DEXA 2015

TIR 2015
12 International Workshop on
Text-based Information Retrieval
In conjunction with DEXA 2015

INDUCTIVE BUILDING OF SEARCH RESULTS RANKING MODELS TO ENHANCE THE RELEVANCE OF TEXT INFORMATION RETRIEVAL

Vyacheslav ZOSIMOV, Cand. Sci[1]
Volodymyr STEPASHKO, Dr. Sci[2]
Oleksandra BULGAKOVA, Cand. Sci[1]

[1]Mykolaiv National University

[2]International Research and Training Centre for Information Technologies and Systems of the National Academy of Sciences of Ukraine

Description of the developed technology

For solving the problem of improving the efficiency of relevant information search on the Internet it is necessary to develop a system that provides:

1. High search precision rates, it means the lack of search spam and artificially promoted sites among search results.
2. Search completeness rates not worse than by current search engines.
3. High performance of search results analysis.
4. Wide capacity of software customization by user.

Proposed technology consists of the three main phases:

Phase 1. Information collecting.

Phase 2. Sifting the commercial information.

Phase 3. Results ranking.

Sifting the commercial information

Main groups of commercial sites:

1. Online stores.

2. Sites that offer access to the information for a fee or watching ads

3. Websites of companies.

4. Message boards.

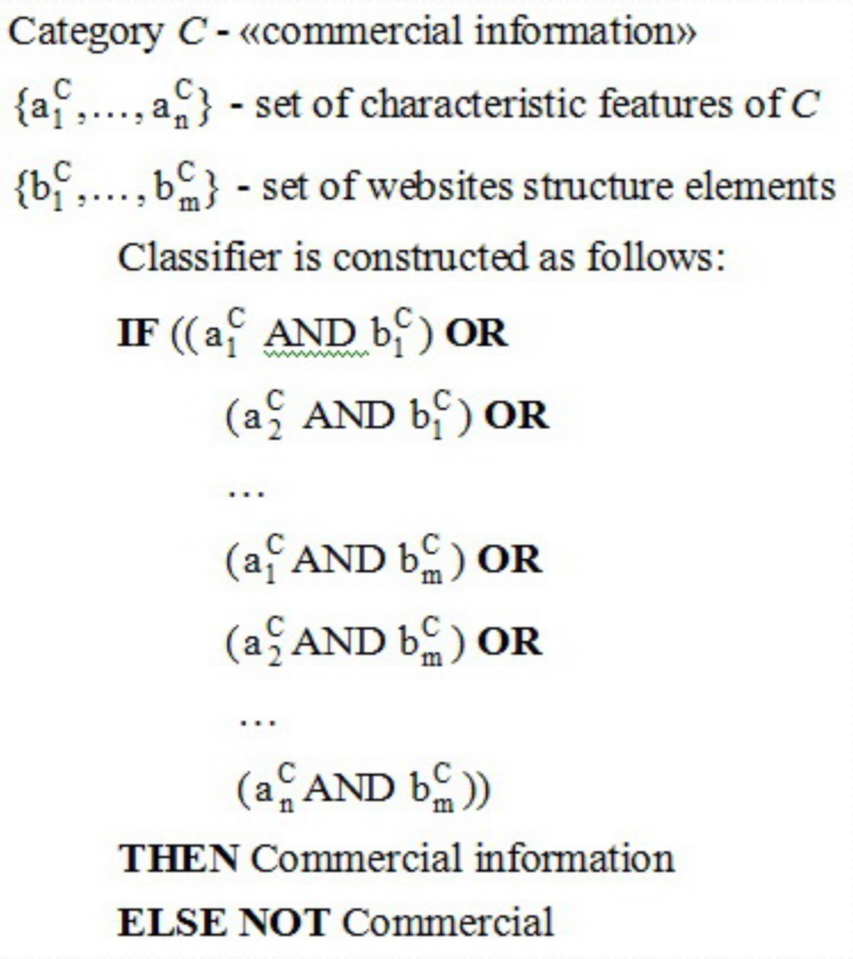
Structural elements of the websites:

- meta tags, the path to Java-script and image of design;
- titles, meta description, keywords;
- shopping cart;
- text on the home page;
- navigation elements.

Features of the commercial sites:

- Presence among the navigation elements the following items: "Services", "Company", "Price List", "Dealer", "Activity of the company", "Employment", "order service", "Jobs“, "Price", "Our customers”;
- Usage of the specialized CMS for creating online stores.

DNF classifier:



Category *C* - «commercial information»

$\{a_1^C, \ldots, a_n^C\}$ - set of characteristic features of *C*

$\{b_1^C, \ldots, b_m^C\}$ - set of websites structure elements

Classifier is constructed as follows:

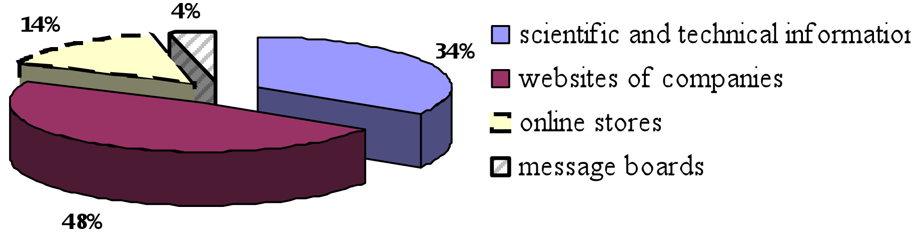
```
IF (($a_1^C$ AND $b_1^C$) OR
        ($a_2^C$ AND $b_1^C$) OR
        …
        ($a_1^C$ AND $b_m^C$) OR
        ($a_2^C$ AND $b_m^C$) OR
        …
        ($a_n^C$ AND $b_m^C$))
THEN Commercial information
ELSE NOT Commercial
```

