Theoretical Fundamentals of Search Engine Optimization Based on Machine Learning

Michael Godlevsky [1], Sergey Orekhov[1], Elena Orekhova [1]

[1] National Technical University “Kharkiv Polytechnic Institute”,
Kharkiv, Ukraine
(god_asu, osv)@kpi.kharkov.ua

Abstract. The theoretical basis of search engine optimization (SEO) process, metrics of its efficiency and the algorithm of performance were proposed. It is based on the principles of situation control, machine learning, semantic net building, data mining and service oriented architecture as IT solution. The main idea of the scientific work is the use of situation control as a learning method for search engine to recognize WEB site content in the Internet. In this case built semantic net of WEB site content is a learning sample to teach search engine. To receive the keyword list (semantic kernel) of web content the modified algorithm of semantic net building was proposed. Developed service oriented IT solution includes PrestaShop CMS, 1C Enterprise component and WEB services done by Google and Yandex. An efficiency of the approach was proved by successful performance of 25 real SEO projects in 2012-2016 for the companies in Ukraine.

Keywords. Semantic net, Search engine, Search engine optimization, Keyword list, Semantic kernel, Machine learning, Web service, Situation control, Data mining.

Key Terms. MachineIntelligence, DecisionSupport.

1 Introduction: Problem and Research Goals

The results of performed SEO projects for Ukrainian companies in 2012-2016 showed some stable trends. Firstly, a lot of companies were trying to build their corporative WEB site or Internet shop. Secondly, an enterprise structurally is being transformed. At the moment the sale process is being built around Internet communications and interactions. The enterprises actively apply the new paradigms such as «virtual office» (WEB 2.0), «semantic search» (WEB 3.0) and «intelligent agents» (WEB 4.0) [1-2]. Thirdly, more and more users work with WEB content via mobile devices. Thus, they create different textual and numeric information streams such as blogs and social network accounts [2]. Fourthly, these streams (news and opinions) remain the sources of marketing information. This information could be applied for advertising or product promotion activity according to classical marketing theory named “4P” [3]. But marketing theory states if a trade mark comes first to the mind of a potential customer it has maximum success. Therefore at the moment if WEB content is the first in search engine answers then a company has success in the market. In the Internet such task is named search engine optimization problem for a WEB site.

At the moment there potentially exist three ways of solving the task. First way means that we manually or automatically put maximum amount of external hyperlinks on our WEB site in the Internet. The more external links on the WEB pages of our site we have the closer we are to the first position on a list of search engine answers according to a well-known method named Page Rank [4]. To automate the process of external link generation we use a software or WEB service to create a set of some hyperlinks (announcements) on web pages of predefined web sites. The disadvantage of this approach is that this generation usually is stopped by CAPTCHA function. We should input manually a secret code for each link. It is impossible to generate unique hyperlink descriptions including a predefined keyword list as well because a search engine has anti-spam and anti-plagiarism sub-systems to prune its search database. Moreover, a search engine uses a recommendation system to range WEB sites and user’s requests. Therefore, all web sites with a high range sell hyperlinks to our WEB site.

The second way is based on using the web services of contextually targeted advertising (pay per click advertising) such as Google AdWords [5] or Yandex Direct. These services provide automatic announcement of the product definition on different web sites by means of payment.

The third way is a promotion in a social network. This technology could be realized for free. In this case we create an internal web page of our product in a social network. As a result we build a copy of our web site but on the platform of social network. The set of new web pages might be a competitor to an existing web site therefore we should input a unique description on both web sites. It is an advantage that the social network forms a new network of personal accounts (web pages of products) and a user can search the product inside at network based on opinions of other users. The second advantage is that the hyperlink from social network has a high range in a search engine.

But the methods mentioned above can not help us to reach the first position on the list of answers during predefined time period. The company can not spend long time to do it. This time is restricted and critical. Usually this time period is not more than three months [6]. According to marketing theory one quarter is the conventional period for a product promotion.

At the moment, there are a lot of practices and good examples of SEO projects but no stable theory was presented. All practices describe how we should change the HTML content or how we can use Web services like Google AdWords. But we need the theoretical basis – the fundamentals of SEO [7]. The main reason of this situation is that the program code of a search engine is secret one. We can work with it as with a “black box”. There is some theoretical description similar to a vector space model to understand how search engine forms the answers.

Thus, at the moment we don’t have a stable theory of search engine optimization but only series of practices how to implement it.

