ENHANCED TECHNOLOGY OF EFFICIENT INTERNET RETRIEVAL FOR RELEVANT INFORMATION USING INDUCTIVE PROCESSING OF SEARCH RESULTS

Vyacheslav Zosimov, Volodymyr Stepashko, Oleksandra Bulgakova

Abstract: *The developed technology consists of three main stages: collection of information from a search engine; sifting irrelevant information by the pre-selected features; ranking the obtained results by relevance to a user's request. The ranking model is built with the usage of inductive1 GMDH algorithms. The article describes the effectiveness investigation of the developed technology improving the search relevance of target information on the Internet compared with the Google search engines. When studying, three experiments were conducted with one search request chosen for each experiment. The search results for every request obtained from Google were subsequently processed with the developed technology.*

The first 100 sites from Google SERP were analyzed to compare the relevance level of Google search and that provided with our technology. Outcomes of the experiments are given in the form of circle diagrams showing the percentage of different types of sites in the search results before and after processing it using the proposed technology. The research demonstrates higher effectiveness of the proposed technology compared to Google search: the developed technology allows achieving the search relevance at the level of 80%. Application of this technology will enable more convenient and relevant search of target information on the Internet.

Keywords: *Information search, target information, search engine, search relevance, inductive modeling.*

ACM Classification Keywords: *H.3.5 Information Search and Retrieval - Information Filtering; H.3.5 Online Information Services – Web-based Services*

Introduction

Active artificial promotion of commercial websites to obtain new customers led to the fact that results of a search engine for majority of requests contain large amount of irrelevant

information at the first positions. Search engines algorithms are constantly improving to deal with the artificial promotion of web resources but despite this the search for relevant information remains to be an increasingly difficult task.

For clarity let us distinguish two classes of all the information located on the Internet: the business as well as scientific and technical ones. Below under the commercial information we mean a promotional one provided on a site to attract new customers, visitors, subscribers, etc. to get a commercial gain as a result.

One of possibilities for increasing the efficiency of relevant information search is separation of the whole information on the Internet into commercial as well as scientific and technical according to some predefined attributes. So it is about a solution of two consecutive tasks: sifting irrelevant information in the web and ranking search results using the model built from a specified training sample.

To construct ranking models, a generalized iterative algorithm of the group method of data handling (GIA GMDH) was chosen as an effective inductive modeling tool among a number of methods and algorithms.

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: information collecting, sifting of commercial information, ranking the results.

Phase 1. Information collecting. It is not necessary to develop an individual search robot to collect and index data from Web pages because today's search engines cope successfully with the task of data collection and its subsequent indexing. So it is appropriate to use information from the Google search engine database being the most popular and providing opportunity to connect directly to its database using Google API Interface. Google database provides not only a list of sites relevant to the request entered

by a user but also a number of ranking attributes of these sites which will be needed at the ranking phase.

Sites list obtained from the Google database is stored and transmitted for processing to the next phase in which the commercial information is sifting out based on characteristic attributes. Marks received from the Google database with a list of sites are not used on the sifting phase and stored for the ranking phase.

Phase 2. Sifting the commercial information. In the phase of commercial information sifting, a classification model based on the selected set of attributes is represented as a set of decision making rules. For sifting commercial information the DNF-classifier is used in which for a category *C* "commercial information" in course of research there was predefined a number of characteristic attributes $\{a_1^C,\ldots,a_n^C\}$ (where *n* is the total number of attributes) and a set of the site structural elements $\{b_1^C,\ldots,b_m^C\}$ (where *m* is the total number of the elements) containing these attributes.

The classifier is built like such an example:

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

The list of the structural elements that are analyzed for the presence of characteristic attributes:

- meta tags, paths to Java-scripts and stylesheet decoration;
- title, meta description, keywords;
- text on the home page;
- navigation elements.

