Shortest path algorithm for graphs in instances of semantic optimization

O Yu Lavlinskaya[1], T V Kurchenkova[2] and O V Kuripta[3]

[1]Voronezh Institute of High Technology, st. Lenina 73-A, Voronezh, 394043, Russia
[2]Voronezh State University, 1 Universitetskaya pl., Voronezh, 394018, Russia
[3]Voronezh State Technical University, 20 let Oktyabrya st., 84, Voronezh, 394006, Russia

E-mail: lavlin2010@yandex.ru

Abstract. This article discusses an example of using the algorithm for finding the shortest path to determine the strength of the connection between different words and phrases of the subject area in the tasks of Internet marketing, such as website promotion, thematic traffic, contextual advertising. The input data is words and phrases obtained from search queries and the frequency of their use in the network. Based on the initial data, a weighted graph is constructed, a adjacency matrix is drawn up. The shortest path is determined by the algorithm Floyd-Warshall. The algorithm finds links between words and the force of communication. This approach allows you to find unique semantic combinations and make the necessary search queries. Based on the analysis of the adjacency matrix and the definition of the coherence of words, marketers get a unique tool for solving problems in the field of digital marketing.

1. Introduction

Let's consider the application tasks of using the optimization algorithm (search for the shortest path) in poorly structured areas, such as digital marketing, which includes such tasks as search optimization (SEO), contextual advertising, attracting site visitors, contextual advertising, thematic traffic. Digital marketing is one of the fastest growing areas of activity, including economy, management, information and telecommunications marketing in a global digitalization of societies and economies.

SEO is a digital marketing area related to the promotion of web resources to the top top of the search results. Consumers of services are active Internet users: Companies and businessmen who have Internet business or use the Internet to search for potential clients. The goal of SEO is to develop a semantic core, evaluate which words and phrases should contain the content of the site and how to gain a competitive advantage in website promotion [1].

Contextual advertising is an opportunity to give an ad to the user in accordance with his search query on the Internet. Contextual advertising is an opportunity to give an ad in accordance with search query on the Internet. Attracting visitors to the site is an activity related to finding consumers who were not interested in a product or service, and the task of the marketer is to interest this category of consumers and attract them to the web resource [2].

Working in this environment requires from internet marketing specialist the use of scientific approaches in the organization of work to achieve a competitive advantage.

All these tasks, as is search engine optimization, selection of contextual advertising or attracting the target audience are solved by methods of semantic analysis. Glossary of semantics is the compilation of the thesaurus of the subject area and the finding of connections between words and the determine the weight of this connection in the form of the number of requests words, phrases and sentences on a given topic [3].

At the same time, marketers analyze the selection of search queries, for instance, using the Service Yandex "WordStat", and choose words or phrases that have the most weight (more requests per month or, conversely, looking for those phrases, which are extremely rare, but whose use makes the content unique). The analysis is done manually or using automation tools on the principles of using the service's API [4].

The problem is that with a wide subject matter area of 500 to 5,000 words, depending on the scope of the business and the breadth of the services or goods offered, the number of combinations is n^n [5, 6]. Evaluate the diversity of links and pick up options that meet marketing goals (the most popular phrases, or sentences, or finding phrases that are not met in queries but can be used in marketing solutions without automated analysis algorithms.

2. Materials and Methods

Consider a web resource that provides internet marketing services. The target audience is network users who are interested in promoting their own web resource and need help of an SEO promotion specialist. At the same time, users find a service on a semantic request, formulated on the basis of the thesaurus of the subject area. Our objectives:

Task 1. Offer users content that includes words in phrases and sentences so that this set of words has a strong connectivity.

Task 2. Build a popularity rating for words from phrase requests.

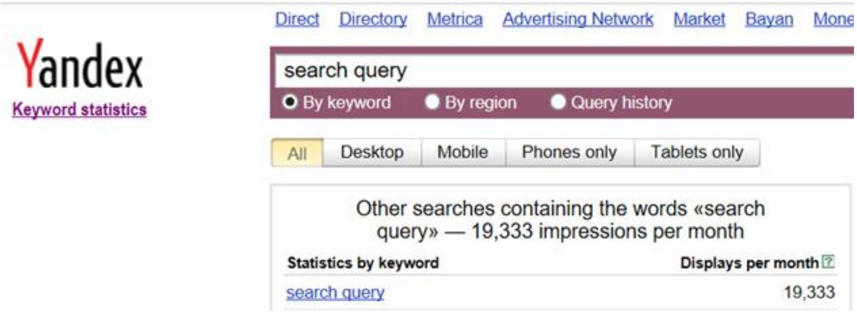
Task 3. Find missing connections between words to get a unique phrase.

