0.019 | 0.193 |
| 5 | $-4.395\ 10^{-3}$ | –0.518 | –0.067 | 0.065 | –0.077 | –0.147 | 0.168 | –0.019 | 0.039 |
| 6 | –0.574 | $-6.205\ 10^{-3}$ | $-6.352\ 10^{-3}$ | 0.199 | 0.13 | 0.163 | 0.157 | 0.251 | 0.539 |
| 7 | –0.308 | $3.728\ 10^{-3}$ | –0.05 | –0.074 | 0.565 | –0.595 | –0.292 | –0.371 | 0 |
| 8 | $-4.395\ 10^{-3}$ | –0.518 | –0.067 | 0.065 | –0.077 | –0.147 | 0.167 | –0.019 | –0.232 |
| 9 | –0.017 | –0.034 | 0.605 | –0.129 | –0.045 | –0.225 | $-2.822\ 10^{-3}$ | 0.252 | –0.048 |
| 10 | $-5.15\ 10^{-3}$ | –0.225 | 0.312 | –0.392 | 0.414 | 0.562 | 0.217 | –0.404 | 0 |
| 11 | –0.122 | $-6.162\ 10^{-3}$ | 0.381 | 0.622 | –0.298 | 0.119 | –0.207 | –0.55 | 0 |
| 12 | –0.017 | –0.034 | 0.605 | –0.129 | –0.045 | –0.225 | $-2.822\ 10^{-3}$ | 0.252 | 0.048 |
| 13 | $-3.264\ 10^{-3}$ | –0.375 | –0.028 | –0.023 | 0.085 | 0.309 | –0.82 | 0.287 | 0 |

$$S := diag(M_1) = \begin{pmatrix} 3.414 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.299 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.268 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.491 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.195 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.983 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.714 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.434 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$S2\_\sin g := submatrix(S,1,2,1,2) = \begin{pmatrix} 3.414 & 0 \\ 0 & 3.299 \end{pmatrix}.$$

$$S2\_\sin g := submatrix(M_3,1,2,1,9) = \begin{pmatrix} -0.426 & -0.046 & -6.326\times 10^{-3} & -0.535 & -4.818\times 10^{-3} & -0.372 & -0.011 & -0.625 & -3.862\times 10^{-3} \\ 4.892\times 10^{-3} & -0,022 & -0.653 & 6.277\times 10^{-3} & -0.585 & 1.894\times 10^{-3} & -0.089 & 7.407\times 10^{3} & -0.471 \end{pmatrix},$$

| | 1 | 2 |
|---|---|---|
| 1 | –0.574 | $6.205 \cdot 10^{-3}$ |
| 2 | –0.34 | $4.148 \cdot 10^{-3}$ |
| 3 | –0.34 | $4.148 \cdot 10^{-3}$ |
| 4 | $-4.395 \cdot 10^{-3}$ | –0.518 |
| 5 | $-4.395 \cdot 10^{-3}$ | –0.518 |
| 6 | –0.574 | $6.205 \cdot 10^{-3}$ |
| 7 | –0.308 | $3.728 \cdot 10^{-3}$ |
| 8 | $-4.395 \cdot 10^{-3}$ | –0.574 |
| 9 | –0.017 | –0.034 |
| 10 | $-5.15 \cdot 10^{-3}$ | –0.225 |
| 11 | –0.122 | $-6.162 \cdot 10^{-3}$ |
| 12 | –0.017 | –0.034 |
| 13 | $-3.264 \cdot 10^{-3}$ | –0.375 |

By according rules matrix multiplication, columns and rows corresponding to smaller singular values give the least contribution to the final result. In our case, according to (6) a two-dimensional singular value decomposition ($k = 2$) is conducted.