2 Problem Statement

Let us summarize the requirements and restrictions mentioned above as follows:

- the time for the performance of search engine optimization process is restricted by one quarter;
- at the moment a search engine is represented as a set of intelligent agents (Internet bots);
- Web site content is a set of keywords and files, merged semantically. The amount of keywords and their combinations are unlimited;
- Web content is not stable. It is being changed. Thus in order to reach the first position in the list of search engine answers we should change the web content step by step. In other words we correct the set of keywords the same way we transform the site from one state (situation) to another.

The web content is unstable because the users in the Internet while searching the product or service apply different keywords constructions as a search engine request. Moreover, these constructions might change in time. Therefore, it is necessary to set up the web content continuously. Also the competition in the market changes Internet environment. The competitors correct their web sites as well.

In the paper the task of search engine optimization is formulated as a problem of machine learning. Let us consider a search engine as an intelligent machine [8]. We ask it by a set of predefined keywords and it answers us via a set of predefined keywords as well. We believe that these sets (request and answer) are based on the annotations of web content of different web sites. Thus, in order to teach a search engine we send to it the web content of predefined web sites (external hyperlinks with a high rank) and our web site. All of them should have the web content including prepared earlier semantic net (semantic kernel or keyword list).

Thus, the way to solve this task is to build correct semantic net of web content of our web site. Actually this approach is well-known among SEO specialists.

Based on classic learning theory such task can be solved if we have the following elements:

- a knowledge representation model,
- learning samples, and
- method of learning.

Let us consider in the capacity of the model of knowledge representation – a semantic net [9]. Therefore, we believe that inside a search engine the semantic net is formed to describe predefined combination of question and answer. This combination leads us to the list of web sites, where our site is placed on the first position.

Thus, as learning samples we can apply the semantic net (keyword list) of our web site(s), but as a learning method – a situation control approach [10]:

$$S_i; Q_j \underset{U_k}{\Rightarrow} Q_l,\ i=\overline{1,n},\ j,l=\overline{1,m},\ k=\overline{1,z}. \tag{1}$$

where S_i - full situation, which characterizes one iteration of learning process;

Q_j and Q_l - current situations, which describe initial and final keyword lists during one iteration;

U_k - logic-transform rule, which defines the changes of keyword list Q_j in order to receive Q_l.

Let us determine a current situation (semantic kernel) in the capacity of a formal model as follows:

$$Q_j =< K, R, A, R_S >, \tag{2}$$

where K - a set of keywords of web site content;

R - syntactic rules to form a core (keyword list);

A - a system of axioms of a core;

R_S - semantic rules to form a core.

In our case logic-transform rule U_k determines the actions of addition and elimination of some keywords from current situation Q_j that leads to the situation Q_l.

Therefore, the task of search engine optimization can be formulated as follows: to teach a search engine for minimal amount of full situations $S_i\ i=\overline{1,n}$ which define the transformations of keyword list Q_j of predefined web site. In other words, the semantic net of a search engine should be changed in a such way that it included the web site keyword list within the minimal amount of iterations.

The solution of formulated task might be compared with a problem of a locked door with a numeric code. We search a number combination to open a lock.

Moreover, our solution should be automated using modern Internet technologies.

3 Task Solution

The following actions were proposed to solve the task in the paper:

1. To propose the algorithm of automated forming keyword list (semantic kernel), which is the annotation of our site web content.
2. To create a learning algorithm based on situation control.
3. To implement the stopping criterion of learning algorithm (the cycle of situation control) which guarantees the web site to be on the first position or on the first page of an answer list. The criterion describes the situation in which a semantic kernel of web site becomes a part of semantic net of a search engine. The criterion

has economic background as well. In case with the Internet shop the stopping criterion describes the situation when the site generates real orders from a user in predefined period of time.

4. To develop software which performs the learning cycle and checks the stopping criterion. The software should be integrated into web services of a search engine.

On the first stage the modified algorithm of semantic net building is proposed. The algorithm is based on Relevance Feedback (Data Mining methodology) [11]:

Step 1. Delete all the words that help to describe main reason of a text. These words are prepositions, pronouns, punctuation. This is about 20% of all words in the text.

Step 2. Build the vocabulary such as Table 1.

Table 1. The vocabulary of semantic net

Keyword	Type	Frequency of occurrence in a text	Amount of links with other words

We put to the vocabulary all the keywords after pruning procedure on the step 1. There are three types of keywords: object, action and notion. As an object we consider all keywords which describe physical issues, things and phenomena. An action is infinitive form. All keywords that we have not classified as objects or actions are the notions. All selected keywords which have the frequency of occurrence more than two might be a part of our semantic net (semantic kernel). But the keywords with frequency more than five, but less than 7 or 10 have the highest priority.