For example, let's limit ourselves to seven of words, although, in fact, the thesaurus is 500 to 3000 words, depending on the scope of activity. In the table 1 is a set of words. Words used: optimization, search, SEO, promotion, site, search query, content.

Table 1. A set of words for semantic analysis.

	Words
1	Optimization
2	Search
3	SEO
4	Site
5	Promotion
6	search query
7	Content

For a given set of words, analyzes the requests in the service Yandex Direct – "WordStat" (word selection). On figure 1 presents the service interface and search results for a combination of "SEO" and "search query". The order of words matters.



Other searches containing the words «search query» — 19,333 impressions per month

Statistics by keyword	Displays per month
search query	19,333

Figure 1. Result of a search query using the service "WordStat".

Compile a table of raw data: A combination of words by serial number, a search query, a connectivity weight – a normalized search query indicator (see table 2).

Table 2. Table of the original data.

A combination of words	Search query (*m*)	Normalized reference, Weight (*w*=*m*/*a*)
1–3	1512	1.512
4–5	133022	133.022
6–3	0	–
2–7	1185	1.185
1–7	214	0.214
3–7	283	0.218
2–4	85407	85.407
3–1	11918	11.918
3–2	412	0.412
3–6	54	0.05
5–2	572	0.572
7–2	0	–

Weight normalization is done according to the formula *w*= *m* /*a*, where is *a*=1000 – normalization option needed to bring the number to a user-friendly view, m – indicator of search query.

3. Mathematical model

Consider a weighted graph $G = (V, E)$, having tops V and arcs E ($|V| = n$, $|E| = m$). We will mark the weights $w_{ij} = w(v_i, v_j)$, $v_i, v_j \in V$. П Let graph G – Oriented (figure 2). Orientation shows the preference for combining words in a query.

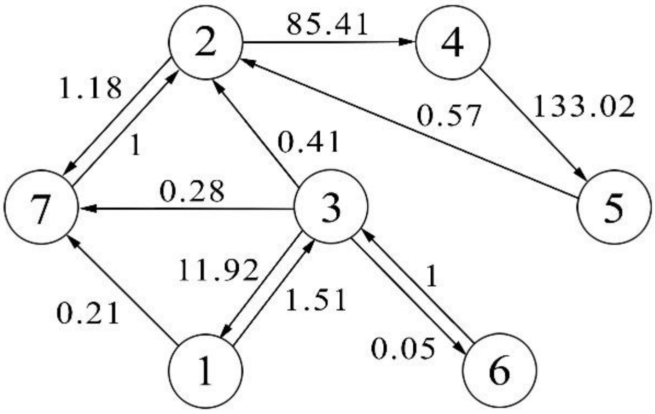


Figure 2. An example of a weighted oriented graph.

Let's solve task 1. Consider the problem of finding a strong connection between words based on the Floyd-Warshall algorithm. Floyd-Warshall algorithm is an algorithm for finding shortest paths in a weighted graph with different weights (unlike other algorithms for finding the shortest path) [7]. The Floyd-Warshall algorithm compares all possible paths through the graph between each pair of vertices. Every combination of edges is tested in graph. Each pair of vertices is compared. Optimization is achieved by gradually improving the estimation of the shortest path between two vertices until the evaluation is minimal.

The vertex adjacency matrix is made

$$A = (a_{ij}), 0 \le i, j \le n - 1\,, \tag{1}$$

The weight of the graph's edges is set

$$a_{i,j} = \begin{cases} w_{i,j}, & \text{if } (v_i, v_j) \in E, \\ 0, & \text{jf } i = j, \\ \infty, & \text{else} \end{cases} \tag{2}$$

Let's make adjacency matrix according to formula 2, and if the link is identified, then $w_{i,j}$ =1.

Table 3. Adjacency matrix.