$$U2\_\sin g := submatrix(M_2,1,13,1,2) =$$

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0.835 | 0.089 | $-9.818 \cdot 10^{-3}$ | 1.049 | $-2.541 \cdot 10^{-3}$ | 0.729 | 0.02 | 1.225 | $-2.084 \cdot 10^{-3}$ |
| | 2 | 0.495 | 0.053 | $-1.598 \cdot 10^{-3}$ | 0.621 | $-2.541 \cdot 10^{-3}$ | 0.432 | 0.012 | 0.726 | $-1.968 \cdot 10^{-3}$ |
| | 3 | 0.495 | 0.053 | $-1.598 \cdot 10^{-3}$ | 0.621 | $-2.541 \cdot 10^{-3}$ | 0.432 | 0.012 | 0.726 | $-1.968 \cdot 10^{-3}$ |
| | 4 | $-1.968 \cdot 10^{-3}$ | 0.039 | 1.117 | $-2.701 \cdot 10^{-3}$ | 1.001 | $2.343 \cdot 10^{-3}$ | 0.152 | $-3.28 \cdot 10^{-3}$ | 0.806 |
| | 5 | $-1.968 \cdot 10^{-3}$ | 0.039 | 1.117 | $-2.701 \cdot 10^{-3}$ | 1.001 | $2.343 \cdot 10^{-3}$ | 0.152 | $-3.28 \cdot 10^{-3}$ | 0.806 |
| X= | 6 | 0.835 | 0.089 | $-9.818 \cdot 10^{-4}$ | 1.049 | $-2.541 \cdot 10^{-3}$ | 0.729 | 0.02 | 1.225 | $-2.084 \cdot 10^{-3}$ |
| | 7 | 0.448 | 0.048 | $-1.382 \cdot 10^{-3}$ | 0.563 | $-2.13 \cdot 10^{-3}$ | 0.391 | 0.011 | 0.658 | $-1.736 \cdot 10^{-3}$ |
| | 8 | $-1.968 \cdot 10^{-3}$ | 0.039 | 1.117 | $-2.701 \cdot 10^{-3}$ | 1.001 | $2.343 \cdot 10^{-3}$ | 0.152 | $-3.28 \cdot 10^{-3}$ | 0.806 |
| | 9 | 0.024 | $5.049 \cdot 10^{-3}$ | 0.073 | 0.03 | 0.065 | 0.021 | 0.01 | 0.035 | 0.052 |
| | 10 | $3.866 \cdot 10^{-3}$ | 0.017 | 0.485 | $4.754 \cdot 10^{-3}$ | 0.434 | $5.134 \cdot 10^{-3}$ | 0.066 | $5.502 \cdot 10^{-3}$ | 0.35 |
| | 11 | 0.178 | 0.019 | 0.016 | 0.223 | 0.014 | 0.155 | $6.499 \cdot 10^{-3}$ | 0.261 | 0.011 |
| | 12 | 0.024 | $5.049 \cdot 10^{-3}$ | 0.073 | 0.03 | 0.065 | 0.021 | 0.01 | 0.035 | 0.052 |
| | 13 | $-1.308 \cdot 10^{-3}$ | 0.028 | 0.809 | $-1.809 \cdot 10^{-3}$ | 0.725 | $1.799 \cdot 10^{-3}$ | 0.11 | $-2.204 \cdot 10^{-3}$ | 0.584 |


DOI 10.15588/1607-3274-2021-3-14

Matrix $X = \widetilde{A}$ (formula (6)). It is seen that it is a good approximation of the original A matrix.

Now, on the descriptors plane ($X$, $Y$), we get the points corresponding to the individual documents $d_i(x_i,y_i)$, where $i$=1, 2, …, 9 (Fig. 3). The coordinates of the documents taken from the $V_2$ *sign* matrix sign for values $\{x_i, y_i\}$($i$=1, 2, …, 9).

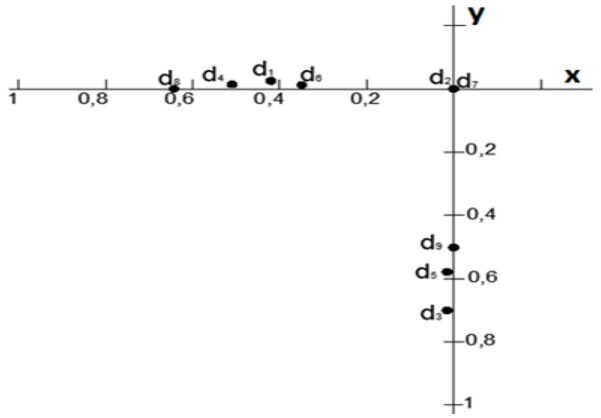


Figure 3 – Placing documents on the descriptors plane ($X$, $Y$)

From Fig. 3 it can be seen that the found documents are several groups (in our case 3), which are united by a group of similar keywords from separate descriptors. This made it possible to narrow down the quantity of documents to find specific information on the introduction of innovative elements in the developed software of intelligent DSS for operational dispatching control of urban engineering networks.