Experimental results on more effective search of the scientific and technical information

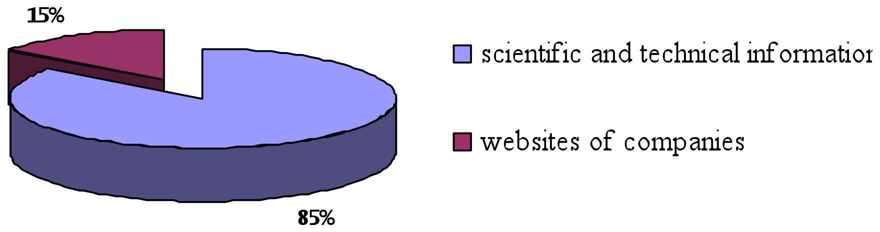
Information is searched by keywords at first using search engine google and then using the proposed technology.

The purpose of the experiment: comparative effectiveness of information search by search engine google and the proposed technology.

Query: "Information Security"

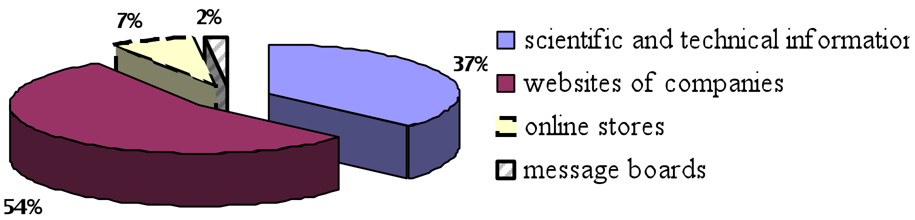


google.com.ua

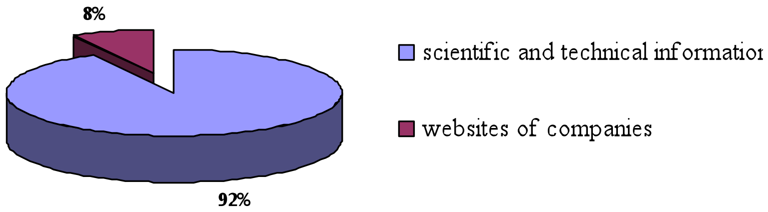


Developed technology

Query: «Programming 1C»



google.com.ua



Developed technology

ON THE GIA GMDH AS A RESEARCH TOOL

The scientific school of Inductive Modelling was originated by Prof. Aleksey Ivakhnenko. The very first article on his Group Method of Data Handling was published by him in 1968.

GMDH as a self-organizing data mining tool is based on the main principles:

- automatic generation of inductively complicated variants
- non-final decisions and successive selection of models of optimum complexity
- using so called external criteria of cross-validation type based on the division of a dataset into at least two parts.

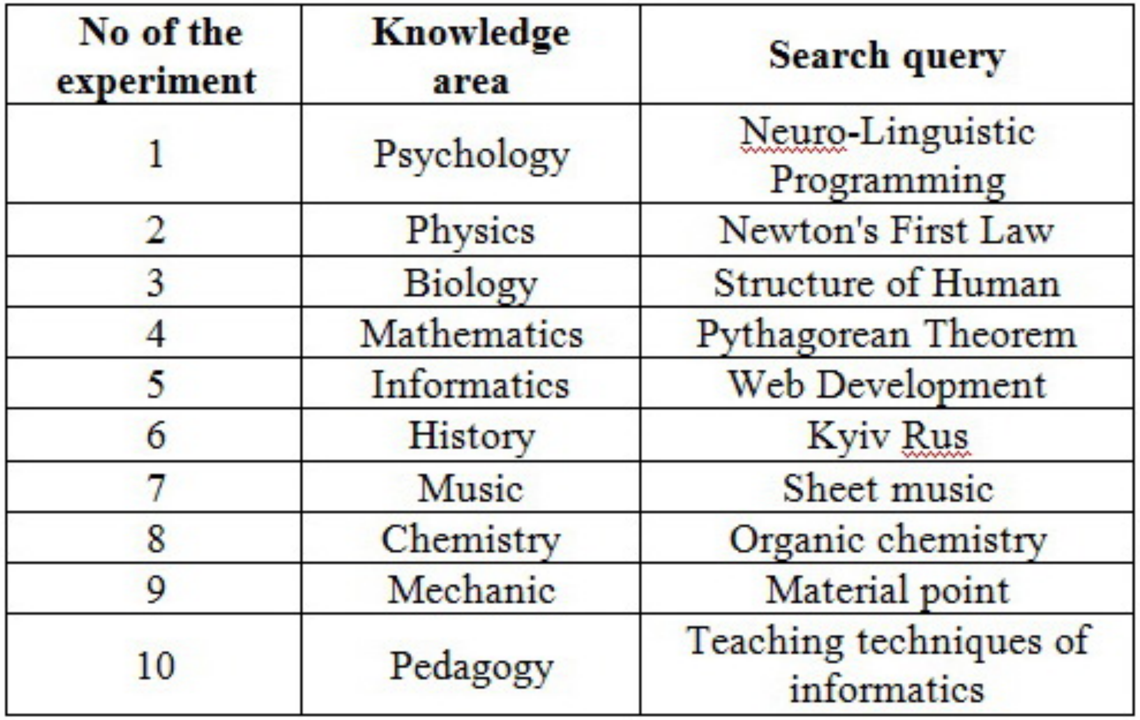
The classical multilayered iterative algorithm MIA GMDH [4] is based on the nature inspired idea of mass biological selection with pairwise account of features. Currently it is considered as a *polynomial neural network* (PNN) notable by self-organization of both its architecture and parameters.