Step 3. Build a semantic net, where the vertexes are the keywords from the vocabulary. But an arc between two nodes has three classes: «isa», «part of» и «kind of». The link «isa» symbolizes a rule «IF-THEN». It takes place between an object and an action or between two actions or two objects. The link «part of» is possible, when one term is a part of another. It occurs between two objects or actions only. The arc «kind of» describes the link between keywords “object” and “notion” only.

Step 4. Calculate the amount of link «isa». This quantity describes the amount of potential requests (questions) to our semantic net. This amount shows how many questions a user will ask a search engine concerning out web site.

When we have a semantic kernel (keyword list) of web site then we have two possibilities. Firstly, we can compare our kernel with the kernel inside the search engine by typing a request in its query box. It allows check the position of the site in an answer list. Next it is possible to organize learning session as follows:

Step 1. Receive a semantic kernel of our web site (to define current situation Q_j). Get the answer of a search engine (annotation) Q_l. Thus, we form the full situation S_i at the moment i. Moreover, a lexical graphic rule shows us how to form the annotation Q_l from Q_j. In other words, the rule describes the principles of annotation in a semantic net.

It is understood that both components Q_l and Q_j can be modified. The first one is influenced by the content manager, and the second - under the influence of learning process of a search engine, which is run together with the implementation of this algorithm. Therefore, it is necessary to change the WEB hypertexts resource in a way that Q_l becomes the annotation of Q_j.

Step 2. Change the semantic kernel of the web resource based on the principle of semantic net annotation.

Step 3. Check the stop criterion of the algorithm, if it doesn't satisfy the conditions, go to Step 1.

In the capacity of stop criterion of the algorithm it is possible to select criteria Moranda, that introduces a limitation by the number of iterations i. If i is greater than a predetermined value, then we stop the algorithm. This criterion is appropriate to be applied to sites that are not Internet shops.

An alternative measure of stopping is a new indicator named an efficiency of online store, as proposed in this paper:

$$P = \frac{O}{V}, \tag{3}$$

where O - a quantity of orders per day, processed by Internet shop;

V - an amount of visitors of web site per day;

Practical application of this algorithm has shown that if index P is in the range of 0.05-0.1, it is possible to stop the learning algorithm.

4 IT Solution

The paper proposed to implement a learning algorithm of search engine based on standard web services of various search engines such as Yandex Metrica, Yandex Wordstat, Bing KeyTool, Google AdWords. The software architecture is shown on Figure 1.

The present software implements the algorithm as follows. Let us assume that we carry out SEO project of an Internet shop. In our case the shop was developed on Prestashop platform (CMS) using PHP programming language.

This is usually a four-level system. For the realization of the algorithm the architecture was complemented by the fifth level - WEB services (Fig. 1). Moreover, CMS functionality is extended by two components: 1C:Enterprise component to determine the current situation and Prestashop plugin to communicate with the web counter and evaluate the value of criterion (3).

