DEVELOPMENT OF A WEB SERVICE FOR FORMING A SEMANTIC KERNEL OF A WEBSITE BASED ON DATA MINING METHODS

Dmitry Mironenko
PhD, Associate Professor[1]
mironenko_ds@ukr.net

Elena Balalayeva
PhD, Associate Professor[1]
balalaevaeu@gmail.com

[1]*Department of Information Technologies*
Pryazovskyi State Technical University (PSTU)
7 Universytets'ka str., Mariupol, Ukraine, 87555

Abstract: A technique has been developed for the formation of the semantic core of a site for Internet resources with already generated or dynamically generated content. A mathematical model based on data mining methods is given. For analysis and further research, an information technology is developed – a web service. The main users of this web service will be SEO specialists, for whom it will become a convenient tool. This web service relies on data mining methods and statistics on the use of search queries obtained from the Google Search interface. Integration of the data allows to qualitatively select the necessary keywords and give a list of the most optimal, relating to the subject of the site.

For research, the site of the Department of Computer Science of the Azov State Technical University was selected. During the experiment, a list of keywords and phrases was obtained. The words in the list are sorted in decreasing order of performance. The automated formation of the semantic core eliminated the subjectivity of the SEO specialist when selecting words and phrases, the time spent on its formation is ten times less than the manual semantic analysis. The result set included only those words and phrases that are often used in the content and are most significant. Using Google Search to adjust the list of words allows to match words with search queries and select those for which users are looking for information. The web service has a flexible mechanism for regulating (limiting) the number of keywords in the result set.

The main advantage of using this service is that only those words and phrases fall into the semantic core, in response to which there is something to offer visitors on the site.

Keywords: data mining, semantic site core, keywords, search engine optimization, web-service, site-content, automation, information technology.

1. Introduction

It is difficult to imagine doing business in modern conditions without using the virtual representation of the company on the Internet [1]. Moreover, the content of the content of the web resource is no longer a simple static text that describes a product, service or contacts with the company. Content filling the site is the provision of information that would be able to make the user think and behave in a direction beneficial to the company [2]. But before the content begins its motivating actions, it is necessary to bring a potential client to this resource. One of the predominant solutions is the transition to the resource from search engines (Google, Yandex and others), it is known that the share of "search traffic" of any site (the ratio of the number of visitors coming from search results to the total traffic to the site) is prevailing [3–5]. Therefore, today, the urgent task is the development of site content taking into account search engine optimization (SEO – search engine optimization) [6]. One of the key stages of SEO is the development of the semantic core of the site, which, as a rule, is performed manually by specialists and is time-consuming [3].

The semantic core includes keywords and phrases that users request in search engines at the stage of searching for the necessary web resource. In fact, the semantic (semantic) core of the site is the basis of internal site optimization, namely, meta-descriptions, texts, heading tags and accent tags [7].

Providence of SEO optimization for an already created web resource or site with dynamic content is of particular difficulty [3, 8]. Manual methods of semantic core formation are inferior to automated methods for a number of reasons. The first of these is that more time is needed for the core to be updated by SEO specialists. The second reason is a subjective approach in the choice of words by each specialist. And the most important third reason is the difficulty of making decisions on accounting for user preferences and actions [8, 9].

The aim of research is in increasing the efficiency of the methodology for the formation of the semantic core of the site through the use of data mining algorithms.

To achieve this aim it is necessary to solve the following objectives:

– to develop a method for the formation of the semantic core of the site, which performs the ranking and filtering of "extra" keywords based on data mining;

– to develop information technology, presented in the form of a web service, conduct research on a real-life site.

A web service should implement: parsing the pages of an existing site, form a set of keywords and phrases. The resulting set of keywords, by ranking and filtering out extra words, will become the basis of the semantic core.

The use of a web service will reduce the time for data processing and eliminate the problem of subjective influence of SEO-specialists [8].

2. Methods

Let's represent the semantic core of the site in the form of five [10]:

$$SK=\{K\},\{P\},\{W_1\},\{W_2\},\{W_3\}, \quad (1)$$

where $\{K\}$ – set of words and phrases (keywords), $\{W_1\}$, $\{W_2\}$, $\{W_3\}$ – set of integral weights of words and phrases (weights) according to the relevant criteria for repeating words in the text of the content, $\{P\}$ – set of priorities.

