Inductive building of search results ranking models to enhance the relevance of text information retrieval

Viacheslav Zosimov
applied mathematics and information computer technologies department
V.O.Sukhomlynsky Mykolaiv National University
Mykolaiv, Ukraine
zosimovvv@bk.ru

Oleksandra Bulgakova
applied mathematics and information computer technologies department
V.O.Sukhomlynsky Mykolaiv National University
Mykolaiv, Ukraine
sashabulgakova@list.ru

Volodymyr Stepashko
Department for ITs of Inductive Modeling
International Research and Training Centre for Information Technologies and Systems
Kyiv, Ukraine
stepashko@irtc.org.ua

Abstract — **The article describes a method for constructing a model for ranking the search engine delivery on the Internet using inductive GMDH algorithms. The method makes it possible to enhance substantially the relevance of scientific and technical information search on the Internet provided to sift spam and the commercial information. The process of discovering the web resources ranking model is described for known search engines and comparing its effectiveness with the constructed model.**

Keywords—information search; target information; search engine; search relevance; inductive modeling; GMDH.

I. Introduction

One of the main reasons for the low quality of scientific and technical information search on the Internet is presence in the SERP of artificially promoted sites and a search spam. Artificial wrapping of ranking parameters leads to enhancing irrelevant information in the search results. Therefore, search engines are continuously improving its search algorithms but there are still a lot of artificially promoted and commercial sites in the delivery.

For example, inserting the "Data Protection" query, a user will on the first positions of search results get links to online shops proposing security equipment and alarm systems, and web security agencies, but do not receive the required links to websites associated with the scientific and technical information on the subject.

It happens because the information is usually retrieved by keywords without regards to a domain and context. Due to that, commercial sites investing significant funds to occupy first positions for many keywords. Accordingly, if the keywords for target search of scientific and technical information and for the commerce sites promotion are the same, a user need to spend o lot of time and attention to view a large number of irrelevant links to irrelevant sites.

One of possible options to improve the efficiency of relevant information search is the solution of two consecutive tasks: sifting irrelevant information in the web and ranking search results using a model constructed based on a given learning set. An approach of this kind was proposed firstly in [1] and is evolving here. To build such model, we have chosen the Generalized Iterative Algorithm of the Group Method of Data Handling (GIA GMDH) [2] among several modeling methods, which is based on an inductive way of model construction from a given data set.

II. Sifting the commercial information

In the phase of sifting commercial information, a classification model based on the selected set of attributes is represented as a chain of decision-making rules. For this task a DNF-classifier is used where for the category C "commercial information" in course of research there were predefined n characteristic attributes $\{a_1^C,\dots,a_n^C\}$ and m site structural elements $\{b_1^C,\dots,b_m^C\}$ with these attributes.

The classifier is built like that: 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 information [3].

The list of structural elements analyzing to detect characteristic attributes contains: meta tags, paths to Java-scripts and stylesheet decoration; title, meta description, keywords; texts on the home page; navigation elements.

III. THE IDEA OF BUILDING A RANKING MODEL

To perform a good ranking of sites remaining after sifting irrelevant commercial ones, it is necessary to identify the features affecting the ranking and their weights.

In this case, algorithms of known search engines cannot be used for ranking of search results in view of the presence some volume of web spam, as they are configured on ranking of sites with large spam amounts among them. In the developed search engines ranking algorithms, external features (number of external links, domain age, PR value, etc.) have considerable weights, as it is harder to counterfeit them than internal ones. In addition, accordingly, fewer weights are imputed to internal parameters (presence of key words in a title, number of keywords on a page, etc.).

Typically, search engine algorithms rank the results well considering the presence of spam but in our case spam is mostly sifted out and therefore the ranking model should be accordingly adjusted. The process of building a new ranking model is described below based on expert opinions.

IV. 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 1967.

GMDH as a self-organizing data mining tool is based on main principles of automatic generation of inductively complicated variants, non-final decisions and successive selection of models of optimum complexity with respect to minimum of 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 [2] 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).