	Optimization	Search	SEO	Site	Promotion	Search query	Content
Optimization	0.00	∞	1.512	∞	∞	∞	0.214
Search	∞	0.00	∞	85.407	∞	∞	1.185
SEO	11.918	0.412	0.00	∞	∞	0.05	0.218
Site	∞	∞	∞	0.00	133.022	∞	∞
Promotion	∞	0.572	∞	∞	0,00	∞	∞
Search query	∞	∞	1	∞	∞	0.00	∞
Content	∞	1	∞	∞	∞	∞	0.00

Let the way $p = (v_0, v_1, \dots, v_i, v_j, \dots, v_{k-1})$. Let this path find the maximum connection or, then the path $p_{0,i} = (v_0, \dots, v_i)$, $p_{i,j} = (v_i, v_j)$, $p_{j,k-1} = (v_j, \dots, v_{k-1})$ will also be the shortest.

Since the length of the path is made up of the sum of the lengths of its parts, it is to the path p add the vertex v_k. There are two options:

If the length (weight-based cost) of the path $p_{i,j}$ less cost of the path $p'_{i,j} = (v_i, v_k, v_j)$, the shortcut won't change.

If length (cost) $p_{i,j}$ more length $p'_{i,j} = (v_i, v_k, v_j)$, then the shortest way will be $p' = (v_0, v_1, \dots, v_i, v_k, v_j, \dots, v_{k-1})$.

Let's change the serial of vertex using indices from 0 to $n-1$.

Identify the matrix $D = (d_{i,j})$, where distances between the vertices without intermediate vertex of the value of the elements which $d_{i,j}$, $0 \le i, j \le n-1$ match with the scales $w_{i,j}$ moving from the vertex i to j. If the edge $e_{i,j}$ missing, then $d_{i,j} = \infty$ moreover $d_{i,i} = 0$.

$$d_{i,j} = \begin{cases} w_{i,j}, & \text{if } e_{i,j} \in E, \\ 0, & \text{if } i = j, \\ \infty, & \text{else} \end{cases} \tag{3}$$

Thus, the original distance matrix coincides with the graph's incidence matrix.

If $d^k_{i,j}$ – it's the length of the path through the intermediate vertex.

k, then D^k – It's a size matrix $n \times n$, where the elements (i, j) coincides with $d^k_{i,j}$. To bypass all

$k-$ tops consistently accepting values $0, 1, \dots, n-1$ calculate D^{k-1} matrix elements D^k, Applying a recurrent ratio

$$d^k_{i,j} = \begin{cases} \min(d_{i,j}, d_{i,k} + d_{k,j}), & \text{if } k = 0, \\ \\ \min(d^{k-1}_{i,j}, d^{k-1}_{i,k} + d^{k-1}_{k,j}), & \text{if } k > 0, \end{cases} \tag{4}$$

Matrix $D^{n-1} = (d^{n-1}_{i,j})$, will be a matrix containing the lengths of the shortest paths for all pairs of tops $i, j \in V$. It's easy to find simple a recurrent ratio:

$$P_{i,j} = \begin{cases} NULL, & if \ \ w_{i,j} = \infty, \\ i, & if \ \ i = j, \\ 1, & if \ \ w_{i,j} < \infty. \end{cases}$$

$$P^0_{i,j} = \begin{cases} P_{i,j}, & if \ \ d_{i,j} \le d_{i,k} + d_{k,j}, \\ \\ P_{k,j}, & if \ \ d_{i,j} > d_{i,k} + d_{k,j}. \end{cases} \quad (k = 0)$$

$$P^k_{i,j} = \begin{cases} P^{k-1}_{i,j}, & if \ \ d^{k-1}_{i,j} \le d^{k-1}_{i,k} + d^{k-1}_{k,j}, \\ \\ P^{k-1}_{k,j}, & if \ \ d^{k-1}_{i,j} > d^{k-1}_{i,k} + d^{k-1}_{k,j}, \end{cases} \quad (k > 0) \tag{5}$$

The complexity of the sequentially Floyd-Warshall algorithm has an order of n^3, which means that when the number of tops in the graph is doubled by 8 times, and so on [8]. To reduce the complexity of computing, the technology of parallelizing the calculation process is used.