The set of keywords $\{K\}$ includes the following subsets: $K_1 \subset K$ – nouns from the site header (<title> tag) and meta tags (<description>, <keywords>); $K_2 \subset K$ – nouns from the headings of the main content (tags <H1>,.., <H6>); $K_3 \subset K$ – nouns from the main content (tags ,).

Since each element of the set is unique, and there cannot be two identical elements in the set, when it is added to the original subset of the next element, it will be checked for uniqueness.

Each element k of the resulting set $\{K\}$ is assigned a priority P_i according to (2).

$$P_i = \begin{cases} 1, & \text{IF} \quad k_i \in K_1, \\ 2, & \text{IF} \quad k_i \in K_2, \\ 3, & \text{IF} \quad k_i \in K_3, \end{cases} \quad \forall i \in [1, n], \quad (2)$$

where n – the power of the set K (the number of keywords and phrases); k_i – another element of the set; $K_1 ... K_3$ – subsets of the set K.

As practice has shown, the most important html tags related to ranking pages on search engines are the <title>, <description>, <keywords>, <h1> tags. Accordingly, they will be assigned a high priority. Following them are the tags <h2>, ... <h6> – medium priority. The and tags that highlight keywords in the text of our document are low priority.

The set of weights for keywords and phrases {*W*} includes the following subsets:

– $w_i^1 \in W_1$ – the relative weight characterizing the occurrence of the query words in the document, which is determined by the formula (3):

$$w_i^1 = \begin{cases} \dfrac{count(K_i)}{n}, \forall\ K_i \in Z, \\ 0, \forall\ K_i \notin Z, \end{cases} \quad \forall i \in [1,n], \tag{3}$$

where n – the power of the set K (the number of keywords and phrases), Z – the set of words in the query string; K_i – keyword;

– $w_i^2 \in W_2$ – relative weight characterizing the occurrence of pairs of query words in a document, which is determined by the formula (4):

$$w_i^2 = \begin{cases} \dfrac{count(K_i \cup K_j)}{n}, \forall\ (K_i \in D) \cap (K_j \in D), \\ 0, \forall (K_i \notin D) \cup (K_j \notin D), \end{cases}$$

$$i \in [1,n],\ \ j \in [1,n], \tag{4}$$

where n – the power of the set K (the number of keywords and phrases) K_i, K_j – keywords; D – set of words in the document;

– $w_i^3 \in W_3$ – the relative weight characterizing the occurrence of the entire query text, which is determined by the formula (5):

$$T = \bigcup_{i=1}^{n} Z_i,$$

$$w_i^3 = \begin{cases} \dfrac{count(T)}{n}, \forall\ T \in D, \\ 0, \forall T \notin D; \end{cases} \tag{5}$$

where n – the power of the set K (the number of keywords and phrases), Z_i – the user's query word; D – set of words in the document, T – the text of the user's entire request.

Based on the obtained relative weights, let's determine the integral weight indicator I_i by the formula (6):

$$I_i = P_i \cdot (w_i^1 + w_i^2 + w_i^3). \tag{6}$$

Based on the calculations, the list of keywords and phrases is supplemented by an integral weight indicator I_i, thanks to which it is possible to sort our list in descending order of this indicator according to formula (7):

$$f(K_i) \geq f(K_j), \quad f: \begin{cases} I_i \geq I_j, \\ i, j = 1, 2 \ldots n, \\ i < j. \end{cases} \tag{7}$$

Using the Google Search service API, for each word or phrase from the list, let's determine F_i^A – the base frequency of the request, U_i^A – the frequency of the request for the exact phrase.