## 6 DISCUSSION

In practice, obviously, the quantity of groups will be much larger (even during the first stage), the phase space will be multidimensional, but the idea of LSA remains the same. In fact, this is a factor analysis that allows to determine the degree of correlation of the found documents with the constructed descriptors, and, consequently, with the technical task.

Given the possibility of using a multi-agent approach to reduce the time to search for the necessary information and increase its semantic value, the process of searching for Internet information in the IRS can represented by the following scheme, shown in Fig. 4 [22].

As a result, the search strategy for the proposed latent-semantic method of weighted descriptors can formulated in the form of an algorithm as follows:

Step 1. To take as an “ideal” the text of the terms of reference for the implementation of some innovative project in a specific subject area.

Step 2. To remove superfluous words, which described above, by using “stop-dictionary”.

Step 3. To construct conceptual descriptors with the help of the thesaurus of this subject area and the technical task.

Step 4. To arrange the conceptual descriptors in descending order of their frequency.

Step 5. To determine the frequency, range of the most significant descriptors (usually 10–20 descriptors).

Step 6. To make a request and get a rectangular matrix “descriptors-documents”.

Step 7. According to the formula (5), to arrange the documents in descending order of relevance.

Step 8. To spend on the LSA k-approximation.

Step 9. To put the selected documents in the knowledge model of the subject area.

Steps 6–9 are implemented by intelligent agents automatically and periodically in the search engines allocated to them.

Developed a method of extracting object-oriented information from Internet on based weighted descriptors with the use of Groebner basis. An automated multi-agent procedure for extracting information from the Internet with a semantic analysis of its semantic content developed.

## CONCLUSION

One of the key points of the search process is the formation of an effective query and an automated procedure for the analysis of information extracted from the Internet for this subject area.

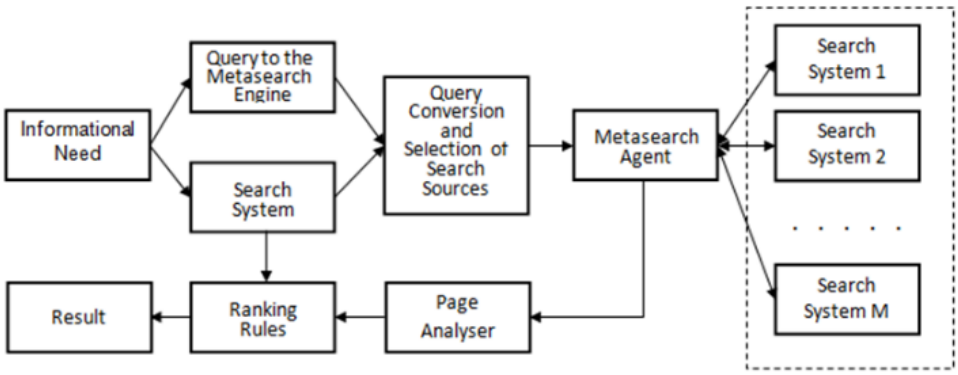


Figure 4 – Multi-agent Search Scheme in Internet Retrieval Systems


DOI 10.15588/1607-3274-2021-3-14

**The scientific novelty** of the results is that for the first time a latent-semantic method of weighted descriptors is proposed, which allows extracting the most significant documents in terms of meaning and meaning, very close to the subject area. The latent-semantic method of weighted descriptors proposed, which allows to extract the most significant in meaning and significance documents that are very close to the subject area. The method based on the idea of Groebner basis, which statistically constructed conceptual descriptors. The choice of the number of descriptors determined by the number of constructed Groebner basis. As the Internet search by the selected of descriptors the initial number of documents turns out to be quite large, used singular value decomposition by *k*-approximation of latent-semantic analyses.

**The practical significance** of the obtained results lies in the fact that the use of the latent-semantic method of weighted descriptors proposed in the work can significantly reduce the search time for the necessary information, taking into account its semantic value for this subject area. To select the documents closest to this subject area carried out on the k-approximation of latent semantic analysis.