GMDH has advantages of automatic formation of the network structure, simplicity and speed of parameters estimation as well as the possibility to "fold" the adjusted network into an explicit mathematical model.

The generalized iterative algorithm GIA GMDH has constructed enclosing typical known and new architectures of the iterative procedures of both multilayered and relaxational type with combinatorial optimizing the partial descriptions (quadratic transfer functions).

BUILDING RANKING MODELS FOR SPECIFIC FIELDS

The complete lists of knowledge areas and search queries participated in modeling experiments

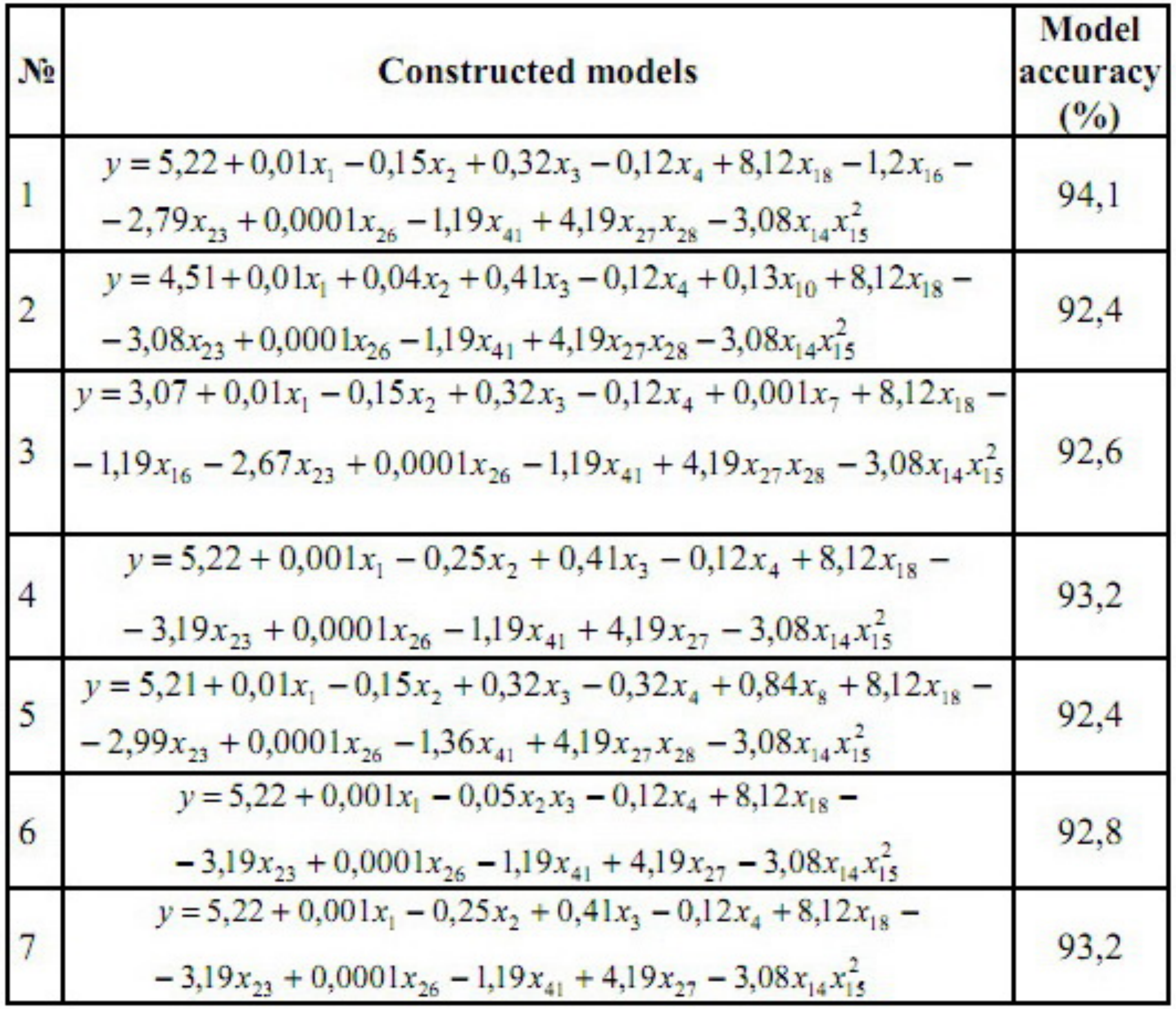


No of the experiment	Knowledge area	Search query
1	Psychology	Neuro-Linguistic Programming
2	Physics	Newton's First Law
3	Biology	Structure of Human
4	Mathematics	Pythagorean Theorem
5	Informatics	Web Development
6	History	Kyiv Rus
7	Music	Sheet music
8	Chemistry	Organic chemistry
9	Mechanic	Material point
10	Pedagogy	Teaching techniques of informatics

In the process of ranking models construction, 64 lecturers from various departments of Mykolaiv National University (Ukraine) have participated. All participants were divided into 10 groups of 6 to 8 people. Each group analyzed one search query in their field of knowledge. Each expert sorts by relevance first 50 sites with scientific-and-engineering information obtained from the Yandex search engine results page (SERP) after sifting the commercial spam information.

