Towards Question on Linguistic Approach to Search Engine Optimization: Clustering, Collocation, Grams

Oksana Akay[1*], *Anna* Kalashnikova[1], *Igor* Kalashnikov[2], and *Alina* Golubeva[1]

[1]Don State Technical University, Rostov-on-Don, Russia
[2]Rostov State Transport University, Rostov-on-Don, Russia

Abstract. The article is devoted to the problem of search engine optimization in the context of nowadays reality. Linguistics can become one of the main instruments to improve its marketing position of the web-product so the linguistic approach to the search engine optimization acts as the modern and effective mechanism. The task of search engine optimization is to achieve the most effective promotion of the site in the search engines while preserving the most possible naturalness of the text. Now the topic of the so-called over-optimized texts is becoming especially urgent. One of the ways of the search engine optimization of a site can be clustering as an automatic search and detection of semantically similar groups of files among a predetermined number of files. Also the method of collocations is an extremely topical linguistic method for solving problems of search optimization. Philological science also offers the developers a method close to the method of collocations - the gram-method. The linguistic approach to search engine optimization obviously allows us not only to increase the effectiveness of SEO in terms of the results obtained, but also to facilitate and accelerate the development processes and ultimately the business processes.

1. Introduction

The Internet has finally confirmed its position in a person's life, having recently occupied the equivalent of the off-line environment. From the viewpoint of modern linguistics, the study of various aspects of computer-mediated communication is a matter of paramount importance [1]. As a result of the emergence of the Internet as a new communication environment, the question about the emergence of a special type of communication arises. The visual-written format provided from the virtual space technology perspective is perceived by the sight. Internet communication, especially in social networks, is characterized by spontaneity and conversationality, and that can be proved by the illegitimacy of interpreting network communication as strictly written [2].

Various business areas continue to be actualized purely on the Internet and simply presented in the network, as a result of which the sphere of web marketing is actively reconstructing. This is due to the fact that the audience of the communicative field of the Internet is, in fact, virtually infinite. And now the issue of search optimization is becoming increasingly important.

In search engines, the indexing method is used to search for documents and sites, that is, drawing up an information portrayal of the document based on keywords, i.e. the selection of many features, in a simple case of key words (thematic elements, terms and sometimes the links between them) that reflect the main theme of the text.

Optimization is understood as a description (model) of the problematic area, in which this region preserves only those essential properties in the resulting representation that are necessary for a given practical problem. Search engine optimization (SEO) is a set of actions to promote a website or an account in social networks aimed at increasing the "visibility" of this site / account in the final search engine delivery list for various user requests [3].

Statistics shows that the user is unlikely to continue to view more than thirty positions of the results. Consequently, the site of the organization should be at the forefront of search engines, so that the company is recognized and, moreover, interested in the subject of its activities. Here a stereotype begins: the higher the position of the site, the most respected is this web resource.

By means of optimization measures, a place is achieved on the maximum possible search result line relative to the first line, as a result of which the site / account receives an increase in the number of visitors and, accordingly, the unique visits and in analyzing the efficiency of the search engine optimization service, taking into account the time of the site exit to the customer's specified position, the value of the visitor from the target audience is considered. And in the context of the SEO development, the new approaches are

* Corresponding author: oksanaakay@me.com

emerged to adequately meet the needs of the marketing Internet environment.

The mechanism of search services in calculating the relevance of a site / account on social networks takes into account such indicators as the citation index (a parameter based on the ratio of the number and authority of sites linked to the promoted resource) and keyword density - a specific algorithm that allows semantic analysis of the text. Semantics is a section of linguistics that examines the meaning of words and their interrelations. And the most relevant areas for the modern SEO development include the linguistic vector.

2. Results and Discussion

It was natural breakthrough when Google quietly rolled out the Hummingbird algorithm in 2013 and the semantic search came with it. The hummingbird was more than a simple algorithm update at its core; it was a fundamental shift in the way Google would deliver results to their users.

The hummingbird was the culmination of 15+ years of data and user analysis, as well as testing and tweaking in order to deliver a substantial search experience for Google. In order to deliver the right results as quickly as possible, Google created semantic search. At its core, the purpose of semantic search is to create a relational connection by delivering the contextualized content.

The main task of linguistics in the context of search engine optimization is the development of methods for working with the semantic field of the site. A semantic field is a collection of semantic units that have a fixed similarity in some semantic layers and are related by the specific semantic relationships.

The semantic analysis is a form of analysis, derived from linguistics. This analysis also plays an important role in the search engine optimization. A search engine can determine webpage content that best meets a search query with such an analysis.

For the significative layer, this similarity is treated as a link to some (the same) set of concepts, for the denotative layer - as a link to the same set of objects in the external world, for the expressive layer - as a link to the same set of conditions for speech communication, for the syntactic layer - as a link to the same set of syntactic relations between the speech segments parts [4]. Thus, there are semantic fields in each semantic layer. Integration into semantic fields and arch-units may be considered (for example, it is not dismembered for the significative-denotative units).