Phase 3. Results ranking. To implement qualitative ranking of sites that remain after the filtering stage, it is necessary to define the weights of ranking attributes obtained from the Google database during data collection phase and get a new ranking model based on these weights.

We do not use the Google rankings data because its algorithms are set up to rank sites taking into account the presence among them a large amount of search spam. Considering this, in Google's algorithm a lot of weights have external attributes (number of external links, domain age, web pages authority, etc.) because it is harder seemingly to forge them than internal attributes. And accordingly minor weights are set to internal attributes (the presence of keywords in the title, number of keywords per page, etc.). Google's algorithm ranks well search results with a lot of search spam but its ranking model should be adjusted. In the developed technology, the search results ranking is implemented according to a new ranking model built using GIA GMDH.

Investigation of the technology effectiveness

The research consists of three experiments. For each experiment was chosen one search request. Selected request was processed at first with the search engine Google and then handled with the developed technology. To simplify of the experiments description, only first 100 sites from Google SERP were analyzed. In what follows we compare the performance of Google search and the developed technology.

Experiment 1. Investigation of the technology effectiveness for the "Information Security" request for search engine google.com.ua.

The tested search query: "Information Security".

Fig.1 shows the percentage of different types of sites among the top 100 results found for this request.

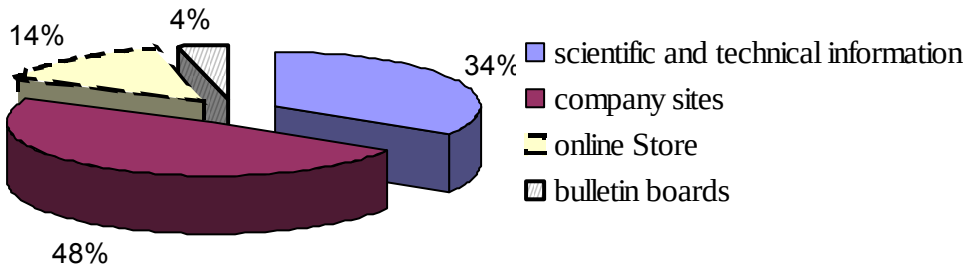

Fig. 1 Percentages of different types of sites from google.com.ua SERP for the "Information Security" request

Fig. 1 demonstrates that the target scientific information forms only 34% of the total number of sites found. The remaining 66% were mostly sites of companies that provide services for the information protection and some online stores selling products to protect information.

Then we processed the same query using our technology. Fig.2 illustrates the obtained results.

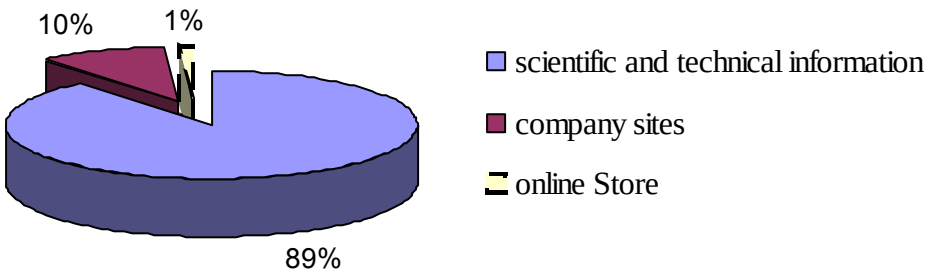

Fig. 2 The percentage of different types of sites from google.com.ua SERP for the "Information Security" request after additional processing

According to Fig. 2, the percentage of scientific information have raised to 85%. This is almost three times better than Google results and updated search results include only 15% of commercial sites. At the same time any site containing the scientific information was not wrongly ignored. This means that the performance of the search delivery was not reduced.

Experiment 2. Investigation of the technology effectiveness for the "Programming 1C" request for search engine google.com.ua

The tested search query: “Programming 1C“.

Fig.3 shows the percentage of different types of sites among the top 100 results found for this request.