BUILDING RANKING MODELS FOR SPECIFIC FIELDS

Results of the experiment for ranking models building



№	Constructed models	Model accuracy (%)
1	$y = 5{,}22 + 0{,}01x_1 - 0{,}15x_2 + 0{,}32x_3 - 0{,}12x_4 + 8{,}12x_{18} - 1{,}2x_{16} -$ $- 2{,}79x_{23} + 0{,}0001x_{26} - 1{,}19x_{41} + 4{,}19x_{27}x_{28} - 3{,}08x_{14}x_{15}^2$	94,1
2	$y = 4{,}51 + 0{,}01x_1 + 0{,}04x_2 + 0{,}41x_3 - 0{,}12x_4 + 0{,}13x_{10} + 8{,}12x_{18} -$ $- 3{,}08x_{23} + 0{,}0001x_{26} - 1{,}19x_{41} + 4{,}19x_{27}x_{28} - 3{,}08x_{14}x_{15}^2$	92,4
3	$y = 3{,}07 + 0{,}01x_1 - 0{,}15x_2 + 0{,}32x_3 - 0{,}12x_4 + 0{,}001x_7 + 8{,}12x_{18} -$ $- 1{,}19x_{16} - 2{,}67x_{23} + 0{,}0001x_{26} - 1{,}19x_{41} + 4{,}19x_{27}x_{28} - 3{,}08x_{14}x_{15}^2$	92,6
4	$y = 5{,}22 + 0{,}001x_1 - 0{,}25x_2 + 0{,}41x_3 - 0{,}12x_4 + 8{,}12x_{18} -$ $- 3{,}19x_{23} + 0{,}0001x_{26} - 1{,}19x_{41} + 4{,}19x_{27} - 3{,}08x_{14}x_{15}^2$	93,2
5	$y = 5{,}21 + 0{,}01x_1 - 0{,}15x_2 + 0{,}32x_3 - 0{,}32x_4 + 0{,}84x_8 + 8{,}12x_{18} -$ $- 2{,}99x_{23} + 0{,}0001x_{26} - 1{,}36x_{41} + 4{,}19x_{27}x_{28} - 3{,}08x_{14}x_{15}^2$	92,4
6	$y = 5{,}22 + 0{,}001x_1 - 0{,}05x_2x_3 - 0{,}12x_4 + 8{,}12x_{18} -$ $- 3{,}19x_{23} + 0{,}0001x_{26} - 1{,}19x_{41} + 4{,}19x_{27} - 3{,}08x_{14}x_{15}^2$	92,8
7	$y = 5{,}22 + 0{,}001x_1 - 0{,}25x_2 + 0{,}41x_3 - 0{,}12x_4 + 8{,}12x_{18} -$ $- 3{,}19x_{23} + 0{,}0001x_{26} - 1{,}19x_{41} + 4{,}19x_{27} - 3{,}08x_{14}x_{15}^2$	93,2

Ranking models was built using GIA GMDH based on 42 features which are typically used in search engines ranking.

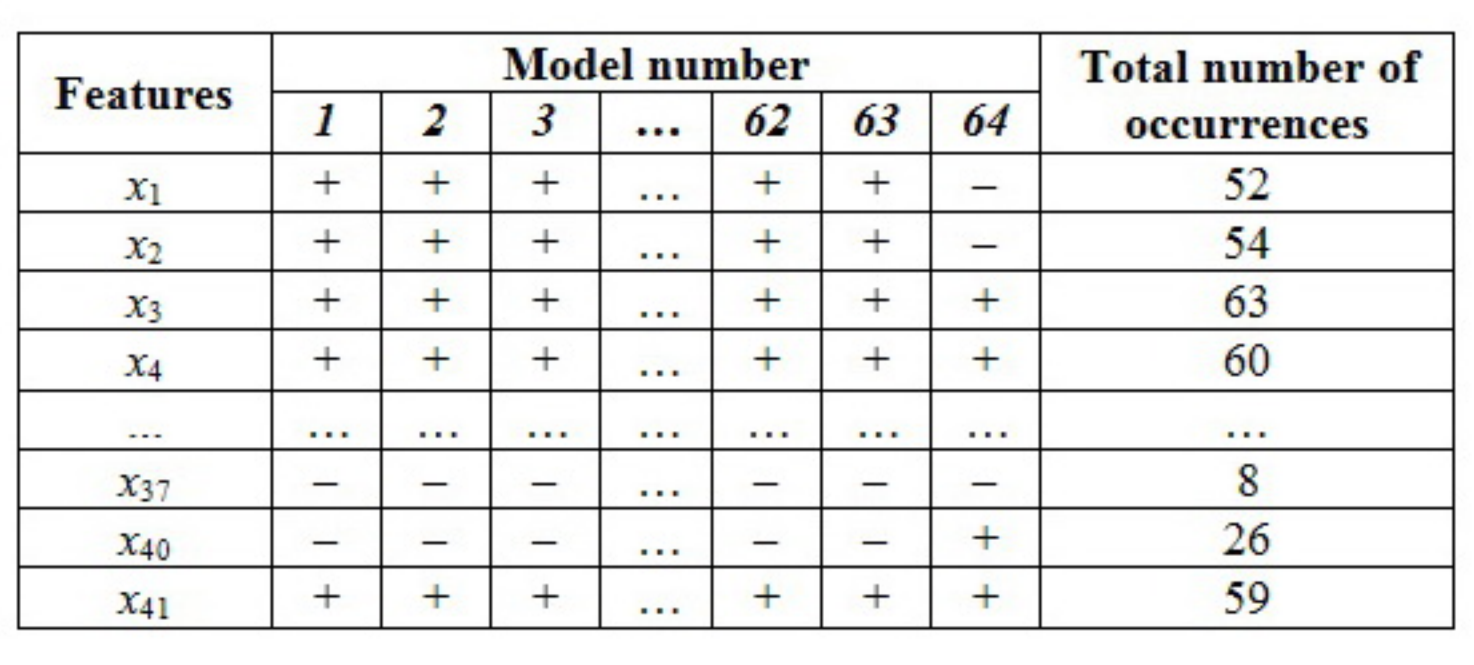
The accuracy of models varies from 92,4% to 95,8%.