The idea behind the parallel Floyd-Warshall algorithm is that the operation of the recurrent search for the shortest paths is carried out independently. In accordance with the overall scheme of the Floyd-Warshall algorithm, it is necessary $n-1$ times perform a similar operation that updates the shortest path matrix.

In order to organize parallel calculations, it is necessary that in one thread in a recurrent ratio the consistency of data preservation for a single stream is observed.

Applied calculations used a parallel implementation of the Floyd-Warshall algorithm based on OpenMP technology, taking into account the nuances described in the [9].

The algorithm continues its work until the current vertex k will not visit all the top of the many vertices V.

We will use the PageRank line-up algorithm in task 2. PageRank line-up algorithm a forward ranking algorithm [10].

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. In PageRank, graph nodes are web pages, and hyperlinks between pages are edges that have weight. The algorithm ranks web pages by popularity. The algorithm can be used in the problems of semantic analysis to rank words and phrases in the construction of the semantic core and determine the popularity of words and phrases.

Warshall algorithm was implemented in the task 3 [11]. Warshall algorithm finds transitive closure of directed graphs. In Warshall's original formulation of the algorithm, the graph is unweighted and represented by a Boolean adjacency matrix.

A transit closure is a data structure that allows to answer questions about feasibility. That is, can moved from node *i* to node *j* directly or through *k* vertices? A binary relationship indicates that there is edge between the two vertices.

Problem 3 is solved by converting the adjacency matrix into a binary matrix, replacing significant connections with a binary unit, and all other values with binary zeros.

4. Results

Let's show the result of the algorithm on the calculated example.

Task 1. The results of the definition of the power of communication between words through the *K* word. For a calculated example, we get a matrix of connectivity to find the force of communication between words, represented in a table 4.

Table 4. Adjacency matrix by the Floyd-Warshall algorithm.

	Optimization	Search	SEO	Site	Promotion	search query	Content
Optimization	0	2	1	3	4	2	1
Search	∞	0	∞	1	2	∞	1
SEO	1	1	0	2	3	1	1
Site	∞	2	∞	0	1	∞	3
Promotion	∞	1	∞	2	0	∞	2
search query	2	2	1	3	4	0	2
Content	∞	1	∞	2	3	∞	0

Task 2. Determining significance based on weights. To determine the significance based on the scales with the PageRank algorithm [12].

Let's imagine a graph that renders the visualization of the scales. From the picture you can see that the most popular in the thesaurus is the word "Site", followed by the word "Promotion" and then "Search".

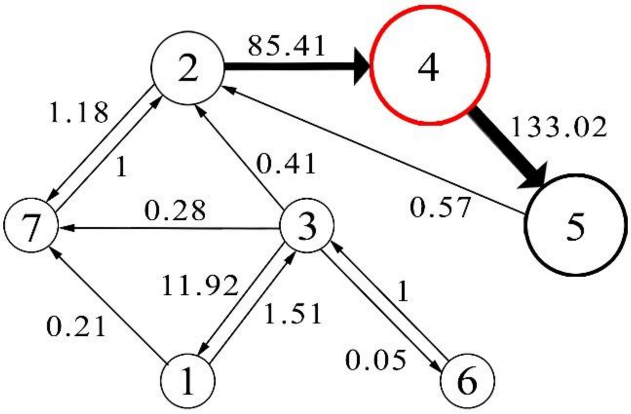


Figure 3. Ranking by weight.

Task 3. The definition of missing semantic relationships on the basis of the search algorithm for transit closure allows you to build unique phrases that were not included in the semantic core. Figures 4a and 4b show a visual result for the example in question, where the algorithm found a missing link between the words "promotion" – "search query". New connections can be used for a unique semantic core or key identification for contextual advertising. As practice shows, this approach allows us to find unique combinations that have new meaning.

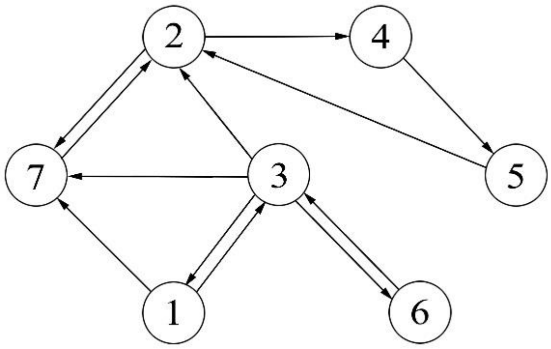


Figure 4a. Original graph (Warshall's algorithm).