Based on the data obtained, let's determine the relative frequency R_i^A by the formula (8):

$$R_i^A = \frac{(F_i^A - U_i^A)}{(F_i^A + U_i^A)/2} \cdot 100\,\%, \quad \forall i \in [1,n]. \tag{8}$$

After the relative frequency has been determined, it is necessary to calculate the efficiency coefficient E_i by the formula (9):

$$E_i = K_i \cdot R_i^A, \quad \forall i \in [1,n]. \tag{9}$$

Words, which performance ratio $E_i < 1$, are removed from the list. This is because when $E_i < 1$ – unsuitable key phrases, $1 \leq E_i \leq 10$ – optimal key phrases with the presence of traffic, $10 \leq E_i \leq 100$ – excellent key phrases that allow to receive a significant share of traffic, $E_i > 100$ – key phrases of the highest category with mega-portions of traffic and a large number of audience [1].

As a result, let's obtain a list of keywords and phrases forming the semantic core of the site.

3. Results

Work with the web service begins with a screen (**Fig. 1**), in which it is necessary to specify the http-link to the site for which let's form the semantic core.

Fig. 1. Web service start screen

After indicating the link to the website, a set of keywords and the frequency of their appearance on the website will be generated automatically (**Fig. 2**).

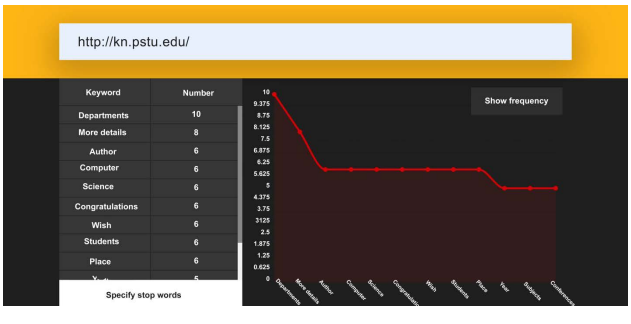
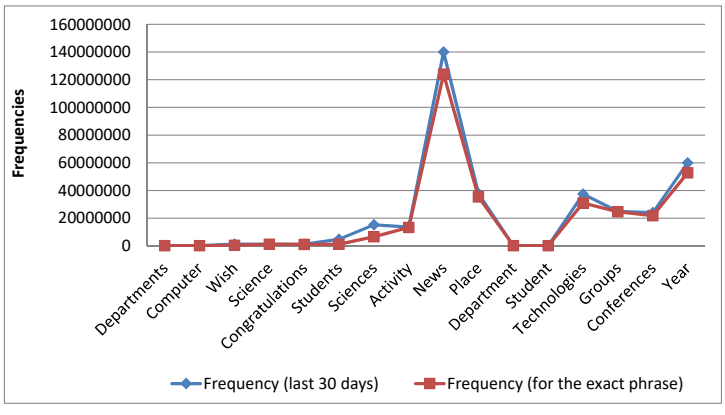


Fig. 2. The stage of obtaining keywords

To exclude keywords that are not significant from the sample, the technique described above is used. Using the Google Search API, let's obtain the frequency of queries for each word over the past 30 days – F^A and for the exact phrase – U^A based on these data, the relative frequency R^A is determined (**Fig. 3**).

The next step is determination of the keywords performance ratio, during which let's obtain the exact value that will show whether to use the keyword or it will be removed from the list according to the restrictions (**Fig. 4**).

Keywords performance coefficient allows to rank the list of keywords by excluding words with a low coefficient $(E_i < 1)$ from it. Thus, at the output there is a semantic core of the site.