V. BUILDING RANKING MODELS FOR SPECIFIC FIELDS

In the construction process of ranking models, 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 analyzes 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) on a selected sample request and available after sifting the commercial spam information.

Group members accounted their ranking lists of sites sorted by the relevance of the target information.

For each expert list, a ranking model was built based on the given features using the GIA GMDH [2].

The table I shows the results of an experiment.

Example. Building ranking models from the learning set formed by experts in the field of psychology.

Knowledge Area: psychology. Number of experts: 7.

Search query: "Neuro-Linguistic Programming."

TABLE I. 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

Table II contains the complete lists of knowledge areas and search queries participated in modeling experiments.

TABLE II. DESCRIPTION OF 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

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

VI. BUILDING THE UNIVERSAL RANKING MODEL

To ensure effective ranking of search results from any field of knowledge, one may build either a specific ranking model for each field of knowledge or a universal model.

To build the universal model, the results of previous experiments were used. Table III represents data on the frequency of ranking features in previously constructed models. To create an effective model, one needs to sift out uninformative features and estimate the weights (coefficients) of the informative features.

TABLE III. THE FREQUENCY OF FEATURES 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 discarded.

Informative features:

x_1 – keywords number on site;
x_2 – keywords number on page;
x_3 – the ratio of total words number on site to the keywords number on site;
x_4 – the ratio of total words number on page to the keywords number on page
x_{14} – frequency of updating site information;
x_{15} – the last update;
x_{18} – the presence of alt-tags for images;
x_{23} – keywords font weight;
x_{26} – How far from the web page head are keywords;
x_{27} – presence of keywords in the title;
x_{28} – presence of keywords in the meta-tags;
x_{41}– number of external links with keywords in the title;

Uninformative features:

x_7 – number of requests for a particular keyword in a given period of time;
x_8 – total number of web pages;
x_{10} – site volume;
x_{12} – site age;
x_{16} – number of images on site;
x_{22} – keywords font size;
x_{35} – matching of site keywords to the search engine directory partition in which the site is;
x_{37} – total number of links;
x_{40} – site depth;

They can be ignored as they are spurious and have fewer weights when ranging.

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:

$$y = 4{,}67 - 0{,}01x_1 + 0{,}02x_2 + 0{,}72x_3x_4 - 1{,}97x_{18}x_{23} + \\ + 2{,}01x_{22} + 0{,}0001x_{26} - 1{,}1x_{41} + 4{,}07x_{27}x_{28} - 3{,}15x_{14}x_{15}^2 \qquad (1)$$

Accuracy of the model is 88.6%, slightly lower as compared to any of the built models based on the learning sample. However, universality of its application to any field of knowledge compensates the loss of accuracy, and the constructed model we use in what follows. If a user wants to build its own ranking model based on personal preferences with higher-ranking accuracy, he could use the implemented technique for automatic construction of ranking model based on a relevant learning sample.

VII. BUILDING AND ANALYSIS OF RANKING MODEL FOR THE YANDEX SEARCH ENGINE

An experiment demonstrating the procedure of the ranking model building for popular in Ukraine search engine Yandex is described below. To do that we activated Yandex ranking process of web resources for the search request "heat exchange" and have built the ranking model.

For this experiment we selected the first 50 sites from Yandex SERP. The data set $W = (X\ y)$ contained 42 variables-features that numerically characterize each of 50 sites. Matrix columns and lines of X correspond to the variable values and to the Web resources respectively. Output variable y corresponds to the position of a web resource among the ranking results.

The matrix was divided into two parts: 2/3 as the training set A used for estimation of model parameters of various generated models and the other 1/3 as the testing sample B to calculate the model quality evaluated as the accuracy on the testing sample.