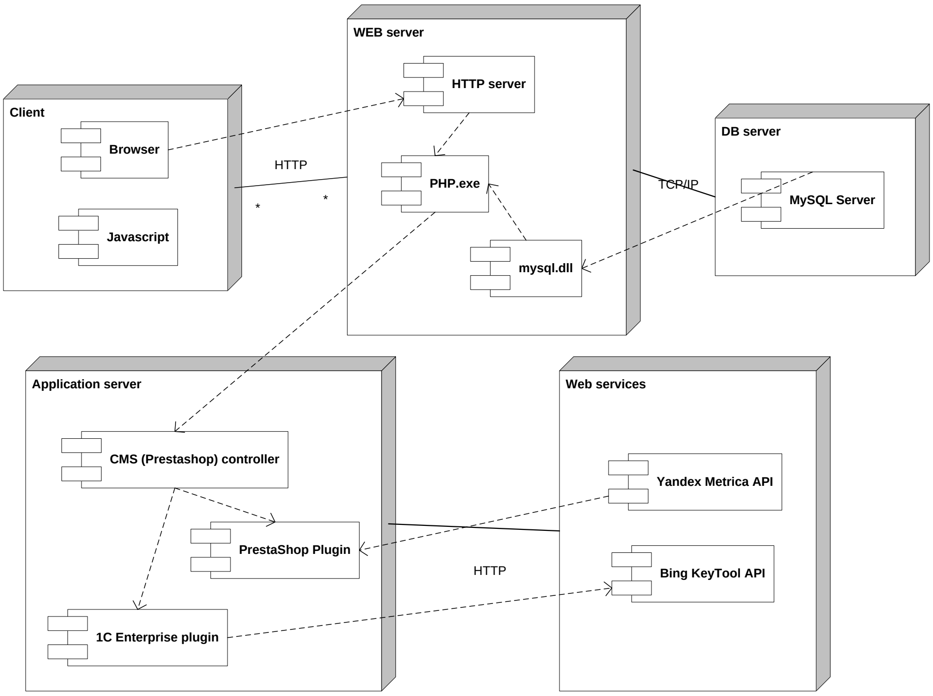
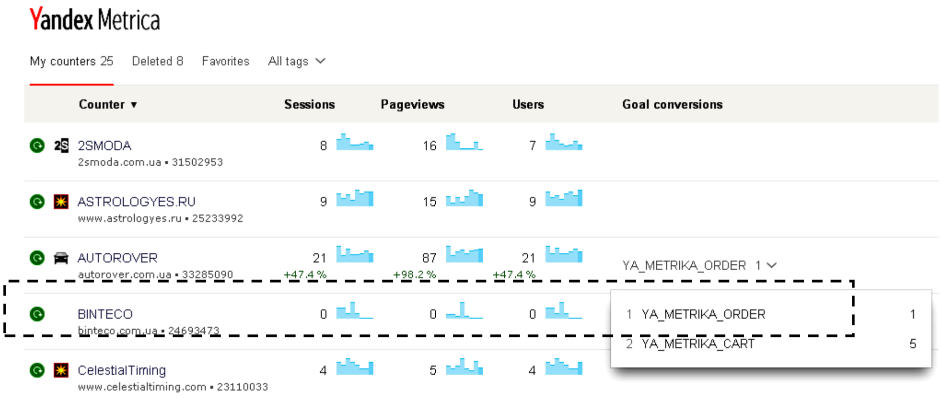


Fig. 1. Software architecture

Thus, we apply our software as follows. At first, using the 1C:Enterprise component we test the semantic net of a search engine and implement the algorithm of semantic kernel building for our Internet shop. Actually we build a semantic kernel of the shop product page. Then 1C component compares the semantic kernel and annotations in a search engine. It allows to change hypertexts on the product page. Having checked the results of changes in position of the site in a search engine list we can use software component (Prestashop plugin) that communicates with a web service Yandex Metrica. On the basis of their interaction, it is possible to obtain values of the criterion (3) - Fig. 2.

All presented screenshots contain the actual data of Internet shop of automobile accessories. This SEO project was based on proposed algorithms and its successful results are confirmed by values of Mail.ru web counter in 2016 - Fig. 3.



Yandex Metrica

My counters 25 Deleted 8 Favorites All tags

Counter ▾	Sessions	Pageviews	Users	Goal conversions
2SMODA 2smoda.com.ua • 31502953	8	16	7	
ASTROLOGYES.RU www.astrologyes.ru • 25233992	9	15	9	
AUTOROVER autorover.com.ua • 33285090	21 +47.4 %	87 +98.2 %	21 +47.4 %	YA_METRIKA_ORDER 1
BINTECO binteco.com.ua • 24693473	0	0	0	
CelestialTiming www.celestialtiming.com • 23110033	4	5	4	

1	YA_METRIKA_ORDER	1
2	YA_METRIKA_CART	5

Fig. 2. The visits of Internet shop (the data from Yandex Metrica display)