**Prospects for further research** proposed by the authors in [23]. They consist in the construction of evolutionary models of knowledge based on the logical evaluation of data obtained from a neural network with neurons having memory and integrated logic. At the same time, the integrated logic is implemented on the basis of a genetic algorithm that processes the updated knowledge model and improves each next generation of "genes" by weighing semantic data based on the superposition of the reference reaction to the situation and the assessment of the situation by the current generation of "genes". It should note that the information obtained in this way serves as the basis for the correct choice of research methods for the innovative development of organizational and technical systems in this subject area [24].

## ACKNOWLEDGEMENT

The work supported by the State Budget Research Project of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" "Developing specialized knowledge bases for recursive parser of weakly connected originated natural language text information and Web-applications" (a state registration number 0110U002409).

## REFERENCES

1. Chu H., Rosenthal M. Search engines for the World Wide Web: A comparative study and evaluation methodology, *Proceedings of the annual meeting-american society for information science: journal*, 2009, Vol. 33, pp. 127–135.
2. Singhal A. Modern Information Retrieval: "A brief Overview", *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2001, Vol. 24, No. 4, pp. 35–43.
3. Gandal N. The dynamics of competition in the internet search engine market, *International journal of industrial organization,* 2001, Vol. 19, pp. 1103–1117. doi:10.1016/S0167-7187(01)00065-0
4. Tarakeswar M. K., Kavitha M. D. Search Engines:A Study // Journal of Computer Applications (JCA) : journal, 2011, Vol. 4, No. 1, pp. 29–33.
5. Mikhalev A. I., Stenin A. A., Shitikova I. G., Lemeshko V. A. Intellectual multi-agent system of formation of the subject-oriented evolutionary model of knowledge, *System technologies*, 2018, No. 3 (116), pp. 57–63.
6. Agirre E., Cer D., Diab M. et al. A pilot on semantic textual similarity, *The 6-th International Workshop on Semantic Evaluation, Atlanta*, USA, 2012, pp. 385–393.
7. Bao J., Shen J., Liu X., et al. Semantic Sequence Kin: A Method of Document Copy Detection, *Advances In Knowledge Discovery and Data Mining. Lecture Notes in Artificial Intelligence (LNAI)*. Sydney, Australia, 2004, Vol. 3056, pp. 529–538.
8. Floridi L. Semantic Web, *A Philosophical Assessment, Episteme,* 2009, Vol. 6, No. 1, pp. 25–37.
9. Berners-Lee T., Hendler J., Lassila O. The semantic web, *Scientific American*, 2001, pp. 29–37.
10. Kalchenko D. Intelligent agents of semantic Web, *Computer press-confer,* 2004, No. 10, pp. 26–32.
11. Etzioni O., Weld D., "Intelligent agents on the internet/ O.Etzioni Weld, *Fact, Fiction, and Forecast", IEEE Expert*, No. 4, 1995, pp. 44–49.
12. Wentia Li. Random Texts Exhibit Zipf's Law, *Like Word Frequency Distribution Santa Fe institute. NM 87501*, 1992, Vol. 38, No. 6, pp. 1842–1845.
13. Kechedzhy K. E., Ustenko O. V., Yampol'ski V. A. Rank distributions of words in additive many-step Markov chains and the Zipf, *Physical review*, 2005, Vol.72, pp. 1–6.
14. Gerdt V. P. Groebner bases and innovative methods for algebraic and differential equations, *Mathematics and computers in modelling*, 1997, Vol. 25, No. 8/9, pp. 75–90
15. Orlov A.I. Organizational and economic modeling. P.2: Expert estimations. Moscow, Bauman Moscow State Technical University 2011, 486 p.
16. Golub, J. Matrix calculus. Moscow, Mir, 1999, 548 p.
17. Alston S. Hausholder Unitary triangularization of an asymmetric matrix, *Journal of New Technologies in Computational Systems*, 1958, ACM, 5 (4), pp. 339–342. DOI:10.1145/320941.320947
18. Jones K. S. Statistical interpretation of term specificity and its application to search, *Journal. MCB University Documentation*, 2004. Vyp. 60, № 5, pp. 493–502.
19. Matthews D., Curtis D., Fink K. Numerical Methods. Using MATLAB. Numerical Methods: Using MATLAB. 3rd ed. Moscow, Williams Publisher, 2001, 720 p.
20. Charles Henry Edwards Penney, David E. Differential Equations and the Eigenvalue Problem: Modeling and Computation with Mathematica, Maple and MATLAB.3rd edition. Moscow, Williams Publishing House, 2007, 1104 p.
21. Alexa M., Zuell C. Text Analysis Software: Commonalities, Differences and Limitations, The Results of a Review, *Springer Netherlands*, 2000, Vol. 34 (3), pp. 299–321.
22. Dubinsky A. G. Model of multi-agent information retrieval system in the global network, *Artificial intelligence,* 1999, No. 3, pp. 271–279.
23. Stenin A. A., Pasko V. P., Lemeshko V. A. Neurosemantic approach to building automated information retrieval systems, *Adaptive automatic control systems*, 2019, No. 1(34), pp. 125–130.