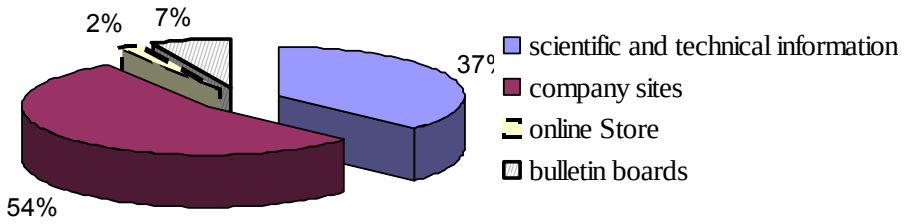

Fig. 3 The percentage of different types of sites from google.com.ua SERP for the ”Programming 1C” request

Fig. 3 shows that the requested scientific information keeps only 37% of the total number of sites found. The remaining 63% were mostly sites of companies that provide services for the 1C programming, bulletin board and some online stores selling products of 1C Company.

Fig.4 illustrates the percentage of various sites categories in the remaining results after processing the same query using our technology.

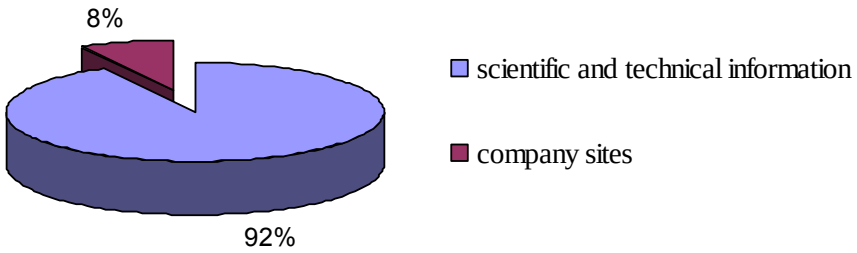

Fig.4 The percentage of different types of sites from google.com.ua SERP for the ”Programming 1C” request after processing

Figure 4 shows that the share of scientific information grew to 92%. This is almost three times as much as Google results, includes only 8% of commercial sites and any site containing scientific information was not wrongly aborted. Hence the performance of the search output was not reduced.

*Experiment 3. Investigation of the technology effectiveness for the request "BOBCAT engine structure" (*bobcat is an American manufacturer of forklifts*) for search engine google.com.ua*

The tested search query: “BOBCAT engine structure “.

Fig.5 shows the percentage of different types of sites among the top 100 results found for this request.

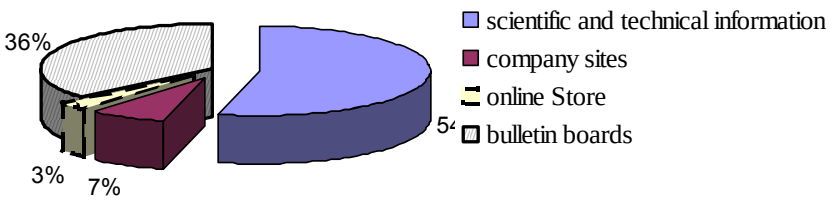

Fig. 5 The percentage of different types of sites from google.com.ua SERP for the ” BOBCAT engine structure” request

As it is evident from Figure 5, the searched scientific information makes up only 54% of the total number of sites found. The remaining 46% were mostly bulletin board, sites of companies that provide services for BOBCAT engines, and some online stores selling engines.

Than the same query was processed using our technology and Fig.6 illustrates the share of various sites categories among the obtained results: the percentage of the scientific information increased to 83% and only the rest 17% of commercial sites. This is more than half much again as Google results and any site containing scientific information was not wrongly excluded, so the performance of search message was not reduced.