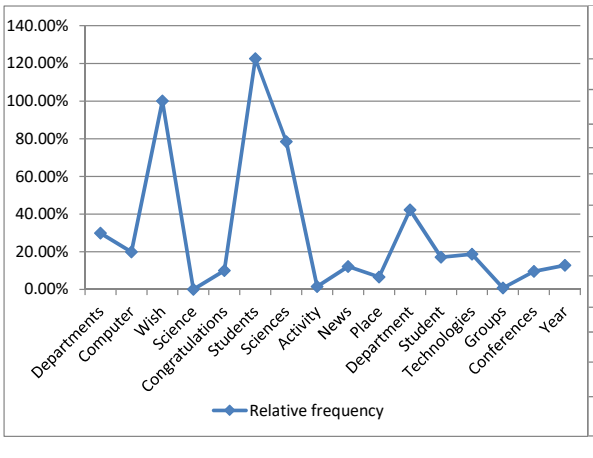


a

Frequency (last 30 days)		Frequency (for the exact phrase)		Relative frequency (keyword effectiveness%)	
Departments	169000	Departments	125000	Departments	29.93%
Computer	127000	Computer	104000	Computer	19.91%
Wish	1310000	Wish	436000	Wish	100.11%
Science	1170000	Science	1170000	Science	0.00%
Congratulation	1150000	Congratulation	1040000	Congratulation	10.05%
Students	4750000	Students	1140000	Students	122.58%
Sciences	15200000	Sciences	6630000	Sciences	78.52%
Activity	13500000	Activity	13300000	Activity	1.49%
News	140000000	News	124000000	News	12.12%
Place	37800000	Place	35400000	Place	6.56%
Department	136000	Department	88600	Department	42.21%
Student	197000	Student	166000	Student	17.08%

b *c*

Fig. 3. Data obtained as a result of a request in Google Search: *a* – frequency graph; *b* – table values of frequencies (for the last 30 days and according to the exact phrase); *c* – relative frequency (keyword efficiency, %)



Efficiency factor	
Departments	3.48
Computer	1.71
Wish	8.00
Science	0.00
Congratulation	0.80
Students	8.54
Sciences	4.68
Activity	0.06
News	0.72
Place	0.36
Department	2.52
Student	1.02

a *b*

Fig. 4. Keyword efficiency factor: *a* – relative frequency graph (keyword efficiency, %); *b* – table values of efficiency factor

4. Discussion and conclusions

The implementation of this technique allows to create a list of the most optimal keywords and phrases for a real-life site. This technique of constructing a semantic core significantly reduces the time of site optimization, and is also quite universal and can be applied by SEO specialists with a few improvements to effectively promote sites with dynamic content.

Using this web service, a list of keywords and phrases that make up the semantic core of the site is ordered in descending order of efficiency coefficient. But, along with this, it isn't recommended to use the entire list, since depending on the subject and volume of the site, the list may include 3–5 keywords, or maybe 500–1000. The choice and final decision is up to the SEO specialist.

The main advantage of using this service is that only those words and phrases fall into the semantic core in response to which there is something to offer visitors on the site.

References

1. How big is the SEO Industry on the Internet? Available at: http://www.bluecaribu.com/seo-industry
2. Egri, G., Bayrak, C. (2014). The Role of Search Engine Optimization on Keeping the User on the Site. Procedia Computer Science, 36, 335–342. doi: https://doi.org/10.1016/j.procs.2014.09.102
3. Ashmanov, I., Ivanov, A. (2011). Optimizatsiya i prodvizhenie saytov v poiskovyh sistemah. Sankt-Peterburg: Piter, 464.
4. Enzh, E. (2017). SEO. Iskusstvo raskrutki saytov. Sankt-Peterburg: BHV-Peterburg, 812.
5. Grohovskiy, L. (2011). SEO: rukovodstvo po vnutrennim faktoram. Moscow: TSentr issledovaniy i obrazovaniya. «TopEkspertRF», 133.
6. Search Engine Optimization (SEO) Starter Guide. Available at: https://support.google.com/
7. Sevost'yanov, I. O. (2010). Poiskovaya optimizatsiya. Prakticheskoe rukovodstvo po prodvizheniyu sayta v Internete. Sankt-Peterburg: Piter, 240.
8. Chung, D., Klünder, A. (2007). Suchmaschienen-Optimierung: Darschnell Einstieg. Heidelberg.
9. Maliy, V., Zolenko, M. (2017). SEO na eksport. Pervaya kniga po prodvizheniyu za rubezhom. Topodin/Ridero, 154.
10. Mironenko, D. S., Kunak, V. A. (2018). The model of building a semantic kernel. Nauka ta virobnictvo, 19, 179–184. Available at: http://eir.pstu.edu/bitstream/handle/123456789/20589/%d1%81.179-184.pdf?sequence=1

Received date 04.10.2019
Accepted date 11.11.2019
Published date 23.11.2019