DOI 10.15588/1607-3274-2021-3-14

24. Kirichok P. O., Strutinskii S. V., Ol'inik V. G. Special methods of scientific research. National Technical University of Ukraine "Kyiv Polytechnic Institute". Kyiv, ArtEk, 2016, 592 p. ISBN 978-617-7264-28-5

Received 26.06.2021.
Accepted 20.08.2021.

УДК 004.91

## МУЛЬТИАГЕНТНА ЛАТЕНТНО-СЕМАНТИЧНА ІНТЕРНЕТ-ТЕХНОЛОГІЯ ФОРМУВАННЯ ПРЕДМЕТНО-ОРІЄНТОВАНОЇ МОДЕЛІ ЗНАНЬ

**Стенін О. А. –** д-р техн. наук, професор кафедри технічної кібернетики Київського політехнічного інституту. Ігор Сікорський, Київ, Україна.

**Пасько В. П. –** канд. техн. наук, доцент кафедри технічної кібернетики Київського політехнічного інституту. Ігор Сікорський, Київ, Україна.

**Солдатова М. А. –** канд. техн. наук, старша викладачка кафедри технічної кібернетики Київського політехнічного інституту. Ігор Сікорський, Київ, Україна.

**Дроздович І. Г. –** канд. техн. наук, старший науковий співробітник Інституту телекомунікацій і глобального інформаційного простору НАН України.

### АНОТАЦІЯ

**Актуальність**. У статті пропонується латентно-семантична технологія вилучення інформації з інтернет-ресурсів, що дозволяє обробляти інформацію природною мовою, а також заснований на ній мультиагентний алгоритм пошуку. Актуальність даного підходу до пошуку предметно-орієнтованої інформації визначається тим, що в даний час пряме лексичне порівняння запитів з індексами документів не повною мірою задовольняє розробника. Об'єкт дослідження – мультиагентний латентно-семантичний алгоритм пошуку предметно-орієнтованої інформації. Мета роботи – підвищення ефективності формування адекватної даної предметної області моделі знань.

**Метод.** Запропонована латентно-семантична технологія, заснована на розробленому авторами методі зважених дескрипторів. Основна відмінність від існуючих методів полягає в тому, що аналіз слів, що зустрічаються в тексті як по частотності, так і з урахуванням семантики, здійснюється шляхом підбору відповідних дескрипторів, що підвищує якість знайденої інформації.

**Результати**. Розроблена латентно-семантична технологія пошуку інформації апробована в задачі побудови моделі знань автоматизованої СППР для оперативно-диспетчерського управління міськими інженерними мережами (ГІС). Проведене моделювання пошуку предметно-орієнтованої інформації даної предметної області показало ефективність розробленого підходу.

**Висновки.** Підвищення ефективності пошуку і семантичного наповнення предметно-орієнтованої інформації моделі знань даної предметної області досягається за рахунок використання в даній технології методу зважених дескрипторів, заснованого на законах Зіпфа. Перспективи подальших досліджень полягають у побудові еволюційних моделей знань і підвищення якості оновлюваної інформації.

**КЛЮЧОВІ СЛОВА**: інтернет-ресурси, інформаційний пошук, закони Зіпфа, базиси Гребнера, інтелектуальні агенти, зважені дескриптори, латентно-семантичний аналіз, мультиагентна процедура автоматичного пошуку.