After that new queries in the same fields can be handled without experts with high accuracy.

This means that the most relevant target sources will already be placed at the top of the modified search delivery.

Building the universal ranking model

The frequency of feature’s occurrence in ranking models built during the experiments



Features	Model number							Total number of occurrences
	1	*2*	*3*	…	*62*	*63*	*64*	
x_1	+	+	+	…	+	+	–	52
x_2	+	+	+	…	+	+	–	54
x_3	+	+	+	…	+	+	+	63
x_4	+	+	+	…	+	+	+	60
…	…	…	…	…	…	…	…	…
x_{37}	–	–	–	…	–	–	–	8
x_{40}	–	–	–	…	–	–	+	26
x_{41}	+	+	+	…	+	+	+	59

To sift uninformative features, the threshold of their occurrence frequency in the formulae was determined at an acceptable level 50. Features with the frequency 50 to 64 were selected as informative and the rest were excluded.

Building the universal ranking model

A new ranking model was built using GIA GMDH based on the extracted informative features from Table III used as input arguments. The following model has built:

$$y = 4{,}67 - 0{,}01\tilde{o}_1 + 0{,}02\tilde{o}_2 + 0{,}72\tilde{o}_3\tilde{o}_4 - 1{,}97\tilde{o}_{18}\tilde{o}_{23} +$$
$$+ 2{,}01\tilde{o}_{22} + 0{,}0001\tilde{o}_{26} - 1{,}1\tilde{o}_{41} + 4{,}07\tilde{o}_{27}\tilde{o}_{28} - 3{,}15\tilde{o}_{14}\tilde{o}_{15}^2$$

Accuracy of the model is 88.6%, slightly lower as compared to any of the built models based on the learning sample.

Yandex ranking model rebuilt using GIA GMDH

$$y = 7{,}12 + 1{,}01\tilde{o}_3 + 0{,}12\tilde{o}_4 + 0{,}000001\tilde{o}_7 - 2{,}69\tilde{o}_{12} + 8{,}12\tilde{o}_{22} + 2{,}79\tilde{o}_{27} +$$

$$+ 0{,}001\tilde{o}_{28} - 48{,}19\tilde{o}_{35} - 2{,}001\tilde{o}_{41} - 12{,}22\tilde{o}_{42}\tilde{o}_6 - 3{,}08\tilde{o}_{14}\tilde{o}_{15}^2 + 0{,}04\tilde{o}_{37}\tilde{o}_{38}\tilde{o}_{39}$$

The most influencing in the Yandex ranking model are the following 16 factors:

***x*3** – the ratio of the total words number on the site to the keywords number on the site;

***x*4** – the ratio of the total words number on the page to the keywords number on the page;

***x*6** – topic’s popularity;

***x*7** – requests number of the particular keyword for a certain period of time;

***x*12** – website age;

***x*14** – frequency of updating site information;

***x*22** – keywords font size;

***x*27** – keywords presence in title;

***x*28** – keywords presence in meta-tags;

***x*35** – match website keywords to the search engine page directory in which site is;

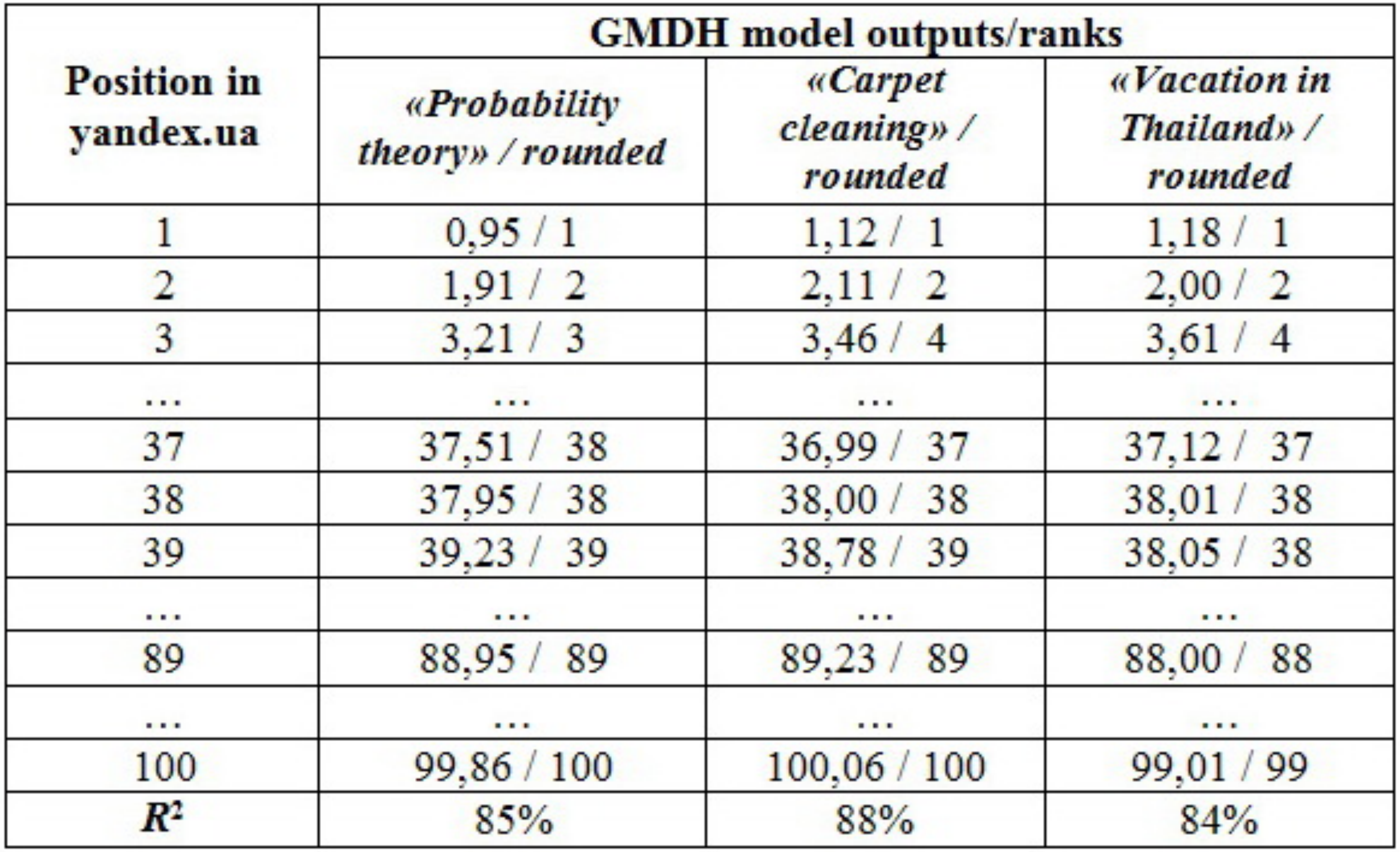
***x*41**– number of external links, containing keywords in its title;

***x*42** – Yandex citation index;

Comment: The most influenced in the Yandex ranking has the external factors (x_4, x_6, x_7, x_{12}, $x\ _{14}$, x_{41}, h_{42}) as compared to internal ones.

This data are automatically collected from several sources: Yandex database, html-code analysis and independent services for the sites content analysis .

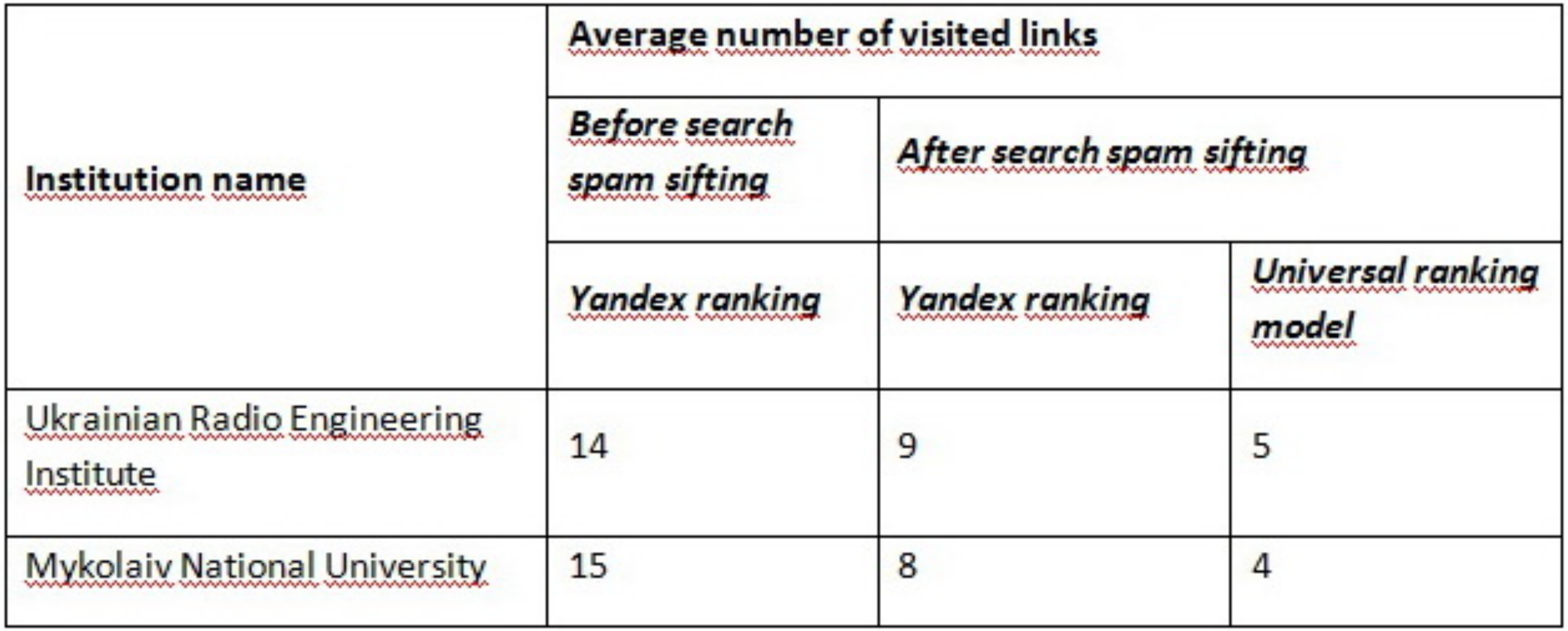
Testing the constructed Yandex ranking model