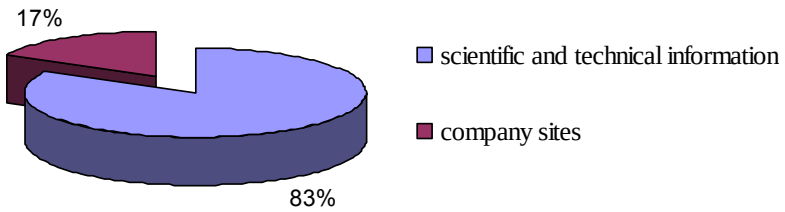

Fig. 6 The percentage of different types of sites from www.google.com.ua SERP for the ”BOBCAT engine structure” request after processing

Table 1 shows the results of application of the developed software system for solving the problem of increasing the efficiency of scientific and technical information search in two state organizations: Ukrainian Radio Engineering Institute (UREI), and Mykolaiv National University (MNU).

Ranking performance of search results was estimated as an average number of links found for a user's request. This evaluation method of ranking is appropriate though not the only possible.

Table 1. Application results of the developed software system

Organization	Search accuracy performance in %		Percentage of not sifted commercial sites	Percentage of wrongly sifted relevant sites
	Search engine	Developed software system		
UREI	83%	97%	3%	0,4%
MNU	54%	85%	8%	0,5%

Table 2 shows the results of comparing the ranking performance by the Google search engine model and that built by GIA GMDH model.

Table 2. Ranking performance results

Average number of visited links		
Without sifting the irrelevant information	After sifting the irrelevant information	
Google ranking model	Google ranking model	Using GIA GMDH ranking model
12	10	5
17	11	6

Building of Google search engine ranking model

To check the quality of bult using GIA GMDH ranking models we decided to rebuilt Google ranking model. For this we simulated (have found the model), Google ranking process of web resources for the search request "web programming".

Input variables:

For the experiment, we selected the first 50 sites from Google search engine result page (SERP).

Model quality was evaluated on a sample *B* as the value of the regularity criterion AR.

$$AR = \left\| y_B - X_B \hat{\theta}_A \right\|^2 \quad (1)$$

The matrix X contains 42 variables-factors that numerically characterize each site and divided into two parts: 2/3 - study A, which is used for coefficients estimation, the other 1/3 - test sample B. Matrix columns are corresponding to the values of the factors, and lines - are corresponding to the Web resource.

To simulate the ranking process of web resources search results, the following attributes were used:

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_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_8 – total number of web pages;
x_9 – amount of text on site;
x_{10} – amount of site;
x_{11} – amount of web page text;
x_{12} – age of site;
x_{13} – the keyword presence in URL of the site (domain name);
x_{14} – frequency of updating site information;
x_{15} – the last update;
x_{16} – number of images on site;
x_{17} – number of multimedia files;
x_{18} – the presence of alt-tags for images;
x_{19} – alt-tags length (in symbols);
x_{20} – usage of frames;
x_{21} – site language (Russian or foreign);
x_{22} – keywords font size;
x_{23} – keywords font weight;

x_{24} – the distance between keywords (in symbols);
x_{25} – written in capital letters or not the keywords;
x_{26} – How far from the beginning of the web page are keywords;
x_{27} – the presence of keywords in title;
x_{28} – the presence of keywords in meta-tags;
x_{29} – the presence of file «robot.txt»;
x_{30} – site's location;
x_{31} – comments in source code;
x_{32} – what type of pages, each page relates: html or asp;
x_{33} – the presence of flash files;
x_{34} – the presence of the same pages on site;
x_{35} – matching of site keywords to the search engine directory partition in which the site is;
x_{36} – the presence of "noise words" ("stop words");
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;
x_{42} – Yandex citation index.

Output variable: ***y*** – web resource position among the ranking results.

Model accuracy was calculated by the determination coefficient formula:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} 100\%, \qquad (2)$$

where $\bar{y}$ is an average value, $\hat{y}_i$ is the model output.