The following model of the ranking process of web resources in the search engine was built using GIA GMDH:

$$y = 7{,}12 + 1{,}01x_3 + 0{,}12x_4 + 0{,}000001x_7 - 2{,}69x_{12} + \\ + 8{,}12x_{22} + 2{,}79x_{27} + 0{,}001x_{28} - 48{,}19x_{35} - 2{,}001x_{41} - \qquad (2) \\ - 12{,}22x_{42}x_6 - 3{,}08x_{14}x_{15}^2 + 0{,}04x_{37}x_{38}x_{39}$$

Determination coefficient of the model: R^2 =89%.

TABLE IV. THE RESULTS OF WEB RESOURCES RANKING BY YANDEX.UA

Position in yandex.ua	GMDH model output	Rounded results
1	0,83	1
2	1,29	2
3	3,08	3
4	5,01	5
5	5,23	5
…	…	…
21	21,23	21
22	21,99	22
23	23,85	24
…	…	…
57	57,02	57
58	58,11	58
…	…	…
99	99,95	100
100	107,12	107

Analyzing (2) one can conclude that the following 13 features have main influence on web resources ranking in the search engine Yandex: the ratio x_3 of total words number on site to the keywords number on site; the ratio x_4 of total words number on page to the keywords number on page; topic's popularity x_6; number of requests x_7 for a particular keyword in a given period of time; site age x_{12}; frequency x_{14} of updating site information; the last update x_{15}; keywords font size x_{22}; the presence of keywords in title x_{27}; the presence of keywords in meta-tags x_{28}; matching the site keywords to the search engine directory partition in which the site is present x_{35}; number x_{41} of external links with keywords in title; Yandex citation index x_{42}.

Hence Yandex ranking model relies mainly on external factors (x_6, x_7, x_{12}, x_{35}, x_{41}, x_{42}).

To verify the correctness of the results for Yandex ranking model (2), we used it for checking its effectiveness for other search queries, including: «Probability theory»; «Carpet cleaning»; «Vacation in Thailand».

Table V shows that the model built using GIA GMDH tracks the Yandex results of web resources ranking with high accuracy. Hence it can be used to further investigation of various ranking methods.

TABLE V. THE RESULTS OF WEB RESOURCES RANKING BY YANDEX

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
$\boldsymbol{R^2}$	85%	88%	84%

Comparison of the built universal web-resources ranking model (1) with Yandex ranking model (2) shows that the latter comprises mostly internal ranking features as compared to the external ones. It shows also that the weight of matching internal features in our model is bigger whereas the weight of external ones is lesser.

These models were used in two state agencies: Ukrainian Radio Engineering Institute (UREI) and Mykolaiv National University (MNU) for testing of the developed software system in tasks of building models for ranking search results. From the results listed in table VI it follows that the model (2) is more efficient. 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 cheat of site popularity but it decreases the relevance of search results.

Efficiency of ranking results was evaluated by the average number of user browsing of links found on request. This method of assessing the effectiveness of ranking is appropriate but not the only possible one. Obviously one cannot clearly define needs of a user by his/her search query and user's criteria for selection of a particular site are not formalized, so we may only propose a list of sites that are most relevant to the entered query. If user visited first five links found at his request, it does not mean that he has found the asked information at the fifth site. He or she could find the needed information on the second site and then visit other sites just to make sure that nothing better is there.

Intuitively, the less number of links found on request the user browses, the less time he or she spends to search for the relevant information. On this basis, it can be argued that the less is the number of visited links sufficient for finding relevant information the better would be the ranking.

Table VI shows the number of links visited by users using different ranking models based on SERP.

TABLE VI. 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***
UREI	14	9	5
MNU	15	8	4

VIII. RANKING MODEL FOR GOOGLE

It is of interest to compare the Yandex ranking model (2) with the model describing the Google ranking process of web resources constructed in [1] using the same technique based on the generalized GMDH algorithm:

$$\begin{aligned} y &= 3{,}24 + 2{,}71x_3 + 0{,}12x_4 + 0{,}00003x_7 - 2{,}69x_{12} + \\ &+ 0{,}012x_{22} - 14{,}8x_{27} - x_{28} - 27{,}29x_{35} + 4x_{40} - \\ &- 0{,}006x_{41} - 7{,}89x_5x_6 + 0{,}06x_{14}x_{15}^2 + 0{,}002x_{37}x_{38}x_{39} \end{aligned} \quad (3)$$

Determination coefficient of the model: R^2 =92%.