Position in yandex.ua	GMDH model outputs/ranks		
	«Probability theory» / rounded	*«Carpet cleaning» / rounded*	*«Vacation in Thailand» / rounded*
1	0,95 / 1	1,12 / 1	1,18 / 1
2	1,91 / 2	2,11 / 2	2,00 / 2
3	3,21 / 3	3,46 / 4	3,61 / 4
…	…	…	…
37	37,51 / 38	36,99 / 37	37,12 / 37
38	37,95 / 38	38,00 / 38	38,01 / 38
39	39,23 / 39	38,78 / 39	38,05 / 38
…	…	…	…
89	88,95 / 89	89,23 / 89	88,00 / 88
…	…	…	…
100	99,86 / 100	100,06 / 100	99,01 / 99
R^2	85%	88%	84%

The results show that the GMDH approximation of unknown true Yandex ranking rules gives quite good fit to original for simulation in several very diverse areas.

Comparison the performance results of the search engine Yandex and built universal ranking model



Institution name	Average number of visited links		
	Before search spam sifting	After search spam sifting	
	Yandex ranking	Yandex ranking	Universal ranking model
Ukrainian Radio Engineering Institute	14	9	5
Mykolaiv National University	15	8	4

In these results, all user queries during two months were taken into account. So we can make an important practical conclusion: the most reputable search engines, attaching more importance to external ranking features, complicate the possibility of artificial enhancing a site popularity but it decreases the relevance of search results for a user.

Google ranking model built using GIA GMDH

$$y = 3{,}24 + 2{,}71\tilde{o}_{3} + 0{,}12\tilde{o}_{4} + 0{,}00003\tilde{o}_{7} - 2{,}69\tilde{o}_{12} + 0{,}012\tilde{o}_{22} - 14{,}8\tilde{o}_{27} - \tilde{o}_{28} -$$

$$- 27{,}29\tilde{o}_{35} + 4\tilde{o}_{40} - 0{,}006\tilde{o}_{41} - 7{,}89\tilde{o}_{5}\tilde{o}_{6} + 0{,}06\tilde{o}_{14}\tilde{o}_{15}^{2} + 0{,}002\tilde{o}_{37}\tilde{o}_{38}\tilde{o}_{39}$$

The most influencing in the Google ranking model are the following 16 factors:

***x*3** – the ratio of the total words number on the site to the keywords number on the site;

***x*4** – the ratio of the total words number on the page to the keywords number on the page;

***x*5** – Google PR;

***x*6** – subject popularity;

***x*7** – requests number of the particular keyword for a certain period of time;

***x*12** – website age;

***x*15** – recently updated pages;

***x*16** – number of pictures on website;

***x*22** – keywords font size;

***x*27** – keywords presence in title;

***x*28** – keywords presence in meta-tags;

***x*35** – match website keywords to the search engine page directory in which site is;

***x*37** – total number of links;

***x*38** – number of internal links;

***x*39** – number of external links;

***x*40** – website depth;

***x*41**– number of external links, containing keywords in its title;

Comment: The most influenced in the Google ranking are the external factors (x_5, x_6, x_7, x_{12}, x_{35}, x_{39}, h_{41}) as compared to internal ones.

Verification of Web resources ranking results

Results of sites ranking using model built with GIA GMDH

Place in google.com.ua	Values by GMDH	Rounded results
1	1,23	1
2	1,89	2
3	4,01	4
4	4,21	4
…	…	…
21	21,23	21
22	22,49	23
…	…	…
57	57,22	57
58	58,15	58
…	…	…
99	98,95	99
100	99,56	100