Using the generalized GMDH algorithm, the following model was constructed that describes the search engine ranking process of web resources:

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

$AR(A) = 2{,}48$; $AR(B) = 3{,}51$, $R^2 = 92\%$

Table 3 shows the results of sites ranking using model (3).

Table 3. 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
5	4,89	5
6	6,02	6
7	6,78	7
8	8,00	8
9	8,52	9
10	9,33	9
…	…	…
21	21,23	21
22	22,49	23
23	22,85	23
…	…	…
32	33,56	34
33	33,56	34
34	33,68	34
…	…	…
57	57,22	57
58	58,15	58
…	…	…
99	98,95	99
100	99,56	100

Analysis of the built model shows that 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 main influence on Google ranking has the external factors (x_5, x_6, x_7, x_{12}, x_{35}, x_{39}, h_{41}) as compared to internal ones.

We verify correctness of constructed model (3) for other search queries:

- «omelet recipe»
- «buy notebook Kiev»
- «expert systems»

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

Table 4. 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
3	3,09 / 3	3,56 / 4	3,01 / 3
…	…	…	…
15	14,89 / 15	15,08 / 15	14,98 / 15
16	16,02 / 16	16,06 / 16	14,99 / 15
17	16,78 / 17	17,21 / 17	14,99 / 15
…	…	…	…

Table 4 continued. 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
21	19,52 / 20	21,11 / 21	21,03 / 21
22	21,33 / 21	22,13 / 22	21,89 / 22
23	23,01 / 23	24,05 / 24	22,99 / 23
…	…	…	…
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
…	…	…	…
62	63,06 / 63	62,12 / 62	62,13 / 62
63	63,56 / 64	63,42 / 64	62,58 / 63
64	64,18 / 64	64,01 / 64	64,02 / 64
…	…	…	…
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%

Table 4 shows that the model built using GIA GMDH accurately follows the Google ranking of web resources and can be used to further investigation of ranking methods.

Conclusion

Results of the experiments described in this article shows that the developed technology allows achieving high precision of search results containing more than 80% of relevant scientific information. This is due to using the results of Google search engine and applying inductive GMDH algorithm. All this indicates that the usage of this technology will

allow making search for scientific and technical information on the Internet more convenient, simple and precise.

Application of the developed software system for solving applied problems improves greatly the accuracy of relevant information search necessary for the research.

The usage of the inductive algorithm for building ranking models is efficient. Models built using GIA GMDH helps greatly reduce the time needed for search of relevant information.

High accuracy of the constructed Google ranking model proves the effectiveness of a generalized iterative algorithm GMDH for solving such kind of problems.

Bibliography

[Stepashko] Stepashko V., Bulgakova O., Zosimov V. Performance of Hybrid Multilayered GMDH Algorithm. – Proceedings of the III International Workshop on Inductive Modelling IWIM-2011, 5-9 July 2011, Kyiv-Zhukyn, Ukraine. – Kyiv: IRTC ITS NANU, pp 109-113, 2011.

Authors' Information

Viacheslav Zosimov *– Lecturer of Mykolaiv V.O. Suhomlynsky National UniversityI, Ukraine; e-mail: zosimovvv@bk.ru*

Major Fields of Scientific Research: Web-technologies, Information Search and Retrieval.

Volodymyr Stepashko *– Head of Department for Information Technologies of Inductive Modeling of IRTC ITS, Professor, Dr Sci, P.A.: 40, Akademik Glushkov Prospect, Kyiv, Ukraine, 03680; e-mail: stepashko@irtc.org.ua*

Main Fields of Scientific Research: Data Analysis Methods and Systems, Knowledge Discovery, Information Technologies of Inductive Modelling, Group Method of Data Handling (GMDH)

Oleksandra Bulgakova *– PhD, Associate Professor of Mykolaiv V.O. Suhomlynsky National University, Ukraine; e-mail: sashabulgakova@list.ru*

Major Fields of Scientific Research: Information Technologies of Inductive Modeling.