УДК 004.91

## МУЛЬТИАГЕНТНЫЕ ЛАТЕНТНОГО СЕМАНТИЧЕСКОГО ИНТЕРНЕТ-ТЕХНОЛОГИЯ ФОРМИРОВАНИЯ ПРЕДМЕТНО-ОРИЕНТИРОВАННАЯ МОДЕЛЬ ЗНАНИЙ

**Стенин А. А**. – д-р техн. наук, профессор кафедры технической кибернетики Киевского политехнического института. Игорь Сикорский, Киев, Украина.

**Пасько В. П.** – канд. техн. наук, доцент кафедры технической кибернетики Киевского политехнического института. Игорь Сикорский, Киев, Украина.

**Солдатова М. А.** – канд. техн. наук, старший преподаватель кафедры технической кибернетики Киевского политехнического института. Игорь Сикорский, Киев, Украина.

**Дроздович И. Г.** – канд. техн. наук, старший научный сотрудник Института телекоммуникаций и глобального информационного пространства НАН Украина.

### АННОТАЦИЯ

**Актуальность.** В статье предлагается латентно-семантическая технология извлечения информации из интернет-ресурсов, позволяет обрабатывать информацию на естественном языке, а также основанный на ней Мультиагентный алгоритм поиска. Актуальность данного подхода к поиску предметно-ориентированной информации определяется тем, что в настоящее время прямое лексическое сравнение запросов с индексами документов не в полной мере удовлетворяет разработчика. Объект исследования – мультиагентный латентно-семантический алгоритм поиска предметно-ориентированной информации. Цель работы – повышение эффективности формирования адекватной данной предметной области модели знаний.

**Метод.** Предложенная латентно-семантическая технология, основанная на разработанном авторами методе взвешенных дескрипторов. Основное отличие от существующих методов состоит в том, что анализ слов, встречающихся в тексте как по


DOI 10.15588/1607-3274-2021-3-14

частотности, так и с учетом семантики, осуществляется путем подбора соответствующих дескрипторов, повышает качество найденной информации.

**Результаты.** Разработанная латентно-семантическая технология поиска информации апробирована в задаче построения модели знаний автоматизированной СППР для оперативно-диспетчерского управления городскими инженерными сетями (ГИС). Проведенное моделирование поиска предметно-ориентированной информации данной предметной области показало эффективность разработанного подхода.

**Выводы.** Повышение эффективности поиска и семантического наполнения предметно-ориентированной информации модели знаний данной предметной области достигается за счет использования в данной технологии метода взвешенных дескрипторов, основанного на законах Зипфа. Перспективы дальнейших исследований заключаются в построении эволюционных моделей знаний и повышения качества обновляемой информации.

**КЛЮЧЕВЫЕ СЛОВА:** интернет-ресурсы, информационный поиск, законы Зипфа, базисы Гребнера, интеллектуальные агенты, взвешенные дескрипторы, латентно-семантический анализ, мультиагентная процедура автоматического поиска.