Analysis of (3) shows that the main influence on the Google ranking model has the following 16 factors:

x_3 – the ratio of total words number on site to the keywords number on site;

x_4 – the ratio of total words number on page to the keywords number on page;

x_5 – Google Page Rank;

x_6 – topic's popularity;

x_7 – number of requests for a particular keyword in a given period of time;

x_{12} – age of site;

x_{14} – frequency of updating site information;

x_{15} – the last update;

x_{22} – keywords font size;

x_{27} – the presence of keywords in title;

x_{28} – the presence of keywords in meta-tags;

x_{35} – matching of site keywords to the search engine directory partition in which the site is present;

x_{37} – total number of links;

x_{38} – number of internal links;

x_{39} – number of external links;

x_{40} – site depth;

x_{41} – number of external links with keywords in title.

After analyzing these factors, one can say that the external factors (x_5, x_6, x_7, x_{12}, x_{35}, x_{39}, x_{41}) has main influence on Google ranking as compared to internal ones.

The correctness of the universal model (4) was verified for several completely different search queries including: «omelet recipe»; «buy notebook Kiev»; «expert systems».

Table VII shows the results of sites ranking by (4).

TABLE VII. THE RESULTS OF WEB RESOURCES RANKING BY GOOGLE

Position in google.com.ua	**GMDH model outputs/ranks**		
	«Omelet recipe» / rounded	***«buy notebook Kiev» / rounded***	***«expert systems» / rounded***
1	0,83 / 1	1,02 / 1	0,78 / 1
2	1,91 / 2	2,11 / 2	2,02 / 2
3	3,09 / 3	3,56 / 4	3,01 / 3
…	…	…	…
37	37,91 / 38	36,99 / 37	36,89 / 37
38	37,95 / 38	38,00 / 38	38,01 / 38
39	38,23 / 38	38,78 / 39	39,05 / 39
…	…	…	…
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%

Like that for Yandex (table VI), experiments were performed (table VIII) regarding the engine Google [1] and results were closely related: to get the needed information, user visited less number of links.

TABLE VIII. 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***
UREI	12	10	5
MNU	17	11	6

IX. 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 is 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.

X. 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.

REFERENCES

[1] V. Zosimov, V. Stepashko, O. Bulgakova, “Enhanced technology of efficient Internet retrieval for relevant information using inductive processing of search results”. – Artificial Intelligence Methods and Techniques for Business and Engineering Applications / G.Setlak, M.Alexandrov, K.Markov (Eds.). – Rzeszow, Poland; Sofia, Bulgaria: ITHEA, 2012. – 345 p. / – P. 99-112.

[2] V. Stepashko, O. Bulgakova, “Generalized Iterative Algorithm GIA GMDH” // Proceedings of the 4th International Conference on Inductive Modelling ICIM-2013, 16-20 September 2013, Kyiv, Ukraine. – Kyiv: IRTC ITS NASU, 2013. – P. 119-123. See http://www.mgua.irtc.org.ua/attach/ICIM-IWIM/2013/2.6%20.pdf

[3] V. Zosimov, “Construction and Application of a Model for Sifting of Irrelevant Sources when Retrieving the Scientific and Technical Information on the Internet” – Cybernetics and computer engineering: Intern. scientific journal. – 2013. – №171. – P. 52-68. (In Russian)

[4] A.G. Ivakhnenko, “Polynomial theory of complex systems”, IEEE Trans. Sys., Man and Cyb., 1, No 4 (1971), pp. 364-378.