The semantic field is divided into the core and the periphery and in case of SEO, the semantic core of the site or social network account is a list of those keywords that will be involved in a search for a specific website / account [5]. Compiling a semantic field for a web product is a very time-consuming and labor-intensive process. Modern computer linguistics is working on the problem of its optimization.

Some search engines like Google apply complex metrics to the websites to determine their visibility [6]. The way the site is built, the keywords that are used, the dynamism of the web presence and the links to and from website are some of the elements that need to be correct.

Currently the selection of keywords (phrases) occurs automatically based on statistical procedures in some search engines [7]. In fact, all the words of the text are the key ones; the most significant of them are selected using the statistical procedure of attributing the keyword or the thematic weight expression. The document / site with this approach is associated with a numerical vector, reflecting the importance of using the term in each document. A similar vector is mapped to the query. The relevance of some document to the query is determined by the distance between the corresponding vectors: the closer the vectors, the more the document corresponds to the user's request. Such a method, based on the frequency of a specific word, ignores the fact that the text usually contains synonymous and anaphoric substitutions. To improve the efficiency of a document search in addition to a purely quantitative approach, the additional linguistic-oriented technologies are used [8].

One of the ways of the search engine optimization of a site can be clustered as the automatic search and detection of semantically similar groups of files among a predetermined number of files. Clustering allows us to cover the semantic aspect of the search as much as possible, rather than it is possible during machining, which, in turn, facilitates both obtaining conceptually important for the SEO clusters and determining the suitable ones for the specifically required semantics of landings, i.e. those pages that fall after the search request user implementation. This allows us to reduce losses in the potential non-attendance of the page due to the incorrectly prepared list of keywords.

Clustering requests or, in other words, grouping a semantic core, involves allocating keywords to groups, based on certain characteristics. Grouping is done on the basis of meaning and not on the required number of pages in a particular section of the site.

Let us consider the possible stages of this process. Clustering of the semantic core begins with the selection of the most important, competitive and high-frequent key phrases that are the basis for promoting the site as a whole («beauty salon in Moscow»). Further, the general category of the service or the goods, provided by the customer, is allocated. Here, the clustering of search queries involves placing the general name of services or groups of goods without specificity («hairdresser's services», «make-up artist services», etc.). Then a narrower group of inquiries is identified, corresponding to specific goods or services (so if it is a beauty salon site, inquiries are: «manicure», «haircut bob», «evening make-up», «face cleaning», «bio-lifting» etc.). Such vertical approach will make it possible to obtain the fullest possible semantic core.

As a separate nuance for the developer, one can note such elements as «blog» and «news» sections. Due to the limited marketing field of the static site and, accordingly, the inability to place all the necessary elements of the semantic core on the main page and in the description of the goods or services, the adequate solution is to create the dynamic sections. The blog or news can be updated quite often; the amount of the text available to the

developer is expanded, so the grouping of the semantic core involves the allocation of a group of information requests, through which pages with relevant articles are promoted [9].

Also by the clustering search queries, it becomes possible to avoid some keywords that are not quite relevant and meaningless. As a result, the resulting data are classified and attract the attention of a much larger audience, which is the goal of the SEO-activity.

There are two ways to implement search query clustering: it can be performed both manually or with the help of the specialized services.

Obviously, manual work is very difficult and takes a considerable amount of time, only basic software is used, such as Excel or Google Sheets. However, this approach has a serious advantage: with the proper level of care and adequate language skills, the developer gets the most qualitative result: in principle, no «junk» requests can enter the semantic core.

Clustering of the semantic core with the use of online services, in turn, is less time-consuming. There is a number of services with different levels of automation. So in the Keyword Assistant the core is compiled almost in manual mode, it is easier to group by levels because groups are created automatically from the marked keywords.

Fully automatic grouping of the semantic core is possible due to a fairly large number of web resources and applications, that are specialized in these processes [10]. There are both free and paid options. For the large projects with a multi-thousand core, the clustering of queries is performed using such systems as Topvizor and Rush Analytics. The operation on the semantic core with this software is very simple: you only need to load a list of key queries, then the program will analyze the ranking in search engines and perform a keyword grouping. Such a scheme has one serious drawback: with obvious time savings on the preparation of a powerful semantic core, there is no guarantee against getting there some inadequate keywords that can reduce the effectiveness of the site.

But in any case, with the help of the clustering mechanism, the time required for processing the semantic field, is reduced and, as a result, the developers get additional opportunities to solve the problem of the website promotion.