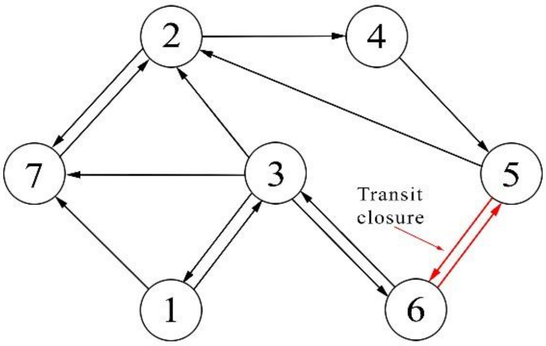


Figure 4b. Search for transit closure.

5. Conclusions

The approach presented in the article was used for semantic analysis as part of a study for the service "Semantica.in", which provides semantic analysis services in a variety of ways: site promotion, contextual advertising, contextual traffic. Algorithms are implemented in the language of the C# in the Visual Studio environment.

The original data is collected on "WordStat Yandex Direct" resource through API, JSON technology. A monthly report of the popularity of queries is formed and algorithm is analyzed on graphs.

The application of this approach allows obtaining additional information on the popularity of words, creating unique phrases, choosing keywords and drawing up a plan in semantic analyzing [13].

Innovation is that the SEO problems did not use the word ranking algorithms when constructing the semantic kernel. It turned out that ranking allows to solve problems of website promotion more qualitatively by means of allocation of priority phrases.

The effect of this approach positively evaluated by marketers. In the future, it is planned to elaborate a service of automated collection of semantic data and their analysis using algorithms on graphs.

References

[1] Joe Wilson Schaefer 2018 *Content Marketing: Essential Guide to Learn Step-by-Step the Best Content Marketing Strategies to Attract your Audience and Boost Your Business* (North Charleston (SC): CreateSpace Independent Publishing Platform) p 246

[2] Kenny D and Marshall J 2000 Contextual Marketing: The Real Business of the Internet *Harvard Business Review* Available at: https://hbr.org/2000/11/contextual-marketing-the-real-business-of-the-internet (accessed 25 October 2019)

[3] Ortiz-Cordova A and Jansen B J 2012 Classifying Web Search Queries in Order to Identify High Revenue Generating Customers *Journal of the American Society for Information Sciences and Technology* **63** (7) p 1401

[4] *Yandex Direct Service* Available at: https://wordstat.yandex.ru (accessed 25 October 2019)

[5] Gergel V P et al 2013 Parallel *Computing: Technology and Numerical Methods: Learning manual in 4 volumes* (N. Novgorod: Nizhny Novgorod State University Publishing House) p 239

[6] Kormen T, Leyzerson C, Rivest R and Shtain K 2006 *Algorithms: construction and analysis* (Moscow: Williams) p 1296

[7] Levitin A V 2006 *Algorithms. Introduction to development and analysis* (Moscow: Williams) pp 345–9

[8] Bernard Roy 1959 Transitivité et connexité *C. R. Acad. Sci. Paris* **249** pp 216–8

[9] Karpov A 2008 Debugging and optimization of multithreaded OPENMP programs *RSDN Magazine* **4** pp 32–6

[10] Schweikin V V and Tanayev I V Application of graph theory in the page rank algorithm PAGERANK *The scientific community of students of the 21st century. Technical Sciences: Sat. Art. On the mat. XLI Internar. Stud. science.-practical conf.* **4** (40) Available at: https://sibac.info/archive/technic/4 (40).pdf (accessed 25 October 2019)

[11] Stephen Warshall 1962 A theorem on Boolean matrices *Journal of the ACM* **9** (1) pp 11–2

[12] Batura T V and Murzin F A 2008 *Machine-oriented logical methods of displaying the semantics of text in natural language: monograph* (Novosibirsk: NSTU) p 248

[13] Lavlinskaya O Yu and Kurchenkova T V 2017 Application of graph theory in structural and topological analysis of information systems *Scientific records of Belgorod State University. Series: Economy. Informatics* **23** (272) pp 105–12