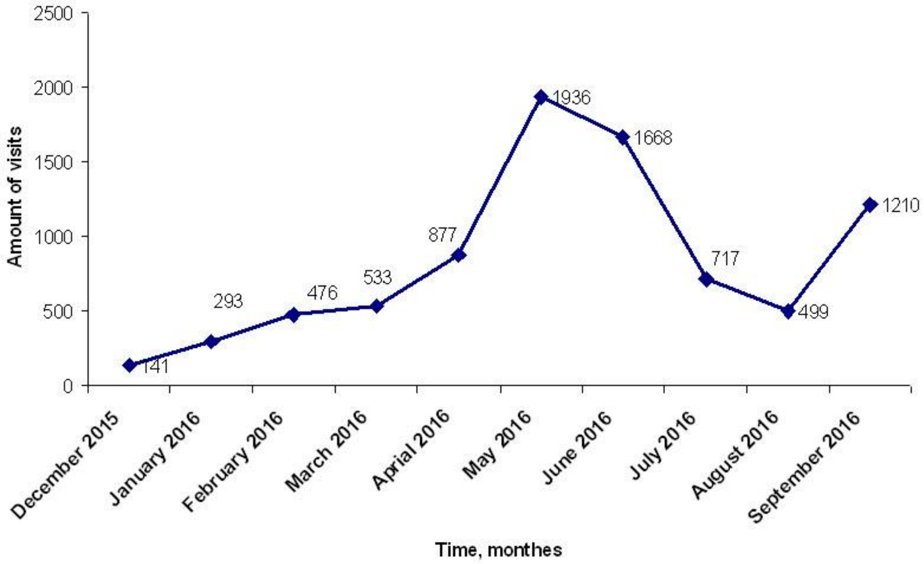


Fig. 3. Mail.ru statistics in 2016

5 Results

Theoretical fundamentals (the problem statement based on machine learning and situation control) and software (the algorithms, architecture and screenshots) are presented in the paper. They were developed in 2012-2016 based on the set of SEO projects done in Ukraine. At least 25 projects were performed to prove the presented results. Moreover, some projects were performed in a manual mode, but some of them were implemented in automated way using the software.

The figures 2-3 demonstrate the real result of SEO project done in 2016 for Ukrainian Internet shop of automobile accessories. In the figure 2 we see the screenshot with initial data to calculate the value of criterion (3). But the data of figure 3 show the amount of visits during the learning process proposed above.

The project was started in March 2016. We performed at least nine iterations of learning algorithm. The medium duration of the iteration was 15 days. During the iteration we have changed the semantic kernel of a web site (product page and home page). The initial amount of keywords was 300 items. Finally it was 17 items. We have placed different hyperlinks on the product page about auto sit covering. Thus, the visitors selected these pages more than others.

As a result, we have prepared the set of announces of the product in Google AdWords. There the keywords from semantic kernel were located. As we see (fig. 3) the amount of visitors grew twice during three months. Thus, the goal of SEO project was reached completely.

6 Summary

Performed scientific work allowed us to get some interesting results:

1. The real time that we spend to learn a search engine is approximately 75 days. But the duration of learning iteration is two weeks for Google and one week for Yandex. But in some projects Yandex learning continued more than Google search engine. Google renews its semantic net for 15 days, but Yandex does it faster – 10 days.
2. The real value of criterion (3) is equal to 0.01. This value is individual for each web site. But this is simple and effective index to evaluate the benefit from SEO project as a whole or to define the goals for SEO specialists.
3. We found a new source of marketing data – web counter statistics. These data allow us to change the classical process of marketing research and rebuild it in automated mode. As a result, we can create marketing information system (MIS) which works automatically. In a new architecture of MIS we propose to include CRM, CMS, and Web services.

Thus, the approach was tested both manually and automatically using different WEB services such as Bing KeyTool, Yandex Metrica, Yandex Wordstat, Google Analytics, Mail.ru WEB Counters and Google AdWords. The scientific novelty of this method lies in the fact that for the first time a metric is proposed that combines data with both the WEB counter (the amount of visits) and the internal information of the enterprise (the number of orders). Due to this, it is possible to establish a criterion for stopping the SEO process itself and estimate its effectiveness. It is revealed that the value of this metric (the stopping criterion) is strictly individual for each WEB project performed earlier.

References

1. Sareh Aghaei, Mohammad Ali Nematbakhsh and Hadi Khosravi Farsani. EVOLUTION OF THE WORLD WIDE WEB: FROM WEB 1.0 TO WEB 4.0: International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.1, (2012)
2. Nupur Choudhury. World Wide Web and Its Journey from Web 1.0 to Web 4.0: (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , (2014)
3. Philip Kotler. Marketing Management. Prentice Hall, New Jersey (2012)
4. Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Stanford University, Stanford (2000)
5. AdWords Study Guide. Naper Solutions Inc. USA. (2012)
6. Sergey Orekhov, Igor Cherenkov. The approach to retrieval events in news stream. East Europe Journal of Forward Technologies, Vol. 1/4 (61), Kharkov (2013)
7. Search Engine Optimization Starter Guide. Google Inc. (2010)
8. Konar Amit. Artificial Intelligence and Soft Computing. Behavioral and Cognitive Modeling of the Human Brain. USA: CRC Press LLC, (2000)
9. Sowa, John F. Knowledge Representation: Logical, Philosophical, and Computational Foundations. – Brooks Cole Publishing Co., Pacific Grove, CA, (2000)
10. Pospelov, D.A.: Situation Control Presentation. Cybernetics and Situational Control Workshop, Columbus, Ohio, 20 – 24 March., (1995)
11. Hastie T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction / Hastie T., Tibshirani R., Friedman J. – Springer, (2011)