Also the method of collocations is an extremely topical linguistic method for solving problems of the search optimization. Collocation is a phrase, consisting of two or more words, that have signs of a syntactically and semantically integral unit, in which the choice of one of the components is realized in meaning, and the choice of the second one depends on the first one's choice. At present, the term «collocation» has found wide application in the corpus linguistics, and within the corpus framework this concept is reinterpreted or simplified in comparison with traditional linguistics.

A person thinks by collocations, they respectively participate in the compilation of search queries. With the help of search collocations, the user shows his interest in something and identifies himself as the target audience in the field of Internet marketing. For example, when composing a semantic field, the site can be collocated as a combination of «soft sofa» and a combination of «mild climate», and the nuances of the meaning of the word «soft». Depending on the main word, the collocation will change, and different collocations will be adequate for their use in various marketing spheres respectively.

The study of the thematic content massive bodies allows us to identify collocations and stable phrases, used in this content. The analysis of these phrases allows us to understand exactly, which questions in this topic are primarily of interest to the potential audience.

Thematic collocation is a tool that has become a part of computer-mediated marketing use on the wave of the devices personalization and the widest possible spread on Android and IOS platforms. The modern search engines are able to calculate the thematic collocations, not only in general for the thematic texts, but also for the thematic collocations from the content resources that are popular with different target audiences. And by taking into account behavioural factors, it is possible to determine, what collocations available in the texts are of interest to a particular social and age group.

Search engines, increasing the amount of information and improving the methods of its processing, collect information about the users, personalizing the requests results, taking into account the previous actions of the recipient. Also the volume of the thematic text corps is being increased, the sets of collocations relevant to a specific user are being specified. So the set of collocations, given out in the line of a potential request from different addressees, can be quite different. A set of collocations for a browser, downloaded to a desktop computer in a rural library and for a modern-model smartphone browser, authenticated with the user's personal account, might be different.

Moreover, if we compare the set of collocations for the same user, but required from the different devices, the differences will be revealed. For the experiment, two devices, belonging to the same user, have been chosen: a smartphone with an operating system IOS and a laptop based on the operating system Windows. Both devices are used by only one person, the user surfs the Internet using a smartphone, uses social networks and browses the mail; the laptop is a working tool for documents, various scientific texts and didactic materials.

The search engine Yandex.ru in the browser on the smartphone when asked «eared» gives the collocations: «eared nanny gel for washing children's underwear», «eared nanny washing powder», «eared hedgehog photo», «eared thief in law». The similar request on the laptop is issued by the following set of collocations: «eared hedgehog», «eared hedgehog photo», «eared lexical analysis».

Thus, the collocation method allows the developer to improve the efficiency in compiling both the semantic core and the periphery of the site / social network account field. Using the collocation method, a SEO-optimizer can combat the «spam» in the text, it is directly aimed at improving the quality of marketing texts, increasing their readability and a perceived user, since human language thinking is aimed at positive perception of the collocated elements.

Philological science offers the developers another method close to the method of collocations. This is a method of gram. Gram is a combination of two (bigram) or several (trigram, etc.) words, that are not related to each other, but are usually included in the semantic field of the marketing resource, promoted by the optimizer. In the fields of computational linguistics and probability, the n-gram is a contiguous sequence of n items from a given text or speech sample. The items can be the phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from the text or speech corpus. When the items are words, n-grams may also be called «shingles» [11].

So, for example, in the semantic field for the store site for the sale of upholstered furniture in Moscow the words «to buy», «Moscow», «sofa» are included. With the help of these words, a trigram is created «to buy a Moscow sofa», which will look illiterate in the text, but the search engine will react to it as a kind of a key, leading the user to the necessary page.

It is important to note, that despite the obvious usefulness of grams in increasing the effectiveness of participation in the search engines, the SEO-optimizer should not overload the text part of the site with grams. In case, when the site is overloaded with grams, the user, even related to the target audience of the project, is unlikely to give an active response since the person's consciousness is «tuned» to the perception of a coherent text rather than a meaningless set of words. In addition, an overabundance of grams in the text can lead to the so-called « text spam» and then search engines can resort to removing the site from indexing for the rules violation.

Some optimizers resort to the use of automatic synonyms; that is software, which allows selecting the synonyms automatically and making the collocations and grams. And if, in the case of grams this in a number of cases (not always though), and it gives a positive result, then in the case of collocations, the use of such techniques is hardly justified.

The task of the search engine optimization is to achieve the most effective promotion of the site in the search engines, while preserving the most possible naturalness of the text. Now the topic of the so-called over-optimized texts is becoming especially up to date. The over-optimized texts are the texts with a critically high content of keywords and queries in the form of grams.

To get the evaluation of the naturalness degree of the text from the point of view of linguistics is possible using the Zipf's law (law rank - frequency). The Zipf's law is an empirical law, formulated the mathematical statistics that refers to the fact, that many types of data, studied in the physical and social sciences, can be approximated with a Zipf's distribution, one of a family of related discrete power law probability distributions. Zipf's distribution is related to the zeta distribution, but is not identical [12].