**ЛІТЕРАТУРА / ЛІТЕРАТУРА**

1. Чу Х. Поисковые системы для Всемирной паутины: сравнительное исследование и методология оценки / Х. Чу, М. Розенталь // Труды ежегодного совещания. Американское общество информационных наук: науч.-техн.сб. – 2009. – Том 33. – С. 127–135
2. Сингхал Амит Современный информационный поиск: «Краткий обзор» / Сингхал Амит // Бюллетень Технического комитета IEEE. Компьютерное общество по инженерии данных. – 2001. – Том 24, № 4. – С. 35–43.
3. Гэндал Н. Динамика конкуренции на рынке поисковых систем Интернета / Н. Гэндал // Международный журнал промышленной организации. –2001. – Том 19. – С. 1103–1117. DOI: 10.1016/S0167-7187(01)00065-0
4. Tarakeswar M. K. Поисковые системы: Исследование / M. K. Tarakeswar, M. D. Kavitha // Журнал компьютерных приложений. – 2011. – Том 4, № 1. – С. 29–33.
5. Интеллектуальная мультиагентная система формирования предметно-ориентированной эволюционной модели знаний / [А. И. Михалев, А. А. Стенин, И. Г. Шитикова, В. А. Лемешко] // Системные технологии: науч.-техн.сб. – 2018. – Вып. 3(116). – С. 57–63.
6. Пилотный проект по семантическому текстовому сходству / [Э. Агирре, Д. Сер, М. Диаб и др.] // Семантические оценки: 6-й Международный семинар, Атланта, США, 2012. – С. 385–393
7. Семантическая последовательность Kin: Метод обнаружения копий документов / [Дж. Бао, Д. Шен, Х. Лью, X. Liu, и др.] // Достижения в области обнаружения знаний и интеллектуального анализа данных: Лекционные заметки по искусственному интеллекту. – Сидней, Австралия, 2004. – Том 3056. – С. 529–538.
8. Флориди Л. Веб 2.0 против Семантической сети / Л. Флориди // Философская оценка. – 2009. – Вип. 6, № 1. – С. 25–37.
9. Бернерс-Ли Т. Семантическая паутина / Т. Бернерс-Ли, Д. Хендлер, О. Лассила // Американская Наука. – 2001. – С. 29–37.
10. Кальченко Д. Интеллектуальные агенты семантического Web'a / Д. Кальченко // Компьютер Пресс. – 2004. – Вып. 10. – С. 26–32.
11. Эциони О. Интелллектуальные агенты в Интернете / О. Эциони, Д. Велд // Факты, вымысел и прогноз, IEEE Эксперт. – 1995. – № 4. – С. 44–49.
12. Wentain Li. Random Texts Exibition Zipf's Law / Li. Wentain // Like Word Frequency Distribution. Santa Fe institute. NM 87501. – 1992. – Vol. 38, №6. – P. 1842–1845.
13. Кечеджи К. Е. Ранговые распределения слов в аддитивной форме. Многошаговые цепи Маркова и закон Ципфа / К. Е. Кечеджи, О. В. Устенко, В. А. Ямпольский // Физическое обозрение. – 2005. – Вип. 72, – С. 1–6.
14. Гердт В. П. Основы Гребнера и инвойтивные методы для алгебраических и дифференциальных уравнений / В. П. Гердт // Математика и компьютеры в моделировании – 1997. – Вып 25, № 8/9. – С 75–90.
15. Орлов А. И. Организационно-экономическое моделирование / А. И. Орлов. – Ч. 2: Экспертные оценки. – М. : МГТУ им. Н. Э. Баумана, 2011. – 486 с.
16. Голуб Дж. Матричные исчисления / Дж. Голуб. – М. : Мир, 1999. – 548 с.
17. Олстон С. Унитарная триангуляризация несимметричной матрицы / Олстон С. Хаусхолдер // Журнал о новых технологиях в вычислительных системах – 1958. – ACM, 5 (4). – С. 339–342. DOI: 10.1145/320941.320947
18. Джонс К. С. Статистическая интерпретация специфичности термина и ее применение в поиске / К. С. Джонс // Журнал. Документация «MCB University. 2004. – Вып. 60, № 5. – С. 493–502.
19. Мэтьюз Д.Численные методы. Использование MATLAB / Д. Мэтьюз, Д. Куртис, К. Финк. – Numerical Methods: Using MATLAB. – 3-е изд. – М. : Изд-во «Вильямс», 2001. – 720 с.
20. Чарльз Генри Эдвардс Пенни. Дифференциальные уравнения и проблема собственных значений: моделирование и вычисление с помощью Mathematica / Чарльз Генри Эдвардс, Дэвид Э. – Maple и MATLAB. 3-е издание. – М. : Изд-во «Вильямс», 2007. – 1104 с.
21. Алекса М. Программное обеспечение для анализа текста: общие черты, различия и ограничения: Результаты обзора / M. Alexa, C. Zuell // Springer Netherlands. – 2000. – Вып. 34 (3). – С. 299–321.
22. Дубинский А. Г. Модель многоагентной информационно-поисковой системы в глобальной сети / А. Г. Дубинский // Искусственный интеллект. – 1999. – № 3. – С. 271–279.
23. Стенин А. А. Нейросемантический подход к построению автоматизированных информационно-поисковых систем / А. А. Стенин, В. П. Пасько, В. А. Лемешко // Адаптивные системы автоматического управления. – 2019. – №1 (34).–С. 125–130.
24. Спеціальні методи наукових досліджень / П. О. Киричок, С. В. Струтинський, В. Г. Олійник ; Нац. техн. ун-т України «Київ. політехн. ін-т». – Київ : АртЕк, 2016. – 592 с. ISBN 978-617-7264-28-5


DOI 10.15588/1607-3274-2021-3-14