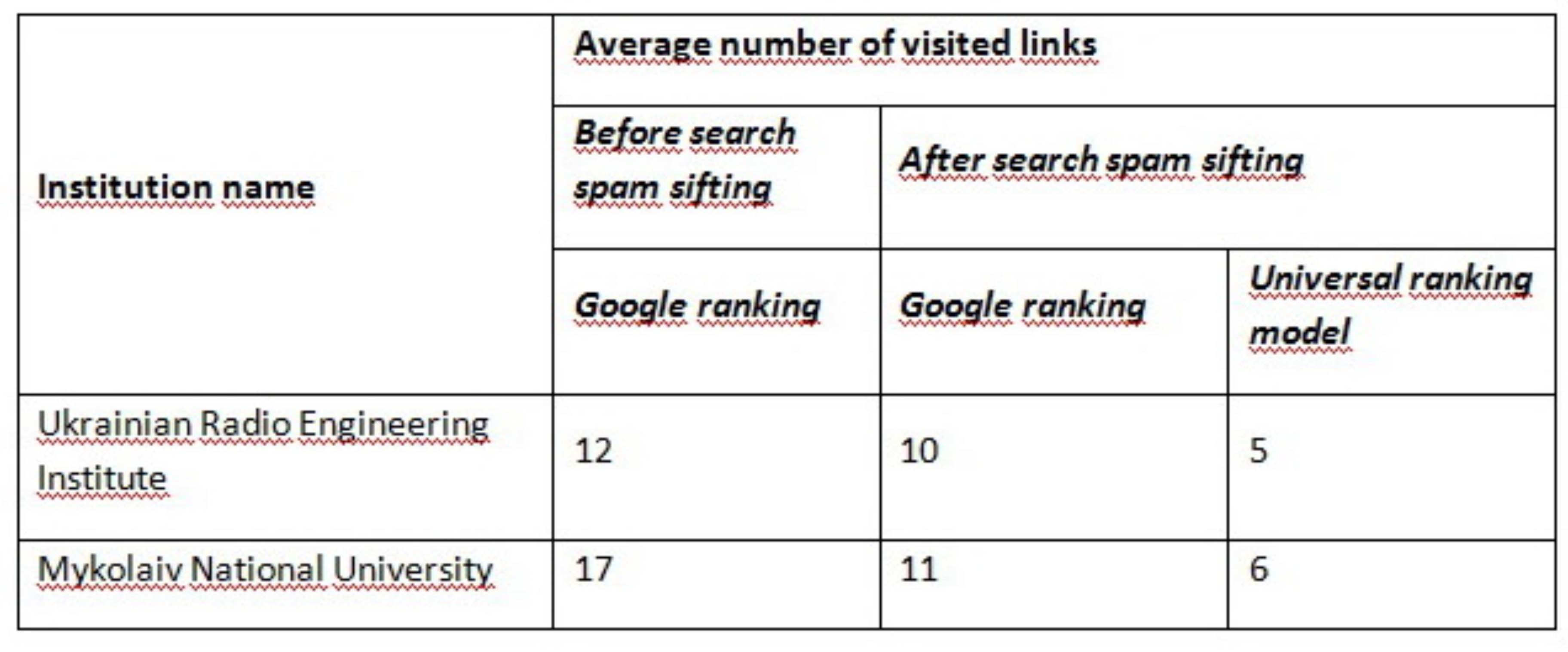
Web resources ranking results

Place in google.com.ua	Values by GMDH		
	«omelet recipe» / rounded result	«buy notebook Kiev» / rounded result	«expert systems» / rounded result
1	0,83 / 1	1,02 / 1	0,78 / 1
2	1,91 / 2	2,11 / 2	2,02 / 2
…	…	…	…
37	37,91 / 38	36,99 / 37	36,89 / 37
38	37,95 / 38	38,00 / 38	38,01 / 38
…	…	…	…
77	77,02 / 77	76,01 / 76	78,00 / 78
78	78,11 / 78	77,72 / 78	78,32 / 78
…	…	…	…
100	99,86 / 100	100,56 / 101	100,01 / 100
***R*²**	87%	95%	93%

We have verified the correctness of the constructed model (3) for other completely different search queries: «omelet recipe» «buy notebook Kiev» «expert systems».

Table 4 shows the successful results of comparing web resources ranking by Google and the model built using GIA GMDH

Comparison the performance results of the search engine Google and built universal ranking model



Institution name	Average number of visited links		
	Before search spam sifting	*After search spam sifting*	
	Google ranking	*Google ranking*	*Universal ranking model*
Ukrainian Radio Engineering Institute	12	10	5
Mykolaiv National University	17	11	6

Discussion

The aim of the paper was to demonstrate the possibility to build automated procedures to significantly improve the target informativity of returns of a present search engine. The developed technique uses the search results of Google or Yahoo as input data. The sifting of irrelevant commercial information is based on data obtained from public API and search results parse. The collected data are analyzed for the availability of pre-defined attributes of commercial sites to remove them from the delivery set.

To enhance the effectiveness of the proposed technique, the sites remaining after sifting are additionally ranked according to the model built using the generalized iterative algorithm GMDH. A procedure of experiments for constructing such a model is described in this article. This enables a comfortable opportunity for a user to find the needed target information on sites placed in the first few positions of the search delivery modified using the proposed technique. Besides that, the results are presented concerning the developed system application in two state institutions of Ukraine to demonstrate the technique efficiency.

An additional task is also solved in this paper on assessing the effectiveness of inductive GMDH algorithms for construction of search results ranking models. For this purpose, the experiments were carried out aimed to "discover" the unknown accurate ranking model for search results of Yandex and Google. Ranking models of the search engines are hidden for users, so the definition of site relevance evaluation parameters of these engines is of great scientific and applied interest. The simulation results showed that the GMDH algorithm successfully builds polynomial models approximating the unknown ranking rules of the popular search engines with great accuracy.

Conclusion

Inductive approach to building ranking models from a user's learning sets can significantly improve the quality of search compared to the ranking models of modern search engines in case of both the presence of search spam in SERP and without it.

Such kind of model for removing commercial information can be configured to churn any other category of information on pre-selected criteria including the sifting any information other than commercial.

The use of the generalized iterative algorithm GMDH makes it possible to build highly efficient ranking models to enhance relevance of search results by any search engine. This shows good prospects for further research in this direction.

Attaching more importance to external ranking features, main search engines complicate the possibility of artificial cheat of site popularity but it decreases the relevance level of search results.