For example, Zipf's law states, that in some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus, the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.: the rank-frequency distribution is an inverse relation. For example, in the Brown Corpus of American English text, the word «the» is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences (69971 out of slightly over 1 million). True to the Zipf's law, the second-place word «of» accounts for slightly over 3.5% of words (36411 occurrences), followed by «and» (28852). Only 135 vocabulary items are needed to be calculated for the half of the Brown Corpus [12].

So according to the Zipf's law, the empirical law of frequency distribution of the words in a natural language looks as follows: if all the words of the language (or simply long text) are ordered in a descending way, according to their frequency use, the frequency of the n-th word in such list will be approximately inversely proportional to its serial number n (the so-called word rank). For example, the second most used word occurs about twice less often as the first one, the third - three times less often than the first one, and so on [13]. That is, the frequency of the word decreases according to its place in the ordinal list in the natural text.

There are special services, allowing us to check the individual pages of the site or some text for compliance with the Zipf's law [14]. However, it is important to note that the law works correctly if it is applied only to the voluminous texts (starting from 5000 symbols without spaces).

3. Conclusion

The search engine optimization is a sphere of interdisciplinary interaction: it involves computer science, cybernetics, sociology and, of course, linguistics in their various aspects, such as computer linguistics, psycholinguistics, semantics. The main parameter of the search engine promotion is content that is highly unique and natural.

Semantic search is the future of the search engine technology. Knowing the semantic analysis can be beneficial for the SEOs in many areas. On the one hand, it helps to expand the meaning of any text with relevant terms and concepts. On the other hand, possible cooperation partners can be identified in the area of link building, whose projects might show a high degree of relevance to the required projects.

The semantic core of the site is, in fact, the foundation with the correct formation, with the help of which it is possible to achieve the visible results. Regardless of the method, chosen by the developer to compose the semantic core of the site, the main requirements for its design are constant: to establish the exact word form for the promotion of the query; to exclude the filler words (dummy words); to remove the keywords, that are intended solely for the wrapping.

When it comes to online marketing, clustering, collocating and gram-making are the key elements in helping one's company's brand to move across the cultural and linguistic barriers, as it is implemented in the well-rounded SEO strategy.

The linguistic approach to the search engine optimization obviously allows us not only to increase the effectiveness of the SEO in terms of the results obtained, but also to facilitate and accelerate the processes of the development and ultimately the business processes.

References

1. S.C. Herring, Relevance in computer-mediated conversation. *Pragmatics of Computer-Mediated Communication*, 245 (2013)
2. O. Akay, I. Kalashnikov, A. Kalashnikova, A. Golubeva, *Proceedings of the 7th International scientific and practical conference "Current issues of linguistics and didactics: the interdisciplinary approach in humanities" (CILDIAH 2017). Advances in Social Science Education and Humanities Research* **97**, 9-14. (2017)
3. V. Sabitha, S.K. Srivatsa. *International Conference on Information Communication and Embedded Systems, ICICES 2017* **7** (2017)
4. A.A. Kalashnikova, O.M. Akay, I.V. Tsarevskaya, M. S. Volodina, E.O. Tsybenko, Man in India **97**, 161 (2017)
5. V. Crescenzi, G. Mecca, P. Merialdo, RoadRunner: *VLDB 2001 - Proceedings of 27th International Conference on Very Large Data Bases*, 109-118 (2001)
6. J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, W.-Y. Ma, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 494 (2006)
7. D. Castellanos-Nieves, J. Tomás Fernández-Breis, R. Valencia-García, R. Martínez-Béjar, M. Iniesta-Moreno, Inf. Sci. **181**, 1517 (2011)
8. A. Sorici, G. Picard, O. Boissier, A. Zimmermann, A. Florea, Comp. Electrical Eng. **44**, 280 (2015)
9. M. Abdou, S. AbdelGaber, M. Farhan, Future Gener. Comput. Syst. **81**, 94 (2018)
10. K.T. Soo, Innovation across cities. J. of Reg. Science **58**, 295 (2018)
11. A.Z. Broder, S.C. Glassman, M.S. Manasse, G. Zweig, Computer Networks and ISDN Syst. **29 (8)**, 1157 (1997)
12. S. Fagan, R. Gençay, An introduction to textual econometrics, *Handbook of Empirical Economics and Finance* (2010)
13. A. Esteban-Gil, J. Fernández-Breis, D. Castellanos-Nieves, R. Valencia-García, F. García-Sanchez, Procedia - Social and Behavioral Sciences **1(1)**, 927 (2009)
14. T. Berners-Lee, J. Hendler, O. Lassila, Sci. Am. **5**, 34 (2001)